Digital Capture and Retention Guidelines

Why digitization standards and guidelines?

The variety and richness of the Yale Library collections (consisting of materials in all formats and from all time periods and parts of the world), combined with the growing number of initiatives undertaken to digitize them, emphasize the need for a unified set of guidelines for consistency and efficiency. Such a document, based on past experiences as well as on federal agencies' specifications and recommendations, outlines specifications for the digital capture and retention of all major format types, as listed below.

Who should use these guidelines?

The documents available below are intended primarily for all parties involved in the planning and implementation of future projects undertaken by the Yale Library, such as

- Yale Library staff
- Partner institutions participating in cooperative digitization projects
- Outsourcing vendors

Why use these guidelines?

The main purpose of these guidelines is to facilitate, not to limit and constrain, and project managers may decide to follow them in full or in part, depending on their particular needs. In providing minimum, rather than overspecific standards, they stress the fact that their adoption may require some degree of flexibility and latitude due to the partidular factors affecting each individual project. These may involve

- *External factors*, such as funding agencies or donors (and their requirements); partner institutions (and their needs, etc.); vendors (and their equipment, workflow, etc.)
- *Project-specific factors*. To some extent, each project is unique due to a combination of factors including: (1) The type of materials that are being digitized, their characteristics and provenance; (2) the audiences that are being targeted; (3) the intended uses and forms of presentation (e.g., online exhibition, temporary or otherwise, versus digital collection, etc.); (4) funding realities; (5) workflow issues (outsourcing, etc.)
- *Technology-specific factors*. These have to do with the constant evolution of all the technological aspects and components that are involved in a digitization project, including DAM solutions as well as file formats and the software applications involved in their management and preservation.

Glossary

Below is a list of terms that recur in the minimum specifications for the content types covered by these guidelines. For a general glossary of terms please refer to the Federal Agencies Digitization Guidelines Initiative (FADGI) Glossary available at: http://www.digitizationguidelines.gov/glossary.php [1]

ACCESS (copy) is also known as the User/Patron format.

BLACKLIGHT

BORN-DIGITAL are assets that originated in digital form, such as Web sites, wikis, e-books, digital sound recordings, and email.

CCITT Group IV is an image compression schema based on the "Comité Consultatif International Téléphonique et Télégraphique"), a telecommunications standard created in 1956

CHECKSUM is a function used for validating data integrity. Also referred to as MD5 (*Message-Digest algorithm 5).* An algorithm or formula is applied against the source (typically a file and its content, such as the image of a scanned page from a book) in order to generate a unique, 128-bit hash value often called a *checksum*. In digital preservation processes, the MD5 checksum from when the content was created is compared to another checksum created after the content has been received or stored over a period of time. The values are compared and, if they match, this indicates that the data (e.g. the scanned page image) is intact and has not been altered.

COLOR (specifications)

FEDORA (http://www.fedora-commons.org/ [2]) (Flexible Extensible Digital Object Repository Architecture) is a software framework to construct and maintain repositories of digital objects.

HYDRA (http://projecthydra.org/ [3]) is an open source respository software solution.

JPEG (Joint Photographic Experts Group) is the name of the group that developed the standard. JPG is a compression method for images.

JPEG 2000 is a wavelet-based image compression standard. It was created by the Joint Photographic Experts Group committee in the year 2000 with the intention of superseding their original discrete cosine transform-based JPEG standard (created about 1991). The standardized filename extension is JP2.

LADYBIRD (http://web.library.yale.edu/lit/projects/ladybird [4]) is Yale University Library's home-grown ingest tool for cataloging non-MARC records for digitized materials for display in a digital library interface on the web.

LZW (Lempel-Ziv-Welch) is a universal lossless data compression algorithm created by Abraham Lempel, Jacob Ziv, and Terry Welch.

MASTER (copy) is also known as Preservation Master or Archival Master.

OCR (Optical Character Recognition), computer software designed to convert images of text (usually captured by a scanner) into machine-editable text

PDF (Portable Document Format) is a file format, created by Adobe Systems, for document exchange in a manner independent of the application software, hardware, and operating system.

PRINT is determined when reformatting a fragile, brittle, or otherwise vulnerable volume characterized as such because of its physical condition.

PROCESSED (copy) is also known as the Mezzanine copy or the Access master or the Working master format.

RATIO

RESOLUTION

RETENTION

TIFF (Tagged Image File Format) is recognized as the best format for preservation and technical longevity.

TXT (.txt) is a file format used for textual documents usually containing very little formatting.

UTF-8 (UCS Unicode Transformation Format—8-bit) is a form of encoding that is backwards compatible with ASCII. The encoding standard is capable of displaying in email and in Internet browsers the standard 128 ASCII characters for English as well as Latin alphabet characters with diacritics, Greek, Cyrillic, Coptic, Armenian, Hebrew, and Arabic characters.

Effective Date: March 19, 2014

All content types

Still Image

Resources predominantly visual in nature, such as two-dimensional visual arts (drawing, painting, graphics, photographic prints and negatives, plans), manuscript books and archival documents, early printed books, or books printed in alphabets or fonts for which Optical Character Recognition (OCR) is cannot be implemented.

Minimum Specifications

	Master	Processed	Access	OCR source	Text	Thumbnail	Print source??
File type	TIFF	TIFF	JPEG	n/a	n/a	JPEG	n/a
Resolution	400 PPI*						
Pixel array	4000 pixels on the long side						
Bit depth	24-bit RGB**						
Color space	eciRGBv2 or Adobe RGB 1998						
Retention	Permanent	Permanent	Create on ingest			Create on ingest	
Ratio	1:1						

* The resolution should be calculated from the dimensions of the object.

** Whenever possible, capture at 48-bit RGB (for color, monochrome, objects with stains or marks) and save at 24–bit RGB.

Textual

Resources that are expressed through a form of notation intended to be read, especially printed documents (books, serials, pamphlets, posters, broadsides, etc.) for which OCR (Optical Character Recognition)-extracted text should be provided as an output of the digitization workflow.

Minimum Specifications

	Master	Processed	Access	OCR source	Text	Thumbnail	Print source
File type	TIFF	TIFF	JPEG	TIFF	UTF-8 compliant .txt file	JPEG	PDF*
Resolution	300 PPI**				n/a	72 PPI	n/a
Bit depth	8-bit grayscale***			Bitonal, adjusted for contrast and			n/a

				brightness			
Retention	Permanent	Permanent	Create on ingest	Temporary	Permanent	Create on ingest	
Ratio	1:1						

* The maximum size for PDF files shlould not exceed 100 MB.

** The resolution should be calculated from the dimensions of the object and the size of the text. For large text, 10pt or higher, 300 PPI should suffice. A document with smaller text, 9pt or lower, may require 400-600 PPI.

*** Whenever possible, capture at 48-bit RGB (for color, monochrome, objects with stains or marks) and save at 24–bit RGB. Or capture at 16-bit grayscale (for most objects where color is not a concern) and save at 8-bit grayscale. Use 1-bit bitonal for clean, high-contrast documents with printed type only.

Audiovisual

Audiovisual materials at Yale University Library include analog audio and video formats, digital audio and video formats, and film. Below is a chart showing AV preservation format recommendations YUL is considering adopting in accordance with a consulting report submitted by Audiovisual Preservations Solutions in June 2013. Two competing standards are currently in use at YUL: (1) uncompressed 10-bit video, which results in large files, and (2) mxf-wrapper motion JPEG 2000, which is not 100% open source.

Minimum Specifications

Format	Preservation Master	Access Master	Access Copy
Audio-Analog	- Broadcast Wav File (BWF) wrapper - PCM uncompressed - 24-bit - 96kHz	n/a	- MPEG Audio Layer 3 (MP3) - Bitrate 256Kbps
Audio-Digital	- Broadcast Wav File (BWF) wrapper - Native uncompressed data at original sample rate and bit-depth	n/a	- MPEG Audio Layer 3 (MP3) - Bitrate 256Kbps
Video - Analog Standard Definition	- Quicktime wrapper (.mov extension) - Video encoded as 10-bit YUV 4:2:2 uncompressed (v210) - Audio encoded as uncompressed PCM, 48kHz - Maintain original aspect ratio, recording standard, interlacing, number of audio channels, and auxiliary information such as original timecode and	- Quicktime wrapper (.mov extension) - Video encoded as DV - Audio encoded as uncompressed PCM, 48kHz - Maintain original aspect ratio, recording standard, interlacing, number of audio channels, and auxiliary information such as original timecode and closed captioning	- MPEG4 wrapper (.mp4 extension) - Video encoded as H.264 - Audio encoded as uncompressed AAC, 44.1kHz, 256Kbps - Bitrate 5.0Mbps

	closed captioning		
Video-Digital	- Native encoding and data rate in Quicktime wrapper (.mov extension), e.g. DV for MiniDV and DVCam - Maintain original aspect ratio, recording standard, interlacing, number of audio channels, and auxiliary information such as original timecode and closed captioning	n/a	- MPEG4 wrapper (.mp4 extension) - Video encoded as H.264 - Audio encoded as AAC, 44.1kHz, 256kbps - Bitrate 5.0Mbps
Film	16 and 8 mm: - 2k 10-bit RGB 4:4:4 DPX log - Uncompressed 96kHz/ 24-bit Broadcast Wav File (BWF) for audio 35 mm: - 4k 10-bit RGB 4:4:4 DPX log - Uncompressed 96kHz/ 24-bit Broadcast Wav File (BWF) for audio	- MXF wrapper - AVC-Ultra	- MPEG4 wrapper (.mp4 extension) - Video encoded as H.264 - Audio encoded as uncompressed AAC, 44.1kHz, 256kbps - Bitrate of 5.0mbps

3D Objects

Objects, artifacts, and three-dimensional works of visual art encountered in archives, galleries, and museums (medals and badges, physical evidence from legal archives, some works of art).

Although the Yale Library collections include three-dimensional objects (such as artists' books in the Arts Library Special Collections, or globes in the Map Collections), guidelines for this format type are to be developed at a later time.

Born Digital

Born-digital materials arrive at Yale University Library from a variety of sources. Curators at different libraries acquire data tapes, floppy disks, CD-ROMs or hard drives from donors. Staff in different Yale offices transfer born-digital records to University Archives. Researchers work with staff at Marx Science and Social Science Library to preserve their datasets. The file formats and genre of these materials vary widely.

Born-digital objects should be preserved in their native file formats whenever possible. In addition, different preservation master copies and/or access copies may be created. The format of these copies will vary depending on the needs of staff and researchers, the library's digital preservation policy, and available resources. Guidelines about preferred file formats may be developed at a later date.

Minimum Specifications

Sample content	Sample formats
Text	.doc, .wpd, .txt, .rtf, .pdf, .xls
Notated music	
Visual	.pdf, .jpg, .tif, JPEG2000
Visual - Cartographic	.tif, JPEG2000, GeoTIFF, .drg
Visual - Plans and diagrams	.dxf, .dwg
Digital audio	.wav, .mp3
Digital video	.mov, mp4
Three-dimensional	
Database	.fp, .dat, .mdb, .mdf, .nsf
Software	
Datasets	
Email	.mbox, .pst, .eml, .msg
Websites	
Video games	

Still image

Definition

Resources predominantly visual in nature, such as two-dimensional visual arts (drawing, painting, graphics, photographic prints and negatives, plans), manuscript books and archival documents, early printed books, or books printed in alphabets or fonts for which Optical Character Recognition (OCR) is cannot be implemented.

External Resources

Minimum Digitization Capture Recommendations [5](Association for Library Collections & Technical Services, June 2013)

DFG Practical Guidelines on Digitisation [6](Deutsche Forschungsgemeinschaft, February 2013)

Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files [7] (Federal Agencies Digitization Guidelines Initiative, August 2010)

Minimum Specifications

	Master	Processed	Access	OCR source	Text	Thumbnail	Print source??
File type	TIFF	TIFF	JPEG	n/a	n/a	JPEG	n/a
Resolution	400 PPI*						
Pixel array	4000 pixels on the long side						
Bit depth	24-bit RGB**						
Color space	eciRGBv2 or Adobe RGB 1998						

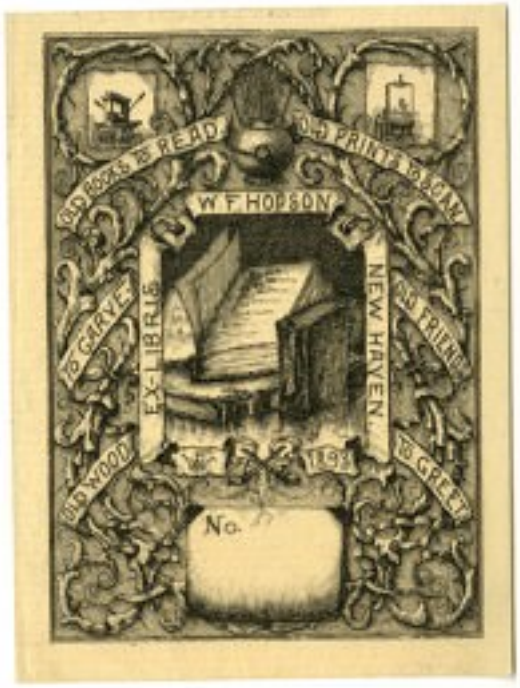
Retention	Permanent	Permanent	Create on ingest			Create on ingest	
Ratio	1:1						

* The resolution should be calculated from the dimensions of the object. See Samples for more information and example images.

** Whenever possible, capture at 48-bit RGB (for color, monochrome, objects with stains or marks) and save at 24–bit RGB.

Samples

The best practice is to have at least 4000 pixels on the longest edge. However, when the original object is small, as is Fig. 1, the project manager should weigh the relative benefits, considering such factors as the maximum optical resolution of the equipment, file size, and server space.



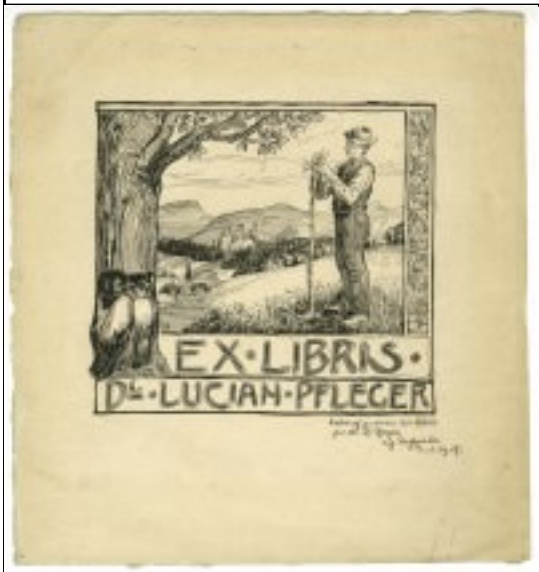
Resolution (PPI) and the resulting pixel array:

300 PPI = 922 x 1201

600 PPI = 1786 x 2366

800 PPI = 2467 x 3303

Fig. 1. [Ex Libris W.F. Hopson, New Haven] by William Fowler Hopson, x 7.0 cm. Collection of Bookplates by William Fowler Hopson, Call # BK



Resolution (PPI) and the resulting pixel array:

300 PPI = 3554 x 3785

400 PPI = 4739 x 5047

600 PPI = 7061 x 7522

[Ex Libris Dr. Lucian Pfleger] by Eug. Moßgraber, 1907, 30.8 x 29 cm. C of Original Artwork for Bookplate Designs, Call # BKP 131.

Textual

Definition

Resources that are expressed through a form of notation intended to be read, especially printed documents (books, serials, pamphlets, posters, broadsides, etc.) for which OCR (Optical Character Recognition)-extracted text should be provided as an output of the digitization workflow.

External Resources

Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files [7] (Federal Agencies Digitization Guidelines Initiative, August 2010)

Minimum Specifications

	Master	Processed	Access	OCR source	Text	Thumbnail	Print source
File type	TIFF	TIFF	JPEG	TIFF	UTF-8 compliant .txt file	JPEG	PDF*
Resolution	300 PPI**				n/a	72 PPI	n/a
Bit depth	8-bit grayscale***			Bitonal, adjusted for contrast and brightness			n/a
Retention	Permanent	Permanent	Create on ingest	Temporary	Permanent	Create on ingest	
Ratio	1:1						

* The maximum size for PDF files should not exceed 100 MB.

** The resolution should be calculated from the dimensions of the object and the size of the text. For large text, 10pt or higher, 300 PPI should suffice. A document with smaller text, 9pt or lower, may require 400-600 PPI.

*** Whenever possible, capture at 48-bit RGB (for color, monochrome, objects with stains or marks) and save at 24–bit RGB. Or capture at 16-bit grayscale (for most objects where color is not a concern) and save at 8-bit grayscale. Use 1-bit bitonal for clean, high-contrast documents with printed type only.

Last modified: Monday, March 24, 2014 - 8:39am

Audiovisual

Definition

Audiovisual materials at Yale University Library include analog audio and video formats, digital audio and video formats, and film.

External Resources

One Format Does Not Fit All: FADGI Audio-Visual Working Group's Diverse Approaches to Format Guidance [8] (October 31, 2013)

Carl Fleischhauer, "Format Considerations in Audio-Visual Preservation Reformatting: Snapshots from the Federal Agencies Digitization Guidelines Initiative [9]," *Information Standards Quarterly* 22, no. 2 (Spring 2010): 34-40.

Library of Congress Audio-Visual Conservation Resources [10](Last updated: 8/3/2007)

Guidelines

Below is a chart showing AV preservation format recommendations YUL is considering adopting in accordance with a consulting report submitted by Audiovisual Preservations Solutions in June 2013. Two competing standards are currently in use at YUL: (1) uncompressed 10-bit video, which results in large files, and (2) mxf-wrapper motion JPEG 2000, which is not 100% open source.

Format	Preservation Master	Access Master	Access Copy
Audio-Analog	- Broadcast Wav File (BWF) wrapper - PCM uncompressed - 24-bit - 96kHz	n/a	- MPEG Audio Layer 3 (MP3) - Bitrate 256Kbps
Audio-Digital	- Broadcast Wav File (BWF) wrapper - Native uncompressed data at original sample rate and bit-depth	n/a	- MPEG Audio Layer 3 (MP3) - Bitrate 256Kbps
Video - Analog Standard Definition	- Quicktime wrapper (.mov extension) - Video encoded as 10-bit YUV 4:2:2 uncompressed (v210) - Audio encoded as uncompressed PCM, 48kHz - Maintain original aspect ratio, recording standard, interlacing, number of audio channels, and auxiliary information such as original timecode and closed captioning	- Quicktime wrapper (.mov extension) - Video encoded as DV - Audio encoded as uncompressed PCM, 48kHz - Maintain original aspect ratio, recording standard, interlacing, number of audio channels, and auxiliary information such as original timecode and closed captioning	- MPEG4 wrapper (.mp4 extension) - Video encoded as H.264 - Audio encoded as uncompressed AAC, 44.1kHz, 256Kbps - Bitrate 5.0Mbps
Video-Digital	- Native encoding and data rate in Quicktime wrapper (.mov extension), e.g. DV for MiniDV and DVCam - Maintain original aspect ratio, recording standard, interlacing, number of	n/a	- MPEG4 wrapper (.mp4 extension) - Video encoded as H.264 - Audio encoded as AAC, 44.1kHz, 256kbps

	audio channels, and auxiliary information such as original timecode and closed captioning		- Bitrate 5.0Mbps
Film	16 and 8 mm: - 2k 10-bit RGB 4:4:4 DPX log - Uncompressed 96kHz/ 24-bit Broadcast Wav File (BWF) for audio 35 mm: - 4k 10-bit RGB 4:4:4 DPX log - Uncompressed 96kHz/ 24-bit Broadcast Wav File (BWF) for audio	- MXF wrapper - AVC-Ultra	- MPEG4 wrapper (.mp4 extension) - Video encoded as H.264 - Audio encoded as uncompressed AAC, 44.1kHz, 256kbps - Bitrate of 5.0mbps

3D Objects

Definition

Objects, artifacts, and three-dimensional works of visual art encountered in archives, galleries, and museums (medals and badges, physical evidence from legal archives, some works of art).

External Resources

Minimum Digitization Capture Recommendations [11](Association for Library Collections & Technical Services, June 2013)

DFG Practical Guidelines on Digitisation [6](Deutsche Forschungsgemeinschaft, February 2013)

Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files [7] (Federal Agencies Digitization Guidelines Initiative, August 2010)

Guidelines

Although the Yale Library collections include three-dimensional objects (such as artists' books in the Arts Library Special Collections, or globes in the Map Collections), guidelines for this format type are to be developed at a later time.

Born digital

Definition

Born-digital materials arrive at Yale University Library from a variety of sources. Curators at different libraries acquire data tapes, floppy disks, CD-ROMs or hard drives from donors. Staff in different Yale offices transfer born-digital records to University Archives. Researchers work with staff at Marx Science and Social Science Library to preserve their datasets. The file formats and genre of these materials vary widely.

Born-digital objects should be preserved in their native file formats whenever possible. In addition, different preservation master copies and/or access copies may be created. The format of these copies will vary depending on the needs of staff and researchers, the library's digital preservation policy, and available resources. Guidelines about preferred file formats may be developed at a later date.

External Resources

Given the relative novelty of this field, standards and guidelines are still being developed.

Links
[1] http://www.digitizationguidelines.gov/glossary.php [2] http://www.fedora-commons.org/ [3] http://projecthydra.org/ [4] http://web.library.yale.edu/lit/projects/ladybird [5] http://www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations [6] http://www.dfg.de/formulare/12_151/12_151_en.pdf [7] http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf [8] http://blogs.loc.gov/digitalpreservation/2013/10/one-format-does-not-fit-all-fadgi-audio-visual-working-groups-diverse-approaches-to-format-guidance/ [9] http://www.digitizationguidelines.gov/audio-visual/documents/IP_Fleischhauer_AudioVisual_Reformatting_isqv22no2.pdf [10] http://www.loc.gov/avconservation/preservation/resources.html [11] http://www.ala.org/alcts/resources/preserv/minimum-digitization-capture-recommendations#3d_objects