Digital Collections Technical Guidelines

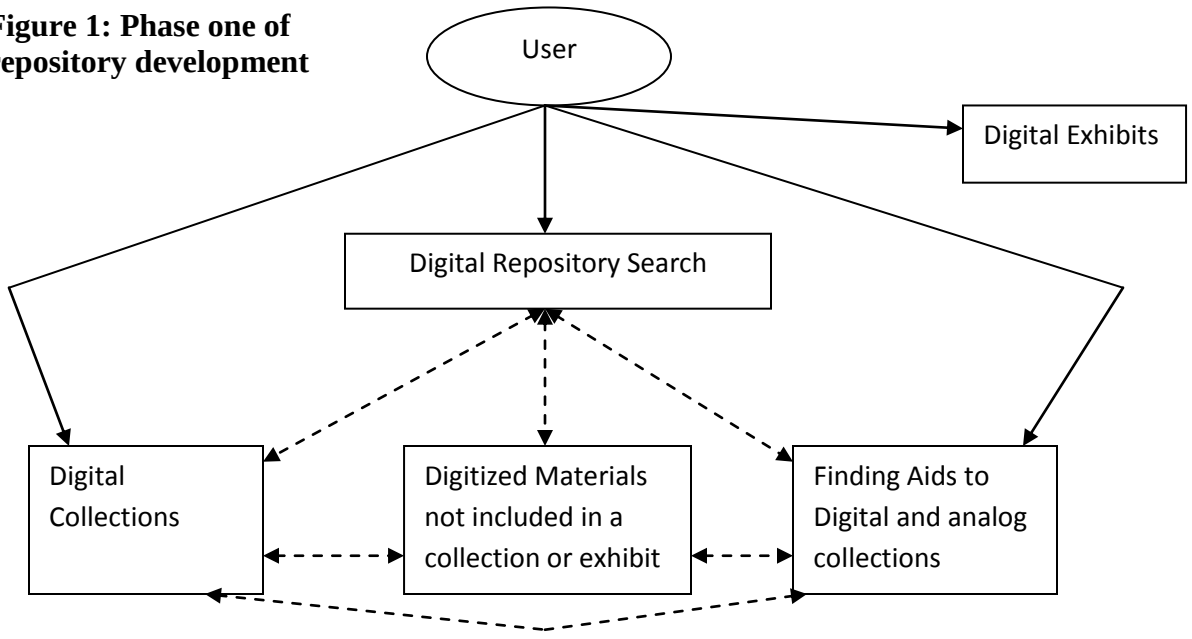
1. Repository Basics

Joyner Library's Digital Collections now utilizes a common repository for the material it digitizes. Built using the Ixiasoft's TEXTML server software, the repository is a native XML storage/retrieval database populated with METS XML records. Each digitized object (text, image, audio, video, or some combination of these) is described in a METS record with links to the digitized content (for a fuller description of METS, see section 2) which is then indexed for searching in TEXTML. A web based asset management system then provides users with access to the repository's content and the digitized objects themselves.

The repository will provide one searchable interface for discovery and access to all of the digitized objects created by Digital Collections from this point forward. Exceptions exist however, for materials digitized for previous HTML-based exhibits and finding aids. Although Digital Collections also manages the creation of EAD encoded finding aids, these are not indexed or searched within the repository. However, metadata records for digital objects contain links back to their collection finding aids when appropriate and possible. In addition, the finding aids contain links to materials in the repository (excepting those in digital exhibits at the current time). Additionally, materials digitized for previous exhibits currently exist only in their isolated HTML files. Future plans include migrating these objects into METS records within the repository.

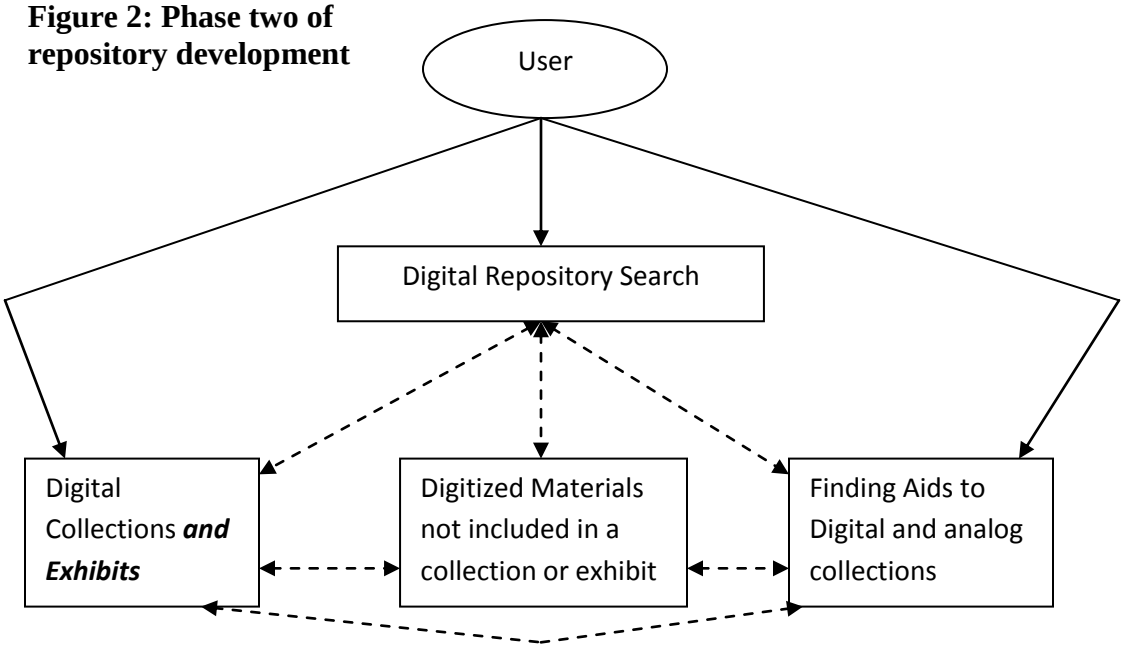
In the current system, a user may interact with digitized objects in four different ways. They may browse through existing digital exhibits, they may search and browse through existing digital collections, they may access digitized items through links in the finding aids, or they may search the repository which includes all the objects except those in the digital exhibits.

Figure 1: Phase one of repository development



Once the migration of the digital exhibits is completed, users will be able to discover objects from those exhibits through a repository search as well as through related finding aids.

Figure 2: Phase two of repository development



1.1.Requests for digitization

All items digitized by Digital Collections from the Library's holdings for user request, staff selection, or digital projects will be stored in the repository and available to users through the repository search and additional outlets as necessary.

Staff wishing to have materials digitized for user request or staff selection can complete the Digital Imaging Request form at http://digital.lib.ecu.edu/request/. At the time of request, the staff member must include:

- their name
- a brief description of the request itself
- the format, resolution, or size desired
- the date the request is needed by.

Staff may also indicate whether the request is for a Staff Pick. If this option is chosen, staff will only need to complete the name, date needed, and identifier.

In addition, staff are asked to supply some descriptive metadata to facilitate the addition of materials to the digital repository as they await full attention from the metadata librarian including:

- an identifier indicating the item's location — for example, a collection/series/box/folder number or a call number (required)

- title (required)
- creator
- date
- description of the item
- the assignment of the item to one of several theme-based collections.

1.2.Digital preservation

Storage in the repository does not constitute digital preservation. Although Digital Collections is committed to providing continuing access to their digitized output, the archiving necessary to constitute digital preservation has not yet been achieved.

2. Metadata

All objects ingested into the digital repository are described using the **Metadata Encoding and Transmission Standard (METS)**. The METS record is divided into 7 sections, 5 of which are required by the digital repository:

1. <metsHdr>
   contains basic identifying information about the digital object including when the record was created, whether it is complete, and who created it
2. <dmdSec>
   contains descriptive metadata about the original object. This metadata may be created in compliance with another descriptive metadata standard and referenced through the use of namespaces, or can be completely original to the METS record. Digital Collections requires a <dmdSec> with MODS descriptive metadata as well as a second <dmdSec> with Dublin Core elements (an initial <dmdSec> is also currently used to house a link to the "finding aid" of the collection for which that particular object is part, in those instances when the object is from a collection in Special Collections).
3. <amdSec>
   contains the technical or administrative metadata describing the digital object. For images, this information should be created according to the MIX standard and referred to by namespaces. For audio elements, a standard set of elements relating to the Library of Congress's AudioMD data dictionary (http://www.loc.gov/rr/mopic/avprot/DD_AMD.html) will be used. No specific DTD or schema is referenced by this standard.
4. <fileSec>
   contains a list of all of the files associated with the digital object including master and derivative files. The location of the files is given using a URL or persistent identifier, and the sequence of the files in the object is given
5. <structMap>
   contains a mapping of the digital files to the structure of the original object. So if files 1, 5, and 8 represent page one of the document, they would be grouped accordingly.

2.1. Other Metadata Schemes used:

2.1.1. *Metadata Object Description Schema (MODS)*

A metadata schema created by the Library of Congress to create bibliographic information for a number of purposes. Each element in the MODS record is repeatable.The MODS record is the descriptive metadata schema used by the digital repository. Specific elements and usages for the digital repository include (Each element in the MODS record is repeatable.):

- titleInfo [@type= uniform]
  - title
- abstract
- name [@ type = personal | corporate | conference]
  - namePart
  - role
  - roleTerm (creator | contributor)
- identifier [@ type = local | doi | job | preferredCitation]
- originInfo
  - dateIssued [@ point = start | end], [@ keydate = yes]
  - publisher
- language
  - languageTerm [@authority=”iso639-2b”]
- typeOfResource [still image | text | sound recording | three dimensional object | moving image]
- physicalDescription
  - form [@authority=”aat” | “local”]
  - extent
- subject [@authority =”lcsh” | “drgrant” | “fast” (hierarchicalGeographic)]
  - topic
  - geographic
  - genre
  - name
  - title
    - *and if hierarchicalGeographic*
  - country
  - state
  - county (for NC only)
  - city
- relatedItem (for “Staff Picks”, Encore “themes”, Child/Parent elements, etc.)
- accessCondition [@type = useAndReproduction]
- accessCondition [@type = pd | copyrighted | orphaned | uncertain | joyner]
  - [These represent the 5 levels of copyright that we are currently recording for all our items.]
- location
  - physicalLocation

2.1.2. *Dublin Core (DC)*

A descriptive metadata element set created by the Dublin Core Metadata Initiative. DC consists of 16 basic elements to describe any digital object. While the standard is very simple, it has great utility when sharing records for this very reason. Digital Collection requires the inclusion of a DC record with the metadata for every digital object to facilitate the sharing of record through OAI harvesting. With the exception of the publisher element, which is a static value represented J.Y.Joyner Library, these DC records are generated from the MODS record and include the following elements:

- title (from <mods:title>)
- creator (from <mods:name> where <mods:role> is "creator")
- contributor (from <mods:name> where <mods:role> is “contributor”, and/or from <mods:publisher>)
- description (from <mods:abstract>)
- date (from <mods:dateIssued>)
- subject (from <mods:subject>[@lcsh])
- coverage (last element in <mods:subject>[@fast])
- format (from <mods:form>)
- publisher (static "J.Y. Joyner Library")
- language (from <mods:language>)
- type (from <mods:typeOfResource)
- rights (from <mods:accessCondition>)
- identifier (from <mods:identifier>)

2.1.3. *Metadata for Images in XML (MIX)*

A technical metadata standard for capturing information about the characteristics and capture of digital images. Much of the creation of MIX data is automated by the JSTOR/Harvard Object Validation Environment (JHOVE). The following elements are created by Digital Collections.

2.1.4. *AudioMD:*

A preliminary schema created by the Library of Congress for the capture of technical specifications of digital audio files. Much of the creation of the AudioMD data is automated by JHOVE’s WAVE-Hul module. The following elements are recorded in our AudioMD:

- <amd:audio_block_size>
- <amd:audio_data_encoding>
- <amd:bits_per_sample>
- <amd:byte_order>
- <amd:checksum>
  - Datetime
  - Type
  - Value
- <amd:first_sample_offset>

- <amd:format_name>
- <amd:num_sample_frequency>
- <amd:duration>
- <amd:num_channels>

Additionally, if the sound recording is not in preservation Broadcast WAVE format, we will still record the following:

- <amd:audio_block_size>
- <amd:checksum>
- Datetime
  - Type
  - Value
- <amd:format_name>

2.1.5. *Text Encoding Initiative (TEI)*

A full text encoding standard that captures structural details of texts (see http://www.tei-c.org for more information). The TEI schema consists of a header containing bibliographic information and a body section containing the marked-up text. In general, we subscribe to a level 3 encoding, as outlined by the Digital Library Federation's *Guidelines for Best Encoding Practice* (http://www.diglib.org/standards/tei.htm) with the addition of the following:

- <pb/> *required*
- <name> *optional*
- <persName> *optional*

Exceptions exist for transcripts of oral histories, which also use:

- <speaker>
- <milestone> *to indicate page breaks in the printed version of the transcript and to enhance readability online*

2.1.6. Encoded Archival Description (EAD)

Information about our implementation of EAD is found at: http://ead2002.pbwiki.com

3. Image/Text Digitization

These specifications are based on the *California Digital Library Guidelines for Digital Images* (California Digital Library 2008) and *BCR CDP Digital Imaging Best Practices Version 2.0 (BCR 2008)* .

3.1.Text

Clean, high contrast, typed material where smallest significant > 1mm

	Master	Access	Thumbnail
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	8 bit grayscale	8 bit grayscale	8 bit grayscale 8 bit indexed color (GIF)
Resolution	400 DPI	300 DPI	72 DPI
Dimensions	100% of original	800 pixels across long dimension	200 pixels across long dimension

Material with poor legibility, handwritten annotations, fading, halftone illustrations, or where smallest significant < 1mm

	Master	Access	Thumbnail
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	8 bit grayscale	8 bit grayscale	8 bit grayscale 8 bit indexed color (GIF)
Resolution	600 DPI	300 DPI	72 DPI
Dimensions	100% of orignial	800 pixels across long dimension	200 pixels across long dimension

Material where color is important to the interpretation or the most accurate representation is desired and where smallest significant > 1mm

	Master	Access	Thumbnail
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	24 bit color	24 bit color	24 bit color
Resolution	400 DPI	300 DPI	72 DPI
Dimensions	100% of orignial	800 pixels across long dimension	200 pixels across long dimension

Material where color is important to the interpretation or the most accurate representation is desired and where smallest significant < 1mm

	Master	Access	Thumbnail
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	24 bit color	24 bit color	24 bit color
Resolution	600 DPI	300 DPI	72 DPI
Dimensions	100% of original	800 pixels across long dimension	200 pixels across long dimension

3.1.1. Text Conversion

3.1.1.1. OCR

Digital Collections uses OmniPage Pro 14 software. This package has an error rate of less than 2% and can support 114 languages. To work well, texts must be clearly printed in machine-created typeface. Fonts with heavy serifs or other highly stylized characteristics may not be compatible with the software. Handwritten material cannot be converted using OCR software. Once automatic conversion is done, Digital Collections will proof material to ensure the error rate is acceptable.

3.1.1.2. Transcription

When OCR is not a possibility, Digital Collections can provide transcriptions for scanned texts based on the project goals and the availability of resources.

3.2.Photographica

3.2.1. Transmissive Originals (Film, Slides, and Negatives)

Material up to 4" x 5" or a total of 20" square

	Master	**Access**	**Thumbnail**
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	8 bit grayscale	8 bit grayscale	8 bit grayscale 8 bit indexed color (GIF)
Resolution	4000 pixels across the long dimension	300 DPI	72 DPI
Dimensions	100% of original	800 pixels across long dimension	200 pixels across long dimension

Material between 4" x 5" and 8" x 10" or 20" to 80" square

	Master	**Access**	**Thumbnail**
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	8 bit grayscale	8 bit grayscale	8 bit grayscale 8 bit indexed color (GIF)
Resolution	6000 pixels across the long dimension	300 DPI	72 DPI
Dimensions	100% of original	800 pixels across long dimension	200 pixels across long dimension

Material > 8" x 10" or 80" square

	Master	Access	Thumbnail
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	8 bit grayscale	8 bit grayscale	8 bit grayscale 8 bit indexed color (GIF)
Resolution	8000 pixels across the long dimension	300 DPI	72 DPI
Dimensions	100% of original	800 pixels across long dimension	200 pixels across long dimension

The following advice about scanning photographic negatives should be followed:

> Often photographic negatives are the most difficult originals to scan. Unlike scanning positives, reflection prints, and transparencies or slides, there are no reference images to which to compare scans. Scanning negatives is very much like printing in the darkroom — it is up to the photographer/technician to adjust brightness and contrast to get a good image. Also, most scanners are not as well calibrated for scanning negatives compared to scanning positives.
>
> To minimize the loss of detail, it is often necessary to scan negatives as positives (the image on screen is negative), invert the images in Photoshop, and then adjust the images.
>
> If black-and-white negatives are stained or discolored, we recommend making color RGB scans of the negatives and using the channel that minimizes the appearance of the staining/discoloration when viewed as a positive. The image can then be converted to a grayscale image. (California Digital Library 2008)

3.2.2. Reflective Originals (Prints)

Material up to 8" x 10" or 80" square

	Master	Access	Thumbnail
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	24 bit color	24 bit color	24 bit color
Resolution	4000 pixels across the long dimension	300 DPI	72 DPI
Dimensions	100% of original	800 pixels across long dimension	200 pixels across long dimension

Material between 8" x 10" and 11" x 14 or 80" t0 154" square

	Master	Access	Thumbnail
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	24 bit color	24 bit color	24 bit color
Resolution	6000 pixels across the long dimension	300 DPI	72 DPI
Dimensions	100% of original	800 pixels across long dimension	200 pixels across long dimension

Material > 11" x 14" or 154" square

	Master	Access	Thumbnail
Format	TIFF	JPEG	JPEG/GIF
Bit Depth	24 bit color	24 bit color	24 bit color
Resolution	8000 pixels across the long dimension	300 DPI	72 DPI
Dimensions	100% of original	800 pixels across long dimension	200 pixels across long dimension

Exclude mounts or borders on long dimension when calculating resolution.
It is good practice to capture the border area in the master scan file. In cases where a small image is mounted on a large board, it may be desirable to scan the image area only at the appropriate resolution for its size, and then scan the entire mount at a resolution that achieves 4,000 pixels across the long dimension.

4. Audio Digitization

Digital Collections is equipped to do a small amount of audio conversion, primarily from the compact cassette format. These guidelines were developed from BCR's CDP Digital Audio Working Group's *Digital Audio Best Practices* (2006).

Type of Recording	Sampling Rate	Bit Depth	Digital File Format
Spoken Language (master)	44.1 kHz	24-bit	WAV
Spoken Language (access)	44.1 kHz	16-bit	MP3
Field Recordings – Spoken Language (master)	44.1 kHz	24-bit	WAV
Field Recordings – Spoken Language (access)	44.1 kHz	16-bit	MP3
Field Recordings – Natural Sounds (master)	96 kHz	24-bit	WAV
Field Recordings – Natural Sounds (access)	96 kHz	16-bit	MP3
Music (master	96 kHz	24-bit	WAV
Music (access)	96 kHz	16-bit	MP3

5. References

California Digital Library. (2008) *CDL Guidelines for Digital Images*. Retrieved 24 November 2008 from http://www.cdlib.org/inside/diglib/guidelines/bpgimages/.

BCR's CDP Digital Imaging Best Practices Working Group. (2008). *BCR's CDP Digital Imaging Best Practices Version 2.0*. Retrived 24 November 2008 from http://bcr.org/cdp/best/digital-imaging-bp.pdf.

BCR's CDP Digital Audio Best Practices Working Group. (2006). *Digital Audio Best Practices Version 2.1*. Retrieved 16 December 2008 from http://www.bcr.org/cdp/best/digital-audio-bp.pdf.