2014-06

Lorraine Cheng Digitization Center Procedures and Best Practices

Project Planning

Proposed projects must be submitted to the Digital Projects Librarian (DPL) via the form and instructions located at http://www.wpunj.edu/library/lcdw.dot. Proposed projects are reviewed by the Digital Projects Librarian (DPL) to determine if they can be completed in-house or need to be outsourced. Proposals are brought before the Digital Projects Committee (DPC) to be considered and prioritized. Requestors are notified of the approval or rejection of a project.

Once a project has been approved, requestor will arrange for delivery or pick-up of project materials.

Upon delivery, the requestor and at least one member of the DPC will review inventory of items that will be digitized, giving description specifics, noting damage, binding condition and issues.

Requestor must provide a spreadsheet with metadata for the project (see metadata section of this document for more information) before work on project may begin.

For in-house projects, DPL will determine equipment and software needs for scanning or conversion of materials. Materials will then be processed in accordance with the below workflow(s) and on an agreed-upon schedule.

For outsourced projects, DPL will write up a set of specifications for the project, and send them to digitization vendor(s) for a quote. The quote must then be approved by the appropriate Budget Officer [through this process] and returned to the vendor. The requestor and the DPL (alone or with a member of the DPC) will create an inventory of items, and assign an accession number to each item. The items and inventory are packaged and sent to the vendor for digital processing. Rare or fragile items may be hand-delivered and picked up by the DPL. Items and digital files are returned to the DPL upon completion. DPL will inspect the returned materials and make sure they meet the specifications.

Outsourcing Workflow

- Before materials are sent to vendor
    - A short list of digitization vendors should be created and annually reviewed/updated by the DPC for future outsourcing projects.
    - An RFP (request for proposal) should be created, as detailed as possible, for each outsourced project. The RFP should be submitted to a minimum of three vendors for quote.

    - An inventory spreadsheet with descriptions, special instructions, and photographs of any delicate or damaged materials should be created before materials are delivered to vendor. The vendor should know the exact condition of materials and specifics for special handling, if any, before receiving
    - Materials should be sorted, prepared (unbiding, etc), and packaged for vendor
    - Provide archival gloves for delicate materials
    - Materials may be either shipped, picked up, or hand delivered, depending on condition of materials and location of vendor
- After materials are returned from vendor
    - Check materials against spreadsheet to verify all has been returned and for any new unreported damage.
    - Check all digital files for quality and to ensure agreed upon imaging specifications were used.

<u>Copyright Considerations</u>

- Nothing under copyright should be scanned unless express written permission from the author is obtained.
- Public domain materials, and (often times) university owned materials, including public relations materials, may be scanned without obtaining permission.

Project Workflow for Text Based Objects

1. Project Management

- Projects are tracked in the TEAMWORK web-based project management software.
- Each new document to be digitized must be added as an individual "Task" on the "Tasks" tab of the LCDC page.
- Sub-Tasks for each Task should be completed in order (i.e. do not start creating CONTENTdm collections and entering metadata until imaging and editing are complete).
- Keep accurate track of status and progress for anyone else who might be working on same project. Tasks should be checked off only when complete and notes should be entered and shared with the appropriate people when necessary.
- Only work on one task at a time. Do not start a second task until the one you are working on is complete.

2. Digital Imaging

File Storage

- Scanned images are stored temporarily on the local machine and permanently on the network drive for archival purposes.
- Files from each scanning session should be temporarily stored in a folder on the desktop. When each scanning session is complete, files should then be transferred to a folder at K: /Groups/Library/Library Digital Content/*Project*/*Object*, where:
    - *Project* is the current project (e.g. "WPUUndergraduateCourseCatalogs") and
    - *Object* is the current document being scanned (e.g. WPUCatalog1980-81).

Choosing a scanner

- For multi-page, loose leaf documents with a clean edges up to 8.5x11 inches, use the Xerox Documate 3125. Volumes which can be unbound and have clean edges should also be scanned with the Xerox.
- For photographs, documents, and bound books and any documents with rough or uneven edges up to 11x17 inches, use the Plustek OpticBook A300. In addition, some items larger than 11x17 inches can be scanned in pieces and merged (See "Digital Editing" below).
- For most maps, photographs, newspapers, and other items over 11x17 inches, use large format digital photography.

File Naming Conventions

- Filenames should be comprised of two parts, an identifier, or prefix (often the LOC call number or the object title), and a counter, or suffix. If a source object has no unique identifier, one should be created. The suffix should consist of a four digit number separated from the prefix by an underscore. (Most scanning software should add this suffix to your file name as it iterates through multiple pages scanned in which case you only need add the underscore to the prefix). File name examples:
    - N5_220_H62_E38_0001
    - hobartManorSideView_0003
- All filenames should be consistent. TIFs, PDFs, and JPGs should have identical names for a single object:
    - N5_220_H62_E38_0001.tif
    - N5_220_H62_E38.jpg
    - N5_220_H62_E38.pdf

Imaging Specifications

- Color photos and documents
    - Should be scanned at *minimum* 600 dpi for archival purposes.
    - Color space should be sRGB, at least 24 bit truecolor. 8 bits per channel is preferred.
    - Archival files should be in TIFF format, encoded with no compression or (where available) a lossless compression algorithm (such as LZW).

- Black and white documents with gray tones, or black and white photos
    - Should be scanned at a *minimum* of 600 dpi.
    - Color space should be grayscale. RGB is acceptable, but contributes nothing to pure grayscale images and only increases file size.
    - Archival files should be in TIFF format, encoded with no compression or (where available) a lossless compression algorithm (such as LZW).

- Black and white documents with pure text
    - Should be scanned at 600 dpi.
    - Color space should be 2 bit color (black and white, or bitmap), but must be converted to grayscale if the documents are to be digitally resized.
    - Archival files should be in TIFF format, encoded with no compression or (where available) a lossless compression algorithm (such as LZW).

- Additional Considerations, or the 3,000-pixel rule
    - Scanning at resolutions much higher than 600dpi may be required if the document is small in physical size.

- All digitized documents should be scanned such that at least one dimension of the visible area of the document (width or height) is comprised of a minimum of 3,000 pixels each. If a scan at 600 dpi does not meet this minimum, then the resolution should be increased and the document re-scanned accordingly. *It is not acceptable under any circumstances to artificially increase the dpi by resampling the image. The original scan must meet the 3,000 pixel minimum.*
- Extraneous image information (such as a calibration strip, record labels, or any part of the scanned image that does not directly pertain to the document itself) should not be included in this calculation. Measurements should be made using only the visible area of the document.
- Example: a 3"x4" color photograph scanned at 600 dpi will render an image that is 1800 x 2400 pixels. As the width does not meet the minimum, a 600dpi scan will be unacceptable. The scan should be increased to 800 dpi to render a 2400 x 3200 pixel image, and thus bring the image to an acceptable level of detail.
  On the other hand, an 8.5"x11" sheet of paper with black and white text will scan acceptably at 400dpi, rendering a 3400 x 4400 pixel image.*

3. Digital Editing

- Edit images in Adobe Photoshop.
- In the most cases, editing of TIFFs should be limited to cropping and rotating/straightening images. Color/contrast/brightness/sharpness correction should only be used when absolutely necessary, as archival digital images should capture the original image as closely as possible, not an improved or idealized version.
- If significant changes to the archival images are necessary (color/contrast/brightness/sharpness correction, inverting of images, degrading of image quality for presentation format/size issues), the new edited TIFFs *in addition to* the original unedited archival TIFFS should be saved. Any presentation formats (see below) should be made using the new edited TIFFs.
- When imaging a multi-page document, all pages should be cropped as closely as possible to the same size.
- Pages should be oriented to the most convenient to read on a computer screen:
  - If a multi-page volume to be imaged contains some pages such as charts or photographs in landscape orientation—where the volume must be turned sideways to view properly—those images should be rotated 90 degrees as to appear in landscape orientation on the computer screen
- Occasionally there will be documents too large for either the oversized flatbed scanner or large format digital photography and must be imaged in pieces. The “photomerge” feature in Photoshop can be used to stitch these two (or more) pieces together to produce a single digital image. Once the separate pieces have been merged, rename the new file in accordance with the standard filenaming conventions and discard the individual pieces.

4. Presentation Format Creation

Because archival quality TIFF files are generally too large to be easily displayed in web pages and digital library software, presentation format files are made from TIFF files for a better end-user experience. The LCDC uses two industry standard presentation formats for their collections, JPG and PDF.

JPGs

- JPGs are used for single page items such as photographs and maps.
- Create JPGs from Archival TIFFs using Adobe Photoshop.
- Standard JPGs are used currently, as CONTENTdm inconsistently displays JPG2000 images. Ultimately LCDC should move towards JPG2000 as its compression algorithm is lossless.
- JPGs should initially be created at the best quality and with as little compression as possible. If JPGs are over 20 MB, compression may be added.

PDFs

- PDFs are used for multi-page text documents such as reports and books.
- Create PDFs using Adobe Acrobat Pro
- If possible, PDFs should be created with no size reduction or optimizing.
- PDFs must be under 20 MB for CONTENTdm to display properly.
- Rather than optimizing to lowest quality, which will most often result in a substandard presentation PDF, resampling archival TIFs from 600dpi to as low as 300dpi and recreating PDF is preferable.
- Orientation of each page should be as per 'Digital Imaging" notes.
- Every PDF must be OCRed.

5. Metadata Creation

Project Metadata

- Every approved project must have a metadata spreadsheet completed by the requesting party before any digitization work may begin.
- The standard metadata spreadsheet can be found here: http://www.wpunj.edu/library/lcdw.dot.
- Columns headed in green must be completed by requestor.

Metadata profiles

The DPL will create a metadata profile for each project based on that project's content and available metadata based on the following standard CONTENTdm metadata profile for the LCDC:

	Field name	DC map	Data type	Large	Search	Hide	Required	Vocab		add field
1	Title	Title	Text	No	Yes	No	Yes	No	move to ▼	edit | delete
2	Creator	Creator	Text	No	Yes	No	No	No	move to ▼	edit | delete
3	Subject terms	Subject	Text	No	Yes	No	No	No	move to ▼	edit | delete
4	Keywords	Subject	Text	No	Yes	No	No	No	move to ▼	edit | delete
5	Description	Description-Abstract	Text	Yes	Yes	No	No	No	move to ▼	edit | delete
6	Notes	Description-Abstract	Text	No	Yes	No	No	No	move to ▼	edit | delete
7	Language	Language	Text	No	Yes	No	No	No	move to ▼	edit | delete
8	Content type	Type	Text	No	Yes	No	No	No	move to ▼	edit | delete
9	Genre	Format	Text	No	Yes	No	No	No	move to ▼	edit | delete
10	Extent	Format	Text	No	Yes	No	No	No	move to ▼	edit | delete
11	Date	Date	Date	No	Yes	No	No	No	move to ▼	edit | delete
12	Rights	Rights	Text	No	Yes	No	No	No	move to ▼	edit | delete
13	Publisher	Publisher	Text	No	Yes	No	No	No	move to ▼	edit | delete
14	Provenance	Provenance	Text	No	Yes	No	No	No	move to ▼	edit | delete
15	Location	Source	Text	No	Yes	No	No	No	move to ▼	edit | delete
16	Source format	Source	Text	No	Yes	No	No	No	move to ▼	edit | delete
17	Filename	Identifier	Text	No	Yes	No	No	No	move to ▼	edit | delete
18	Operating system	None	Text	No	Yes	No	No	No	move to ▼	edit | delete
19	Creating software	None	Text	No	Yes	No	No	No	move to ▼	edit | delete
20	Creating hardware	None	Text	No	Yes	No	No	No	move to ▼	edit | delete
21	Full text search	None	Full Text Search	Yes	Yes	No	No	No	move to ▼	edit | delete
	Field name	DC map	Data type	Large	Search	Hide	Required	Vocab		add field

Metadata Guidelines

- *Title*: Titles should be taken from a corresponding library catalog record. If no catalog record exists, title should be taken from the document title page, or lacking that, the document cover or first page, whenever possible. Item-in-hand title creation can be used if no item title exists.
    - Authority: APA style capitalization and punctuation.
- *Creator*: Any authors, artists, photographers, cartographers, or other creators should be listed here. When multiple creators exist, names should be separated by semi-colons.
    - Authority: Library of Congress Name Authority File
- *Subject terms*: A standard set of subject headings for a collection should be developed with the WPU cataloger for every item. Additional subjects can then be added to individual items based on content.
    - Authority: Library of Congress Subject Headings

- *Keywords*: Free text keywords or tags can be added here if no authorities can be found.
- *Description*: A short description of the item. If item has an abstract or summary, that can be used (or adapted if very long).
- *Notes*: Information about the item itself, including handwritten markings, date or version information, size in centimeters of a map, or anything else of particular interest that may not be included in the description.
    - Authority: None
- *Language*: The code for the language the item is written in
    - Authority: ISO 639-2 three letter codes
- *Content type*: The form the information in the item takes (e.g. text, image, dataset, etc.)
    - Authority: Dublin Core Metadata Initiative (DCMI) Type Vocabulary
- *Genre*: The type of document (e.g. memo, newspaper clipping, report, map, etc.)
    - Authority: Art & Architecture Thesaurus
- *Extent*: Number of pages in the digital object and relevant information (e.g. 114 p., charts, illustrations)
- *Date*: Enter the latest date on the item (e.g. if revised edition of a book, used the revised date). If item has no date, you may use an inferred or questionable date or a date range but there must be an accompanying note in the *Note* field explaining decision.
    - Authority: ISO 8601
- *Rights*: All items where WPU holds the copyright use "This work is licensed under a Creative Commons Attribution-Noncommercial-ShareAlike 3.0 United States License". In the instance WPU does not hold the copyright for an item, written permission must be obtained from the copyright holder and an appropriate copyright statement should be entered.
- *Publisher*: The publisher of the document being digitized.
    - Authority: Library of Congress Name Authority File
- *Provenance*: Enter the ownership history of the source item if available.
- *Location*: Where source item is located in WPU (e.g. library archive, stacks, etc.)
- *Source format*: The form the original object takes; the carrier.
    - Authority: RDA term list for carrier types
- *Filename*: The filename of the PDF or JPG used as the presentation file.
- *Operating system*: OS for the workstation images were created and edited on (currently Windows 7 Enterprise)
- *Creating software*: Any software used in the creation or editing of images, with version (e.g. Adobe Acrobat 9 Pro v9.0.0)
- *Creating hardware*: Any scanners or digital cameras used in imaging, separated by semicolons.

- *Full text search*: This field will auto populate when items are uploaded to CONTENTdm. In instances where OCR of an object is unsatisfactory, a different full-text record can be pasted to this field

6. CONTENTdm Collection Creation and Management

Creating and Editing Collections

Here are some basic steps guidelines for c

reating and editing collections in CONTENTdm Administration. Admin home page is broken into three tabs:

- Server Tab
    - Add new collections here
    - Collections must be created in Admin before they can be created in the Project Client

- Collections Tab
    - Set up metadata field properties
    - Configure the website look for a particular collection

- Items Tab
    - Approve new items item, make changes to uploaded items, or delete items
    - Always index after creating a project on the server tab, and creating a metadata template on the collection tab
    - Approve and Index items simultaneously. This is the easiest option after uploading items from the local client
    - You may edit collection items from the items tab without downloading to local client.
    - Compound objects count as total number of pages in document +1 for JPGs, but only as one *total* for pdfs

**Some digital imaging specifications from: Rutgers University. "About scanning and capturing documents." DCRC Docs. http://dcrc.rutgers.edu/about/scanning-and-capturing-documents/ (accessed June 23, 2014).*