CARLI Consortium of Academic and Research Libraries in Illinois

GUIDELINES FOR THE CREATION OF DIGITAL COLLECTIONS

Digitization Best Practices for Text

This document sets forth guidelines for digitizing text. Topics covered include: image quality, file formats, optical character recognition, text encoding, storage, and access. Examples of text content in CARLI Digital Collections, a sample workflow, and links to scanning and digitization guides and vendor services are provided as appendices.

This guide was created by the CARLI Digital Collections Users' Group (DCUG).

For questions about this document, please contact CARLI at support@carli.illinois.edu

Introduction

Text materials may include printed matter such as books, magazines, and newspapers, or handwritten or typed original manuscripts, letters, notes, or other documents. For the purposes of this document, "text" refers to any manifestation of words that have been affixed to a physical carrier, paper or otherwise.

Depending on the purpose of the collection, different approaches to digitizing text content may be used. In some cases, libraries may only be interested in the information that the text conveys, and the medium of expression is irrelevant. However, in most collections, it is desirable not only to create a digital representation of the information within the text content itself, but also the visual aspects of the text, such as type, formatting, layout, or paper quality. Text is also often accompanied by image content such as line drawings, photographs, graphic illustrations, manuscripts, music scores, blueprints, plans, etc.

Due to this dual nature, the digitization of texts is very similar to the digitization of image content. To facilitate full-text searching or indexing of the actual text content, additional steps must be taken so that the text can be rendered machine-readable. Text materials also have a further complication in that they are often made up of many pages (as in the case of a book) or may have multiple articles on a single page (such as a newspaper). Decisions must be made as to what unit constitutes a "work"—a single page? an individual article? an entire issue or volume?—and the digitization process should be carried out accordingly.

The sections below provide guidance on the processes of creating digital images, producing machine-readable texts, and combining the two components into a single digital object. Libraries will need to determine which approaches are most appropriate, based on the nature of the project and the importance of the materials being digitized.

Creating Digital Images

In the most basic approach, the physical media to which the text is affixed is scanned to create a digital image that reproduces the content of the work. While the digitized facsimile conveys all of the visual information contained in a text, a digital image does not allow the text to be indexed and searched; additional steps must be taken to provide this functionality.

Digital Image Basics

A digital image is a two-dimensional array of small square regions known as pixels. For each pixel, the digital image file contains numeric values about color and brightness. There are three basic types of digital images: bitonal, grayscale, and color. In the case of a bitonal (monochrome) image, each pixel is either black or white – there is no gradation. Grayscale images typically contain values in the range from 0 to 255 where 0 represents black, 255 represents white, and values in between represent shades of gray. A color image can be represented by a two-dimensional array of Red, Green and Blue triples, where 0 indicates that none of that primary color is present in that pixel and 255 indicates a maximum amount of that primary color.

Bit-depth refers to the amount of detail that is used to make the measurements of color and brightness. (It can be thought of as the number of marks on a ruler.) A higher bit depth indicates a greater level of detail that is captured about the image. Most digital images are 8-bit, 16-bit, or 24-bit.

The size and resolution of digital image files is measured in pixels per inch (ppi, also commonly referred to as dpi—dots per inch). The higher the ppi the greater the resolution and detail that will be captured.

Scanning Basics

Due to the wide varieties of scanners and scanning software available, a comprehensive discussion of best practices for scanner operation is not possible in this guide. "The Art of Scanning" by Paul Royster (http://digitalcommons.unl.edu/ir_information/67/) provides a solid introduction to scanning and image editing techniques for text-based and image-based digital collections.

Scanners generally offer three different modes of image capture, which correspond to the three types of digital images: black-and-white, grayscale, and color.

- **Black-and-White (aka bitonal or monochrome)**: One bit per pixel representing black or white. This mode is best suited to high-contrast documents such as printed black-and-white text, line art, or illustrations.
- **Grayscale**: Multiple bits per pixel representing shades of gray. Grayscale is best suited to older documents with poor legibility or diffuse characters (e.g. carbon copies,

- Thermofax/Verifax, etc.), handwritten documents, items with low inherent contrast between the text and background, stained or faded materials, and works with halftone illustrations or photographs accompanying the text.
- **Color**: Multiple bits per pixel representing color. Color scanning is best suited to materials containing color information, such as an illuminated manuscript or other documents where the color and texture of the paper is an important part of the work.

Scanning in color will produce the largest file sizes (in terms of bytes), grayscale the second largest, and bitonal the smallest. Libraries should choose the mode that best suits the material. If there is no advantage to scanning in grayscale or color, then bitonal mode is acceptable assuming there is no significant loss of information. Master copies can also be created in color or grayscale and then converted to bitonal for access images.

Creating Images

For each object or page being scanned or photographed, a high-resolution master or archival file should be created. From that master file, lower-resolution derivative files will be created that are better suited to be delivered and viewed online or compiled into a file containing all the pages of a work.

The chart below describes the differences between master images and two types of derivative files: an access image and a thumbnail image.

Master Image	**Access Image**	**Thumbnail Image**
• Represents as closely as possible the information contained in the original • Uncompressed, or lossless compression • Unedited • Serves as long term source for derivative files and print reproductions • Can serve as surrogate for the original • High quality • Large file size • Stored in TIFF file format	• Used in place of master image for general web access • Generally fits within viewing area of average monitor • Reasonable file size for fast download time; does not require a fast network connection • Acceptable quality for general research • Compressed for speed of access • Usually stored in JPEG or JPEG2000 file format	• A very small image usually presented with the bibliographic record • Designed to display quickly online; allows user to determine whether they want to view access image • Usually stored in GIF or JPEG file formats • Not always suitable for images consisting primarily of text, musical scores, etc.; user cannot tell what content is at so small a scale

from Western States Digital Standards Group, Digital Imaging Working Group, *Digital Imaging Best Practices*, http://www.mndigital.org/digitizing/standards/imaging.pdf, January 2003.

Master Images

The digital master image represents, as accurately as possible, the visual information in the original object. This image's primary function is to serve as a long-term archival record, as well as a source for derivative files and printed materials. A high-quality master image eliminates the need to re-digitize, and therefore re-handle, the same potentially fragile physical materials again in the future. A master image should also support the production of a printed page facsimile that is a legible and faithful stand-in for the original when printed at the same size.

Some general guidelines for creating digital master files:

- Each library should develop specific guidelines for the size and resolution of digital master files based on individual collection needs and requirements.

- When scanning text documents, the scanning resolution may need to be adjusted according to the size of text in the document. Documents with smaller printed text may require higher resolutions and bit depths than documents containing large typefaces (see "Recommendations" below). A higher resolution may offer increased accuracy for Optical Character Recognition (OCR) processing.

- Scanned master images should not be edited for any specific output or use, and should be saved as large TIFF files with lossless or no compression.

- Where possible, scanning guidelines for the creation of digital master files should follow the specifications outlined in the Federal Agencies Digitization Initiative (FADGI) - Still Image Working Group's Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files (http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf).

- CARLI member libraries using CONTENTdm **should not** upload full resolution TIFF files to the CARLI server as a file-storage solution. Archival image file storage is the responsibility of each contributing institution and must be managed locally. The CONTENTdm Project Client can automatically convert TIFF files into JPEG2000 or JPEG display images. (see "Derivative Images: Access Images" below).

Specific recommendations for size, resolution, and file format are provided below.

Derivative Images

Derivative files are used for editing and enhancement, conversion to different formats, and presentation or transmission over networks. In the case of text works that comprise more than one page, derivative images can be compiled into a single file that represents the entire work.

Derivative images can be created using image editing applications such as Adobe Photoshop, GIMP (a freely available open-source image editing program), or Microsoft Office Picture Editor. Some applications, like Adobe Acrobat, can automatically downsize images when compiling a file made up of multiple page images, eliminating the need to create derivative copies by hand.

Access Images

Access images represent the version of the image that users viewing the digital items online will interact with. Access images should be of sufficient size and resolution to allow for detailed study, but not so large that they take too long to load in the browser. Access images may also be edited to improve the viewing experience for the user, through such processes as cropping, straightening, color correction, sharpening, or descreening. These edits can be made using image editing software such as Adobe Photoshop or GIMP.

In the case of collections using CONTENTdm, the software can also be configured to automatically generate access images from the master file. The default settings of the CONTENTdm Project Client convert imported TIFFs files to either JPEG2000 or JPEG for display.

(Note when using CARLI's installation of CONTENTdm: importing large file formats, like TIFFs, will make upload times longer and will not address an institution's need to store an archival master. Archival image file storage is the responsibility of each contributing institution and must be managed locally.)

Thumbnail Images

Thumbnail images are small, low-resolution versions of the content—usually displayed in the search results view of online digital collections—that give the user a preview of the larger image. Most digital asset management systems will automatically generate a thumbnail image for each item loaded into the software.

Recommendations

	File Format	Pixel Array and Resolution	Bit depth
Master Image	TIFF (.tif)	300 ppi for black-and-white text; 600 ppi for grayscale or color materials, or materials with finely printed text	1-bit bitonal mode, 8-bit grayscale, or 24-bit color
Access Image	JPEG (.jpg) or JPEG2000 (.jp2)	72 – 200 ppi	1-bit bitonal, 8-bit grayscale, or 24-bit color

Based on Federal Agencies Digitization Initiative (FADGI) - Still Image Working Group's Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files (http://www.digitizationguidelines.gov/guidelines/FADGI_Still_Image-Tech_Guidelines_2010-08-24.pdf)

File Naming Conventions

Each digital object in a collection should be assigned a unique identifier. Unique identifiers should follow a consistent naming format to ensure ongoing identification and retrieval of digital files. Guidelines for file names will vary by collection and will be based on local needs and specifications. Each library should develop specific file naming conventions based on individual collection needs and local requirements.

Machine-Readable Text

Machine-readable text results from either a scanning and conversion process (OCR) performed on textual materials or from manually transcribing or re-keying text with word processing software to produce some form of machine-readable text file that can be indexed and searched, offering users better access to the intellectual content of the work.

Text-based materials may be handled in various ways. Methods will depend on factors such as library resources, quality of the original materials, software requirements, and end user needs.

Digital Text Basics

Digital representations of text are based on the concept of character encoding, which is the assignment of a numeric code for each character in a given repertoire to a sequence of bit patterns in order to facilitate the transmission and storage of text in digital form. The character encoding used in a file will determine the type of characters that can be represented in the file. Currently, 8-bit Unicode Transformation Format (UTF-8), is the generally accepted standard for digital texts. UTF-8 encoding can accommodate not only Latin-based language characters, but also Greek, Cyrillic, Hebrew, Arabic, and much more. For these reasons, it is recommended that all textual documents be encoded as UTF-8.

Most computer programs can save text-based documents (plain text files, XML, or HTML) as a UTF-8 encoded document. Additionally, some document formats, such as XML and HTML, provide a way to explicitly declare the file as UTF-8 encoded within the markup, which a parser can then use to interpret the rest of the document. In XML, this can be seen easily in the first line of the file, where the type of file is declared (XML) and so is its encoding (UTF-8). Before saving a text file, check the software's save options to make sure that UTF-8 encoding is being used.

Optical Character Recognition (OCR)

OCR is the process of electronically translating a scanned bitmapped image of text material into machine-readable text. A computer program "reads" the character content within the image and creates a digital version of the text, usually in a separate file. This allows the text to be searched and indexed, or used in other processes such as data mining or machine translation.

The accuracy of the OCR process depends on a number of factors, including the quality of the image being scanned, the language that the text is written in, and the type of font used in printing. Poor quality images where the text is not clearly contrasted with the background, text in non-European foreign languages (or non-Latin character sets), and text rendered in serif fonts can all decrease the accuracy of the resulting text file. At this time, hand-printed manuscripts are extremely difficult for OCR software to interpret, and those written in cursive are basically impossible. However, with a clear typeset image, an accuracy of 80%-90% may be achieved through the use of readily available and relatively inexpensive software.

The advantage of OCR is that it eliminates the need for costly, time-consuming transcription. For most libraries transcription may not be an option, and so even an inaccurate rendering as produced by OCR is still an advantage over having no digital representation of the text at all. OCR routines can also be set up as part of the digitization workflow and do not require a significant time investment. For documents where the accuracy of the machine-readable text is of primary importance, the OCR-produced text can be manually corrected.

Software Options

Libraries using the CARLI installation of CONTENTdm have the option of purchasing the CONTENTdm OCR Extension. This extension can be used to generate a searchable full-text transcription as files are being imported into the CONTENTdm Project Client. The resulting text is then stored as the value of a metadata field in the item record. The OCR Extension is not necessary; the same results can be achieved with transcript files created using standalone OCR software. Other OCR software applications include Adobe Acrobat, ABBYY FineReader, and OmniPage. (Disclaimer: The preceding references to specific applications do not necessarily constitute or imply endorsement or recommendation by CARLI.)

Transcriptions

Text that is difficult to read or that cannot be reliably OCR'd, especially handwritten manuscripts, should be considered for transcription. However, transcription presents its own problems—it can be labor intensive and cost prohibitive—so libraries will need to make a decision as to when the importance of providing full-text searching of the content makes the time investment worthwhile.

In CONTENTdm, the unformatted transcribed text from an image can be entered as the value of a metadata field in the item record, making it full-text searchable.

Text Encoding & Markup Languages

Transcribed text can also be encoded with markup languages, such as XML or XHTML, to provide a digital representation of the semantic and physical document structure. Text encoding provides a machine-readable means of denoting structural text elements such as italics, bold type, line breaks, stanzas, paragraphs, page breaks, chapters, etc. Semantic elements of the text, such as geographical locations or personal names, can also be marked.

The most widely used standard for encoding text-based cultural materials is an XML-based schema developed by the Text Encoding Initiative (TEI). The TEI *Guidelines for Electronic Text Encoding and Interchange* "define and document a markup language for representing the structural, renditional, and conceptual features of texts," with a focus on primary source materials for research and analysis.

Like transcription, text encoding requires a significant investment of resources, and encoded texts require specialized systems and applications to parse, process, index, and display the content in any meaningful way. Currently, CONTENTdm does not provide any special functionality for encoded texts; therefore it is not recommended that libraries pursue this effort if they will be using CONTENTdm to collect and provide access to text-based materials.

Combining Multiple Files into a Single Digital Object

As discussed previously, text materials often consist of many pages that collectively comprise a work. Therefore, a digital facsimile of such a work must include a way to compile many separate scans and images into a single file that maintains the order and structure of the original object. A plethora of digital formats can provide this functionality, including Adobe PDF, DjVu, and ePub.

CONTENTdm also has its own format, called a "compound object." A compound object is two or more individual files bound together with an XML structure. CONTENTdm includes several compound object types. The "document" and "monograph" types support a sequential structure that mimics the paginated nature of text objects, however, the "monograph" type also supports a hierarchical structure, akin to the chapters of a book. CONTENTdm can also be configured to automatically create compound objects from a single Adobe PDF file imported into a collection. (If the PDF file is full-text searchable, the text content will also be imported into CONTENTdm.)

Appendix A. Examples of Text Content in CONTENTdm Collections

American Journeys: Eyewitness Accounts of Early American Exploration and Settlement
Wisconsin Historical Society
http://www.americanjourneys.org/

Arabic Papyrus, Parchment, and Paper Collection
University of Utah
http://content.lib.utah.edu/cdm/landingpage/collection/uuappp

Claremont Coptic Encyclopedia
Claremont University Consortium
http://ccdl.libraries.claremont.edu/col/cce

Florence Nightingale Letters Collection
University of Illinois at Chicago
http://collections.carli.illinois.edu/cdm4/index_uic_fnlc.php?CISOROOT=/uic_fnlc

The Free Soil Banner
Indianapolis Marion County Public Library
http://digitallibrary.imcpl.org/fsb.php

John Muir Correspondence
University of the Pacific
http://digitalcollections.pacific.edu/cdm/search/collection/muirletters

Pamphlet and Textual Documents Collection
University of Washington Libraries
http://content.lib.washington.edu/ptecweb/index.html

Wesleyana Yearbooks
Illinois Wesleyan University
http://collections.carli.illinois.edu/cdm4/index_iwu_yearb.php?CISOROOT=/iwu_yearb

Appendix B: Sample Workflow

This sample workflow demonstrates how to apply the recommendations and best practices outlined in this document. In this example, the item to be digitized is a handwritten letter printed on an 8 ½" x 11" standard sheet of paper. It is possible to divide the workflow into two sections:

Creating the Digital Master Image

1) On the computer, open the scanning application and place the document on the scanner bed.
2) Verify that the scan settings are set to scan the image in grayscale (16-bit) at 600 dpi (ppi). These setting will create a quality archival master file. (A 600 dpi resolution scan of an 8.5" x 11" document would produce an image that is 5100 x 6600 pixels in size.)
3) If possible, preview the item to be scanned and use the scanning software to crop the previewed image to the proper size. Leave a small margin beyond the item's borders to ensure that nothing is accidentally missed when scanning.
4) Scan the image.
5) Once the image is scanned, immediately save the image as an uncompressed TIFF image, giving it a unique name. (Example: JohnDoeCorrespondence_letter1_master.tif)

Creating Two Derivative Images: Access Image and Thumbnail Image

6) Open the scanned digital master TIFF images in an image-editing program..
7) Crop, straighten, and adjust the brightness and contrast of the image, as necessary.
8) Create the access image: Resize the image, maintaining the original proportions, so that the image is now 200 ppi. (A 200 dpi resolution version of an 8.5" x 11" document would produce an image that is 1700 x 2200 pixels in size.)
9) Optional: Modify the image to improve legibility and overall appearance using filters such as sharpen, unsharp mask, descreen, etc.
10) Use "**Save as**" to save the corrected/manipulated image as a JPEG 2000 (.jp2) file at high quality. **Be careful not to overwrite the digital master TIFF image.** Save the access image in a location that separates it from the original and/or append an additional letter or number to the filename to identify it as the access image. (Example: JohnDoeCorrespondence_letter1_access.jp2)
11) Create the thumbnail image: Maintaining the image's proportions, resize the image again, changing the image resolution to 72 dpi, and the size to somewhere between 100-200 pixels on the long side. (Most digital asset management systems, including CONTENTdm, have a feature that automatically generates thumbnail images from uploaded items, so it may be unnecessary to create them manually.)
12) Optional: "Sharpen" the image again, if appropriate.
13) Use "**Save as**" to save the corrected/manipulated image as a JPEG file at high quality. **Be careful not to overwrite the digital master TIFF image or the access image.** Save the thumbnail image in a location that separates it from the digital masters and/or append an additional letter or number to the filename to identify it as the access image. (Example: JohnDoeCorrespondence_letter1_thumb.jpg)

Appendix C: Links and Further Reading

The Art of Scanning
http://digitalcommons.unl.edu/ir_information/67/

California Digital Library: Digital File Format Recommendations: Master Production Files
http://www.cdlib.org/gateways/docs/cdl_dffr.pdf

California Digital Library: TEI Encoding Guidelines
http://www.cdlib.org/groups/stwg/

Federal Agencies Digitization Initiative (FADGI): Technical Guidelines for Digitizing Cultural Heritage Materials: Creation of Raster Image Master Files
http://www.digitizationguidelines.gov/guidelines/digitize-technical.html

A Gentle Introduction to XML
http://www.tei-c.org/release/doc/tei-p5-doc/en/html/SG.html

TEI P5: Guidelines for Electronic Text Encoding and Interchange
http://www.tei-c.org/release/doc/tei-p5-doc/en/html/index.html

Renear, Allen. "Text Encoding," in Schreibman, Susan. *A companion to digital humanities*. Malden Mass.: Blackwell Pub., 2004.

University of Michigan Digital Library Production Services: Digitization Specifications
http://www.hathitrust.org/documents/UMDigitizationSpecs20100827.pdf

U.S. National Archives and Records Administration: Technical Guidelines for Digitizing Archival Materials for Electronic Access: Creation of Production Master Files - Raster Images
http://www.archives.gov/preservation/technical/guidelines.html

Western States Digital Standards Group: Digital Imaging Working Group: Digital Imaging Best Practices
http://www.mndigital.org/digitizing/standards/imaging.pdf

Appendix D: Text Digitization Grant Programs & Vendor Services

(Disclaimer: References below to specific products, resources, or services does not necessarily constitute or imply its endorsement or recommendation by CARLI.)

Digitization Vendors

Backstage Library Works
Provo, Utah & Bethlehem, Pennsylvania
(800) 288-1265
http://www.bslw.com/

Kirtas Technologies
7620 Omnitech Place
Victor, New York 14564
(877) 547-8279
http://www.i2s-digibook.com/

Luna Imaging
2702 Media Center Drive
Los Angeles, CA 90065-1733
(800) 452-5862
http://www.lunaimaging.com/

Northeast Document Conservation Center
100 Brickstone Square
Andover, MA 01810-1494
(978) 470-1010
http://www.nedcc.org

Northern Micrographics
2004 Kramer Street
La Crosse, WI 54603
(608) 781-0850
http://www.normicro.com/

Trigonox, Inc.
1501 Barré, #201
Montréal, QC, Canada H3C 4J1
(514) 874-0443
http://www.trigonix.com/en/html/trigonixProfilcorporatif.html