Guidelines

Creating and managing text collections using CONTENTdm

Text materials are an integral component of multi-format, primary source repositories. Such materials should be readily available to end users through the same Web interfaces used for other material types and genres, such as photographs, posters, broadsides/broadsheets, glass negatives, lantern slides, maps, audio, video and more. CONTENTdm® Digital Collection Management Software has helped institutions like yours manage all their digital resources, no matter what the format, since its inception in 1996.

Guidelines for managing text collections

We're offering these guidelines to help you to create and manage text collections. However, please note that these are strictly to be viewed as guidelines and with every digitization choice you make, you also must consider the access needs and material condition.

Context. First, plan your project within the context of your strategic plan and the needs of your end users. While your text collections may include 'born digital' materials, reformatted items from print, analog tape and microforms, your goal is to enable users to discover your repositories, and then easily retrieve relevant sources in all formats. When digital documents are located, they need to be fully searchable online; the user should be able to read the text retrieved—either in the image itself, and/or in its transcript.

Collaboration. As with most digital projects, text collections are best approached collaboratively. Your institution's departments or branches may want to work together to store materials in a shared digital repository for easy search and retrieval via the Web.

Preparation. Be prepared to reconsider collection description, metadata requirements and item maintenance as necessary to facilitate access.

Testing. Once you've built a prototype of your digital repository, you'll want to do user testing. Then, based on the test results, revise your prototype as needed. Consider prototyping different solutions for each project. Test from start-to-finish, paying attention to usability for source materials in terms of viability for scanning and full-text searching.

Access now and later. Digitize comprehensively: that is, consider digitizing complete collections for long-term preservation, providing access now and in the future.

Implementation. All implementation decisions should be made according to the intended use by your target audiences and the scope of the project, as well as the quality and condition of your originals.

Summary. In the table on page 2, we provide recommendations for filetype/scanning resolutions and CONTENTdm options and file formats to consider, such as OCR (Optical Character Recognition)and JPEG2000, for a variety of source material formats.

Arabic Papyrus, Parchment and Paper Collection at the J. Willard Marriott Library, The University of Utah.

The N. A. Chandler Gold Rush Era Letters collection from the Claremont Colleges Digital Libraries

Guidelines

Creating and managing text collections using CONTENTdm

General recommendations for access using CONTENTdm

Source material	Source file formats	Bit-depth	CONTENTdm options and derivative files to consider[1]
Books High image values	High-resolution TIFF	Color (24-bit)	OCR [2], Monograph, JPEG2000, with PDF generation
Books Text with few illustrations	Medium-resolution TIFF	Grayscale (8-bit) or bitonal (1-bit)	OCR, Monograph or Document, JPEG2000, with PDF generation
Manuscript material Handwritten works (not typeset)	High-resolution TIFF	Grayscale or Color	Monograph or Document, JPEG2000, manual entry of metadata (typescript)
Research papers, theses and dissertations Analog (print or film)	High-resolution TIFF	Grayscale or bitonal	OCR, Monograph or Document, JPEG2000, with PDF generation
Research papers, theses and dissertations Electronic (electronic source available or "born-digital," "ETDs")	PDF	N/A	Load PDF[3] (if text has been embedded, CONTENTdm will extract it for searching across documents and collections.)
Administrative records proceedings, minutes Analog (printed)	Medium-resolution TIFF	Grayscale or bitonal	OCR, Monograph or Document, JPEG2000, with PDF generation
Postcards	High-resolution TIFF	Color or grayscale	Postcard, manual entry of metadata (typescript)
Yearbooks and compilations Scrapbooks, albums, etc.	High-resolution TIFF of entire page	Color or grayscale	OCR, Monograph or Document, JPEG2000, with PDF generation
eSerials Journals, periodicals, magazines, etc.	PDF	N/A	Load PDF (if text has been embedded, CONTENTdm will extract it for searching across documents and collections.
Newspapers—professionally processed	High-resolution TIFF w/METS-ALTO; option for segmentation	N/A	Digitization vendor process and load to CONTENTdm; specify for compatibility with NDNP requirements or better
Newspapers and newsletters—processed in-house	High-resolution TIFF	Grayscale	OCR, Monograph or Document, JPEG2000, with PDF generation
Finding aids Analog (print or film)	Medium-resolution TIFF	Bitonal (1-bit) or grayscale	OCR, Monograph or Document, JPEG2000, with PDF generation
Finding aids 1. EAD 2. Word processed	1. EAD DOC-TYPE, import XML file 2. Convert to PDF, ensure embedded text	N/A	Import XML w/XSL for display. Import PDF

Footnotes

1. Each collection should have one searchable field of data type "Full text search."
2. OCR (Optical Character Recognition). CONTENTdm has integrated ABBYY FineReader® into the Project Client. This Extension is licensed from OCLC as an optional add-on to CONTENTdm. Resolution and bit-depth will vary for optimum accuracy.
3. Before multipage PDFs are imported, set Collection to treat as if compound objects for optimum speed of search and display.

For more information

Please contact OCLC at contentdm@oclc.org. Or, if you are a current CONTENTdm user and have technical questions, please contact CONTENTdm support at contentdmsupport@oclc.org or 877-797-0887.

213348USF 0902, OCLC