Effective use: Newspapers

Managing and presenting digital newspapers with CONTENTdm

OCLC
The world's libraries. Connected.

CONTENTdm® Digital Collection Management Software has many features and optional settings to help you manage your digital newspapers.

When planning to present your digitized newspapers, there are a number of important things to consider at the outset. Before you digitize and process them, whether from film or originals, you'll need to decide the best way to present them for your user community. Knowing how your users will search and browse is key. You will select the practices for scanning and text processing that optimize your collections and best serve the searching needs of your users. CONTENTdm has many features and optional settings to facilitate your management and presentation of digital newspapers. We've described some of them in this tip sheet, including:

- Using facets to help users refine searches
- Changing the default user view of results sets
- Searching by date—entering and displaying dates
- Searching newspapers—metadata and full text

Using facets to help users refine results

Facets display on Results pages in the far-left column. Facets enable users to narrow their results by the fields that you choose to display as facet options. Configuring your website to enable users to narrow results by the Date facet is useful for newspaper collections.

For more information about facets and configuring your website, see Results Pages.

Searching by date, entering and displaying dates

You must designate the date field in the collection as a Date (ISO compliant) data type before loading of the digital files. Likewise, you must designate the field as searchable.

Dates are converted to yyyy-mm-dd format in a date data type field when that item is saved. The date is then always stored in that format regardless of how it is displayed. (In the Project Client and in CONTENTdm Administration, dates will be displayed in the yyyy-mm-dd format.)

Searching newspapers—metadata and full text

End-users can search the metadata supplied at the newspaper issue level, as well as at the page level. Some metadata may be automatically extracted and generated at the time of scanning; other metadata may be entered manually after the collection is built.

A minimum set of metadata fields for newspapers might include:

Title (of series)
Edition (volume, number)
Source (digital file)
Geographic coverage (of originally published newspaper)
ISO date (of issue)
Type (form of original)
LCCN
Date (of digitization)

Full-text scanning and preprocessing for newspapers
End-users can also search on the full text found in articles, across pages and across collections on a CONTENTdm Website. As part of the preparation process for building digital collections, the original paper or microfilmed version will be digitally scanned and processed to extract the text for searching.

For additional information

Entering Dates

Using OCR

About CONTENTdm Flex Loader
Efficiently batch import large quantities of XML data using this CONTENTdm Add On.

File naming and organizing

If compound objects are to be created, then there are several CONTENTdm wizards that enable building them in batches. In this case, the pages will be alpha-numerically sorted, and can have descriptive titles created from the file names, if you have them named according to the CONTENTdm display convention. It is important to specify the file naming conventions and organization with your digitization department or vendor to make sure that you can take advantage of this automatic feature.

For example, files organized and named:

Daily Herald, April 1, 1900 (folder name)
0001_Title page.jpg
0002_Masthead. Jpg
0003_Page 1.jpg

Will appear in the navigation pane as:

Daily Herald, April 1, 1900
Title page
Masthead
Page 1

Processing with optical character recognition

The accuracy of optical character recognition (OCR) can vary widely depending upon the quality of the source images. The National Digital Newspaper Program offers advice on specifying scanning resolution, bit-depth, etc. The Library of Congress sponsors this program and technical guidelines for scanning and OCR processes may be found at: http://www.loc.gov/ndnp/guidelines/digitizing.html

CONTENTdm can ingest TIFF, JPEG2000 or JPEG images and offers an integrated OCR Extension that derives the full text of the newspaper page for searching and display. This process generates ASCII characters and bounding box coordinates for each word so that a user's search terms will display in highlights in the image, the page description, and the Text view. (Even without the use of the OCR Extension, the item description and Text views will still highlight the term if a transcript has been supplied through other means.)

If you choose to use a digitization vendor instead of processing the digital images in-house, you may want to specify delivery of files that have been pre-processed to also contain article-level coordinates. When displaying files created with such segmentation software, CONTENTdm will not only highlight the retrieved text on the image, but the articles also will be automatically selectable.

Note: Article segmentation display support is only available in the CONTENTdm 6.x Website. The CONTENTdm Resonsive Website does not support article segmentation display.