

CONTENTdm 4.3 Features

CONTENTdm 4.3 includes the following changes and new features. More information about Version 4.3, including updated help files, download kits for upgrading, and a recording of the CONTENTdm 4.3 Web sessions, is available on the User Support Center.

PDF

CONTENTdm 4.3 includes significant improvements for working with PDF files that make processing born-digital documents quicker and easier. Additionally, new display options provide more flexibility for end-users when viewing and printing PDF files.

Thumbnail generation — Automatic thumbnail generation from PDF files is now supported. Thumbnail images are created from PDF files when items are approved and added to the server.

Text extraction — Improved text extraction now converts all extracted text to UTF-8 encoding. Text embedded in PDF files is extracted and added to the full text search field when the PDF is approved and added to the server.

Search term highlighting — Search terms are now highlighted in PDF files displayed in the item, compound object, and page & text viewers.

Inline display — PDF files are displayed inline in the compound object and page & text viewers for both PC and Macintosh browsers. Collection administrators have the option to display PDF files inline in the item viewer by editing the S_SHOW_ITEMVIEW_PDF variable in the global style file.

Compound object conversion — Multiple-page PDF files are automatically converted into compound objects. This feature can be enabled per collection by editing configuration settings under the Collections tab in CONTENTdm Administration. When compound object conversion is enabled and a multiple-page PDF file is added to a collection, CONTENTdm:

- Creates a compound object based on the page order of the PDF file.
- Generates a page-level metadata record for each page.
- Extracts text from the PDF file, converts it to UTF-8, and inserts it into the full text field of the associated page-level record.
- Generates thumbnail images of each page of the PDF file. The thumbnail image of the first page is used for the compound object.
- Retains the original PDF file for export and printing.

After conversion, the PDF compound object is displayed in a compound object viewer with each page of the PDF accessible from the left navigation menu. Search terms are highlighted in the PDF and end users have the option to access the complete PDF for printing or saving.

A **subset of print version** view option enables end users to select a subset of pages from the PDF to view, save, or print. To generate the subset of the PDF, the end user can select all pages with search hits or pick individual pages or page ranges. The **subset of print version** view option is also available for non-PDF compound objects when they have been processed using the OCR Extension.

Additionally, pages of a compound object automatically generated from a PDF file do not count towards the total number of items on the server.

Large file download — If a PDF file is over 20 MB, it does not load inline in any of the viewers but is downloaded instead. A download prompt displays giving the user the option to save or

open the file. This allows the user to download or open the file in the background while continuing to work within the browser. In the compound object and page & text viewers, the download prompt is triggered when the individual page being accessed is over 20 MB; in the subset viewer, the download prompt is triggered when the selected subset of pages is over 20 MB.

More information: www.contentdm.com/help4/acq-station/importing2b.html
www.contentdm.com/USC/tutorials/using-pdfs.pdf (available November 1, 2007)

Conversion scripts — Two command line conversion scripts are available to update PDF files in existing collections. The pdfprocesscollection script converts existing multiple-page PDF files that are not part of a compound object to compound objects. Text is extracted and thumbnail images are generated during the conversion for all PDF files in the existing collection. The pdfcollection script extracts text and generates thumbnail images for all PDF files in an existing collection but does not perform compound object conversion. Pointers for all PDF items in the collection will remain the same when the command line scripts are run. This means that the reference URL for all PDF items remains the same after conversion.

More information is available in the 4.3 Update Guides: www.contentdm.com/USC/guides/index.asp

Metadata

Administrative Fields — Users can now view and configure the properties of six administrative metadata fields allowing them greater control of the metadata that can be searched and viewed within their collections. The six configurable administrative fields are:

- Full resolution
- OCLC number
- Date created
- Date modified
- CONTENTdm number
- CONTENTdm file name

Each field can be designated as searchable and mapped to Dublin Core. Field names can be changed and exposed in the collection interface. Additionally, content in the Full resolution and OCLC number fields can be edited.

More information: www.contentdm.com/help4/collection-admin/fields2.html

Controlled Vocabulary — Controlled vocabularies can now be shared between fields both within the same collection and across different collections, eliminating the need to maintain duplicate versions of the same vocabulary for separate fields. This new feature enables collection administrators to designate any controlled vocabulary as a shared controlled vocabulary. It can then be selected for use in other fields and edited from any field using it.

More information: www.contentdm.com/help4/collection-admin/fields3.html

Collection templates — Two additional metadata templates are available when creating a new collection. A Qualified Dublin Core template generates a collection with Qualified Dublin Core metadata fields pre-defined and mapped for cross-collection searching. A VRA Core 3.0 template generates a collection with the VRA Core 3.0 fields pre-defined and mapped to Dublin Core.

Full text search field — Each collection is now restricted to one full text search field. If a collection already has a full text search field designated, a second one cannot be created. This

aligns collection field properties with the full text search field display options, which only support a single full text search field.

Server

Server Administration — Several user interface changes have been made to CONTENTdm Administration to make it easier to access and configure server and collection settings. Settings can now be edited from the administration pages and no longer require users to access text files stored on the server to make changes.

- **OAI** — A new interface for configuring OAI settings has been added to the Server tab. Server administrators can enable OAI harvesting for the entire server or select specific collections, specify whether compound object pages should be included in the harvest, and define the repository name. More information: www.contentdm.com/help4/server-admin/oai.html
- **Stop List** — The stop list used to prevent common words from being hyperlinked in the metadata displayed in the item and compound object viewers can now be edited by using CONTENTdm Administration. An interface for viewing, adding, and deleting stop words is now accessible to server administrators under the Server tab. More information: www.contentdm.com/help4/custom/stopwords3.html
- **Viewer Settings** — New interfaces for configuring the item and object viewers, at both the server and collection level, have been added to CONTENTdm Administration. Administrators can turn on and off the image toolbar, define default display sizes for all viewers, and select colors for word and navigational thumbnail highlighting. Server administrators can define viewer settings server-wide under the Server tab, and collection administrators can define settings for specific collections under the Collections tab. More information: www.contentdm.com/help4/custom/zoompan.html
- **About Page** — The CONTENTdm license code can now be edited from the administration About page under the Server tab, making it easier to input new license codes when upgrading your license level.

Additionally, full resolution settings have been moved under Configuration in the Collections tab, and new links to access collection configuration have been added to the Collections page.

Thumbnails — The quality of thumbnail images generated when items are added to a collection using the Web-based Add page has been improved. Additionally, the display of thumbnail images in the browse and results pages is now more consistent. The aspect ratio of all thumbnail images is retained within a defined width and height so that non-standard thumbnail images now display true to their size.

Custom Web pages — A new method for creating custom Web pages is now supported. This method enables new pages to be created and stored outside of the */cdm4* directory without requiring any edits to the config.php page to recognize them.

More information: www.contentdm.com/help4/custom/custompages.html

Connexion Digital Import

This new feature enables catalogers using the OCLC Connexion client to add digital items to their CONTENTdm collections during the Connexion cataloging process. Connexion catalogers can

attach a digital file (or files, if more than one digital file is associated with the same record) to their WorldCat record by using the Connexion client. The digital file is then added to the designated CONTENTdm collection along with its associated metadata record. Metadata records are mapped to Qualified Dublin Core when added to the CONTENTdm collection, TIFF images are converted to JPEG2000, and OCR is performed for TIFF images comprising compound objects. PDF processing occurs if enabled in the CONTENTdm collection. Additionally, the OCLC number from the WorldCat record is added to the new OCLC number administrative field in CONTENTdm.

Digital items added to CONTENTdm using the Connexion client are discoverable from FirstSearch, WorldCat.org and WorldCat Local. Each digital item added to CONTENTdm using the Connexion client is associated with its WorldCat record via a persistent URL based on the OCLC number (WorldCat accession number) of the WorldCat record. The persistent URL of the form **www.worldcat.org/oclc/<oclc#>/viewonline** now provides one-click access to digital content.

By integrating digital collection creation with standard cataloging workflows, Connexion digital import provides an additional option for organizations looking to expand participation in growing and maintaining their digital collections. The Connexion digital import requires both an OCLC Cataloging subscription and the use of CONTENTdm Hosting Services.

More information: www.contentdm.com/USC/tutorials/connexion.pdf (available November 1, 2007)