How It Works

The DocuShare crawler is a custom Java servlet running inside the Tomcat servlet container. On startup it launches two separate threads: a sender thread and a receiver thread.

The sender thread searches for candidate PDF files which lack embedded text layers. When found, those PDF files are sent to the PDF generator via a watched folder. The PDF generator can be either ABBYY Recognition Server (now called ABBYY FineReader Server), or ABBYY FineReader Corporate edition. Because of the limited performance and page count allowed for processing by watched folders (typically 5000 pages per month) by FineReader Corporate edition, Recognition Server 4 or FineReader Server 14 is the recommended PDF conversion platform.

The receiver thread watches for PDF files returning from the PDF generator. When found, those files are checked into DocuShare as new versions of the original PDFs.

The send and receiver threads have no direct communication with each other, and operate asynchronously from each other.

Sender Thread Details

How does the sender thread find candidate PDF files? It searches the DocuShare "DSObject_table" table directly, via SQL.

The sender is given its search parameters from two different sources: the crawler's global properties (stored in the "CFDSCrawler.xml" config file), and a sender profile. There must be at least one sender profile, but there can be several. The sender thread iterates over all of the profiles, searching with each one separately (if the profile has enabled=true). The profiles are stored in individual XML files stored in a location specified in the CFDSCrawler.xml config file.

Using the database connection information from the CFDSCrawler.xml config file, the sender thread uses the information from each profile to search for candidate PDFs. The profile specifies:

1. the "root" of the search (all candidates must have the specified collection as an ancestor),
2. the type of DocuShare objects (default is "Document") to search,
3. the timestamp of the last search of this profile (all candidates must be modified on or after that timestamp).

In addition to those search criteria, the sender thread will also test that:

4. the candidate has not been deleted,
5. the candidate has not been previously rendered by the crawler,
6. the candidate's MIME type is PDF ("application/pdf"), and
7. the candidate's Abstract is empty, which implies there is no text layer on the PDF.

There may also be custom SQL "where clause" components specified in the profile. If present, those where components are added to the profile's search.

The SQL query is constructed to order the candidates in "last modified" timestamp order. This is done to minimize the expense of subsequent searches of candidates (only recently added or modified candidates will be checked). It also has the useful side effect of handling throttling by the conversion service: if the conversion service is too busy, the search can be aborted (recording the timestamp of the last checked candidate in the profile). By recording the timestamp in the profile, the next time the profile is used for a search, the previously examined candidate can be skipped.

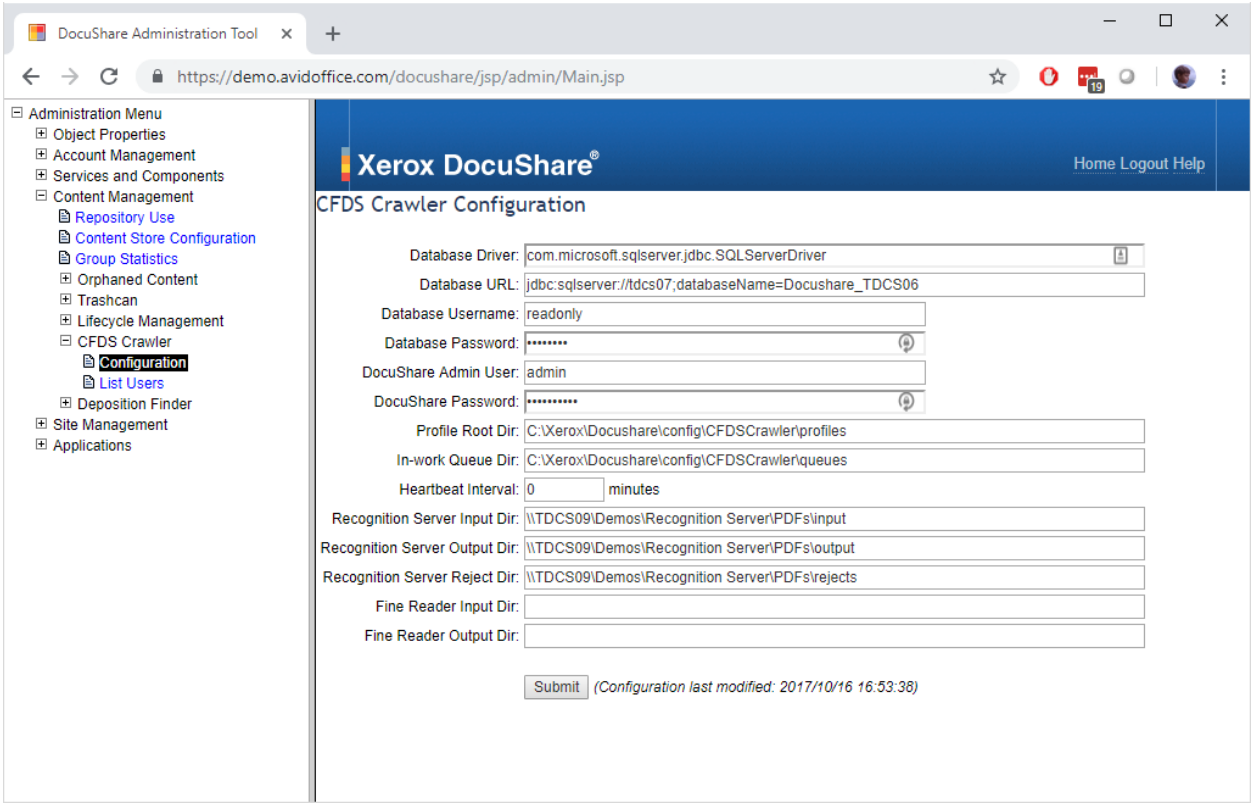
Once a candidate has been qualified, the sender extracts the PDF content into a file in the PDF generator input folder (defined in the CFDSCrwaler.xml config file). The name of the PDF file is the handle of the document (with ".pdf" as the extension).

In addition, the sender creates a zero-length marker file in the “queues” folder. This file servers as an indication that we are expecting a fie back from the PDF generator for that document.

Receiver Thread Details

Compared to the sender thread, the receiver thread’s duties are vastly simpler. On every wakeup, it scans the queue folder for expected returning files. It then looks in the PDF generator’s output folder for any of those files. If found, then the file’s name (which is the document’s handle) is used to find the document. The generated PDF is checked in as a new version of the document. The new content has a revision comment indicating that the PDF was generated by the crawler.

Initial configuration



The database driver is always Microsoft’s: com.microsoft.sqlserver.jdbc.SQLServerDriver

This value is case-sensitive.

The Database URL is the information to reach the DocuShare server’s SQL database. It is always in the form:

Jdbc:sqlserver://*serverName*;databaseName=*sqlDatabaseName*

Where *serverName* is the IP name of the machine hosting SQL Server, and *sqlDatabaseName* is the name of the SQL Server database on that server.

The “Database Username” and “Database Password” are the SQL Server login credentials to use to connect to the database. All of the database actions performed using this credential are queries, so the login credentials only need readonly access to the database. The password is stored in an encrypted form in the XM configuration file.

The “DocuShare Admin User” and “DocuShare Password” are the DocuShare login credentials used to access the existing version and check-in the new version of the PDFs processed. It must be a credential

within the “Docushare” domain, and it should have full access rights to the potential candidate PDFs. The password is stored in an encrypted form in the XM configuration file.

The “Profile Root Dir” is a directory on the DocuShare server where the crawler’s profile XML files will be stored.

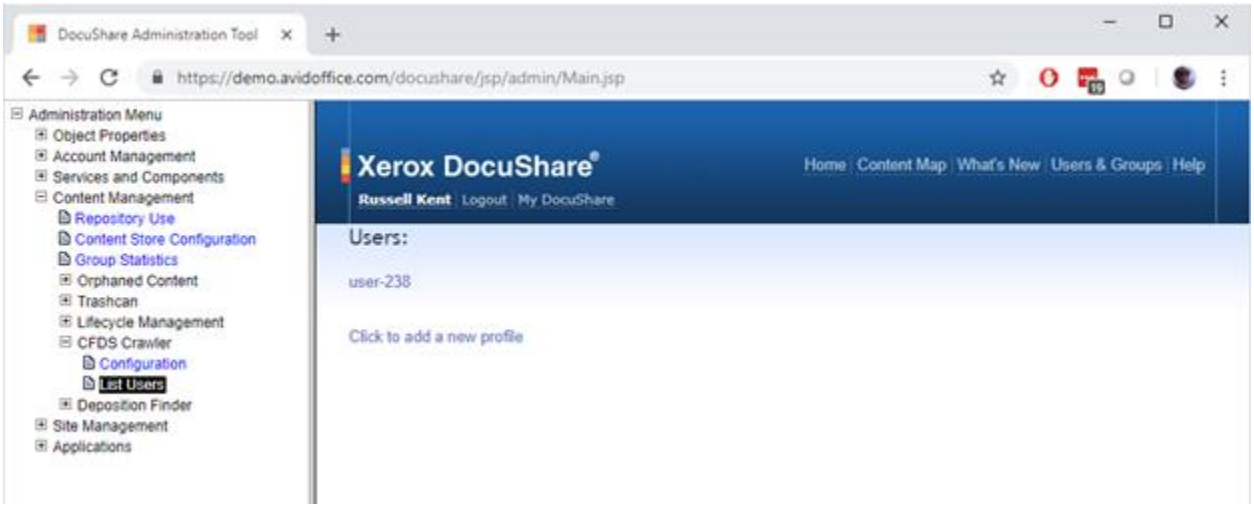
The “In-work Queue Dir” is a directory on the DocuShare server the crawler’s zero-length marker files are created. These marker files record what PDFs have been sent to the PDF generator program (ABBYY Recognition Server or ABBYY FineReader Corporate).

The “Heartbeat Interval” is the number of minutes between executions of the sender and receiver threads. If set to 0 (zero) then the crawler is disabled. This value is only checked when DocuShare is started.

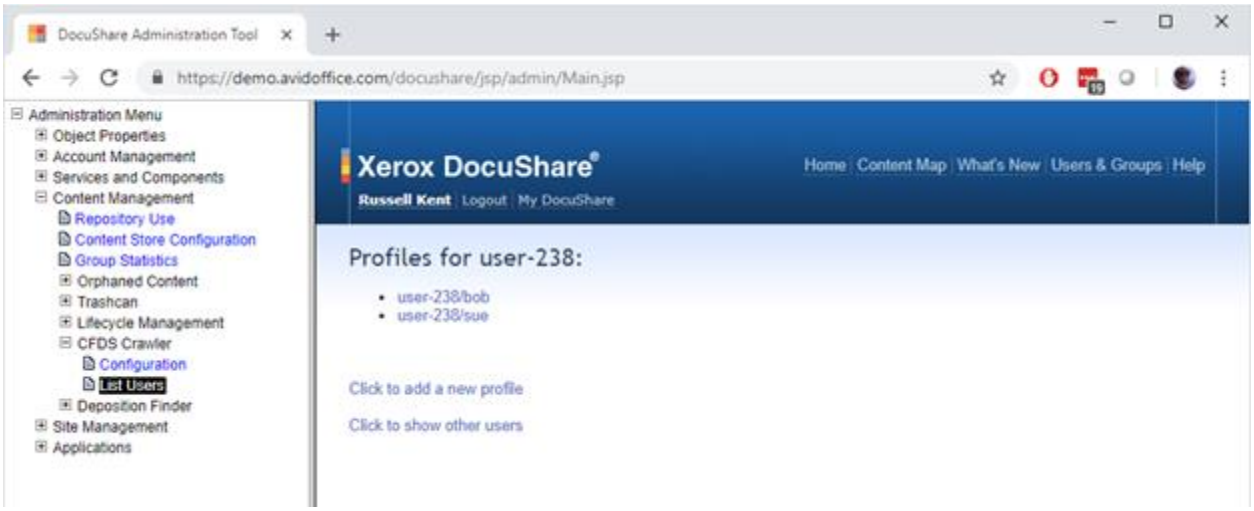
The “Recognition Server Input”, “Output”, and “Reject” are the directories used to send work to the ABBYY Recognition Server. “Input” means from the crawler to the Recognition Server; “output” means from the Recognition Server back to the crawler.

The “Fine Reader Input” and “Output” are the directories used to send work to the ABBYY FineReader Server. “Input” means from the crawler to the FineReader Server; “output” means from the FineReader Server back to the crawler.

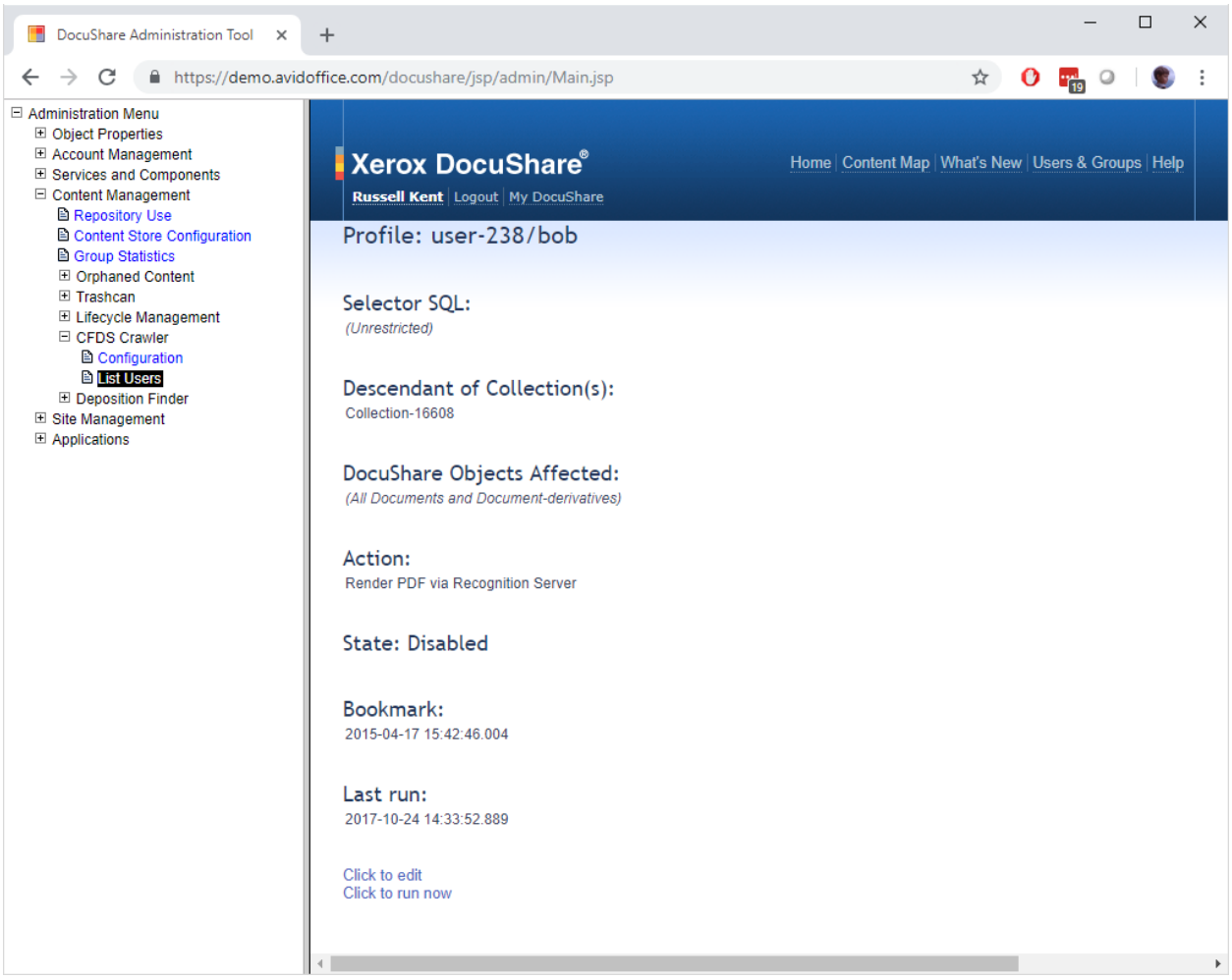
Setting up Profiles



Profiles are stored under the current user’s handle within the profile root dir. (This is an artifact of an early design choice.) To create a new profile, select “Click to add a new profile”. Otherwise, click on the user whose profiles you wish to view.



Bob's profile would appear as …



The "Selector SQL" is the custom "where clause" described in the "Sender Thread Details" above. You should probably leave this blank unless told to put something in it by Criteria First.

The “Descendants of Collection(s)” is a comma-separated list of collection handles that are the root(s) of the tree to be searched.

“DocuShare Objects Affected” is a comma-separated list of DocuShare objects that are to be searched. If blank, then all “document” objects and all objects derived from “document” are searched. Only document derivatives may be listed.

“Action” specifies whether discovered candidates should be sent to Recognition Server or to FineReader Server. If other PDF generators are added in the future, then this list may change.

“State” is a Boolean flag indicating whether the Sender thread should process this profile.

“Bookmark” is the “last modified” timestamp of the most recently examined object. It is used so that the Sender thread can interrupt and resume its search.

“Last run” is the timestamp of the last time this profile was examined by the Sender thread.

Installation, Support and Maintenance

Criteria First provides installation, support, and maintenance for the DocuShare OCR Crawler. Installation is available via professional services and takes about two hours to install plus about two hours to install and configure either ABBYY FineReader or FineReader Server to work with it. Your IT staff is welcome to install and configure the ABBYY component. Please call or write for advice regarding configuration or other concerns.

Support: 972-492-4428

support@criteriafirst.com