Tag Extractor For JGC

Installation, Configuration & Use

Revision - 31 August 2016

Table of Contents

Overview

The Tag Extractor is designed to extract Tags or Part identification numbers from documents.

This is accomplished by leveraging Adlib PDF to normalize all input PDF's to text-searchable PDF, while also extracting text data to an XML-based PDFInfo file.

Some text may be vertically oriented and testing has proven that rotating a page 90 degrees clockwise and counter-clockwise allows Adlib PDF to extract vertically ascending as well as descending text horizontally for a higher degree of accuracy.

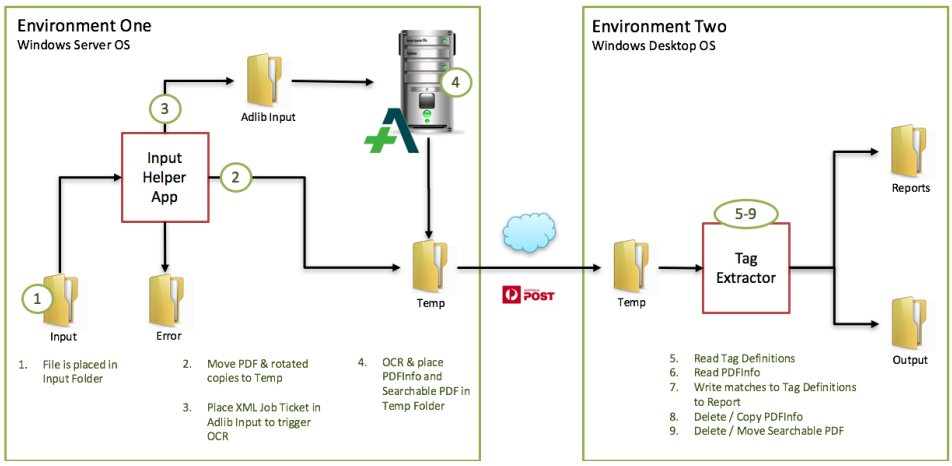

Figure 1 - Solution Overview

To accomplish this, an Input Helper App monitors the folder where PDF documents to be analysed are placed. Upon placement of a PDF, the PDF is moved as well as 2 copies, one rotated clockwise and the other rotated counter-clockwise into the Temp folder.

Next, Adlib PDF will apply OCR when necessary, and then extract the text from the document. A searchable PDF and PDFInfo (xml) file are created and placed in the Temp folder.

The Temp folder is monitored by the Tag Extractor process.

Upon launch, the Tag Extractor process will read the Tag Definitions file. If changes are made to the tag definitions file, the Tag Extractor process should be re-started.

Upon placement of an XML PDFInfo file, the Tag Extractor process will check all text records for matches to tag definitions as defined in the Tag Definitions File.

Matches are determined by comparing each text record against each Tag Definition which includes a Tag Name, Tag Group and Regex pattern.

As matches are found, they are written to a Tag Index report as Comma Separated Values including:

1. File Name
2. Tag Group Name

3. Tag Name
4. Orientation
5. Page Number
6. Tag Value

Installation

System Requirements

The Tag Extractor was designed and tested to work with Adlib PDF Enterprise 5.4, and should work with any version of Adlib PDF Enterprise with Major version 5.x.

Please see the Adlib PDF Enterprise documentation for your version for the system requirements and pre-requisites of Adlib PDF.

The Tag Extractor requires only a Windows OS that supports and also has installed the .NET framework 4.0 or higher, client profile as a minimum.

It is also important to note that the user account under which the Tag Extractor will operate must have read and write access to the folders defined in the application configuration.

Deployment

To deploy the Tag Extractor, simply un-zip the Tag Extractor bundle to the root Adlib folder on the target machine.

This is typically C:\Adlib\, and may be shared as \\localhost\Adlib\

Deploying to a remote environment

If the Tag Extractor is to be deployed to an environment that is not local to an Adlib PDF server, then there will not be a root Adlib folder.

In this event, simply create an Adlib folder somewhere on the target machine, and un-zip the Tag Extractor bundle to that folder.

By sharing the new Adlib folder as 'Adlib' on the local machine, you may be able to avoid reconfiguring a number of settings.

Configuration

Application Configuration Settings

In the \Bin\ folder you will find the executable as well as a file titled app.config. By modifying this file, you can change the configuration of the application.

The Tag Extractor must be re-started after any changes are saved to the app.config file.

App.config contains the following settings:

ScanInterval
This is an integer and indicates the number of milliseconds between each scan of the Input and Temp folders.

Default value is 5000.

TextMode
This defines the level of granularity for text records.
Valid values include: *Adjacent*, *SingleSpace*, *Line* and *Word*.

Default value is *Line*. This is recommended as this will create fewer records resulting in higher performance when analysing PDFInfo files.

FolderToMonitor
This defines the path to monitor for Input PDF documents. Any PDF documents that are located in or placed into this folder will be moved to the temp file once detected.

If the setting RetainOriginalPDF is not 'True', input PDF documents will be deleted upon completion of analysis of the PDFInfo file.

AdlibInputFolder
This defines the path that is monitored by an Adlib PDF Enterprise Folder Connector. This folder must support the processing of XML Job Tickets.

TempFolder
This defines the folder in which to store all temporary files used by the Tag Extractor. This includes the original PDF, rotated copies of the original PDF, a searchable PDF and a PDFInfo file for the original and rotated PDF source documents.

When the Tag Extractor process has completed processing the PDFInfo file, the Originals (Including Rotations), Searchable PDF and PDFInfo files will be deleted or moved to the Reports folder depending on the configuration of settings as listed below.

TagDefinitionsFile
This defines the path to the Tag Definitions File. The specification for this file is in this document.

ReportsFolder
This defines the path to the root of the output from the Tag Extractor. Tag Index reports as well as folders containing retained PDFInfo, Searchable PDF's, Original PDF's and XML Job Tickets will be found here as well.

ErrorFolder
In the event of a failure with picking up and creating rotations of each source PDF, the PDF in error is placed in this folder, as is a log file indicating any errors that prevented successful operation.

RetainXMLJobTicket
This is a Boolean, true or false value. If set to true, the XML Job Ticket that is used to trigger the OCR process by Adlib PDF is retained, and a copy stored in the Reports Folder.

RetainSearchablePDF
This is a Boolean, true or false value. If set to true, the Searchable PDF that is created by Adlib PDF is retained, and can be found in a subdirectory in the Reports Folder.

RetainPDFInfo
This is a Boolean, true or false value. If set to true, the XML PDFInfo file that is created by Adlib PDF is retained, and can be found in a subdirectory in the Reports Folder.

This option is useful if you wish to re-run the Tag Extractor after modifying the Tag Definitions without having to re-OCR all of the documents.

RetainOriginalPDF
This is a Boolean, true or false value. If set to true, the original PDF as well as the rotated copies are retained and can be found in a subdirectory in the Reports Folder.

ReportSizeLimitMB
This is an integer in Megabytes for the maximum size of the Report file. Once the report file exceeds this maximum size, it will be renamed to include the date and time, and a new report file will be started.

Tag Definition File

This file is a tab-delimited file where each line is a record, and each record contains the following fields:

1. Tag Group Name
2. Tag Name
3. RegEx Pattern

The order in which the tags are assessed against each document is determined by the order in which they appear in the Tag Definition File.

Longer, more complicated tags should be higher in the list to prevent shorter tags from matching on subsets of larger tags.

Additionally, more specific tags, such as those with limited character rangs or have a specific character expected in the tag should be higher in the list so that specific matches can be made before more generic patterns are assessed.

XML Job Ticket Template

This is a plain-text file that is mostly XML, but contains some variable place-holders that are replaced at run-time by the Input Monitor.

You can use the XML Job Ticket to change a number of aspects of how the job will be processed to get a different behaviour or output in the PDFInfo and Searchable PDF, such as target PDF version, metadata, etc.

Please refer to the Adlib XML Job Ticket Guide with the Adlib PDF documentation for specific elements and attributes that are valid for your version.

Usage

Launching the Process

You will find the executable in the folder \bin\ along with the app.config and XML Job Ticket Templates.

To launch the process, simply open a command prompt, navigate to the Adlib\Tag Extractor\Bin folder and type TagExtractor.exe

Alternatively, you can double-click the TagExtractor.exe icon from the \bin\ folder in Windows Explorer.

Command Line Arguments

There are 2 optional command line arguments, used to support running the process in 2 phases or steps:

-noinput

Use this command line argument as 'TagExtractor.exe –noinput' to turn off the input folder monitoring feature. This allows for just the execution of the TagExtractor which will monitor the defined Temp folder and create the Tag Index report files.

A shortcut is located in the \bin folder called 'Process PDFInfo' that will launch TagExtractor.exe with -noinput option.

-noextract

Use this command line argument as 'TagExtractor.exe –noextract' to turn off monitoring of the temp folder for PDFInfo files. This allows for just the execution of the Input Folder monitor to detect incoming PDF's and generate the rotated PDF's as well as the Adlib XML Job Tickets.

This would be used as the 1st phase in a 2-phase operation and is useful if you want to create PDFInfo files and Searchable PDF's to be used at a later time or a different physical location.

Stopping the Process

Press any key in the command window, or if you wish to stop the process remotely, place an empty (0-byte file) in the Input directory with the filename stop.

If the –noinput argument has been used, place the stop file in the temp folder, or press any key in the command window.

Please note that any single file currently being processed will be completed prior to the application shutting down.