Autobahn DX Cookbook

Aquaforest

Version 1.1 March 2016

Table of Contents

1   Dealing with Large Data sets

1.1   Requirement
Convert a large data repository to searchable PDF.

1.2   Background
Whether processing documents in a single folder or a large tree structure Autobahn will copy the contents of the source location to the work folders where the conversion is carried out. Once all of the files have been processed they are then moved to the defined target folder.

This approach generally works well however it can become challenging for users when a failure occurs midway during the conversion or for those users who would like access to the searchable PDF files quickly as the output file will only become available when the whole conversion has completed.

1.3   Solution
Use batch processing mode using the "Include with Document Count Limit" option with the scheduler to process a X number of files every X seconds or minutes.

In the screen capture below we have chosen the step "OCR Image-Only PDF" we have also applied the filter to only pick up the first 10 files with the extension .PDF (*.PDF ; 10)

We need to utilise the option "Move to Archiving after Processing" to ensure we don't repeatedly process the same 10 files.
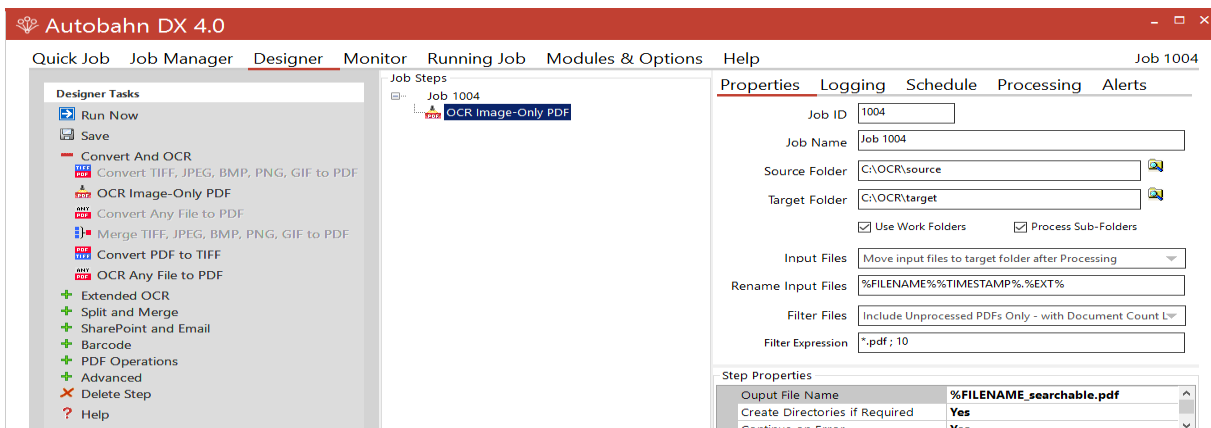
The job has been configured to run continuously, it will start to run every 40 seconds, if the job is already running the service will attempt to run the job after another 40 seconds.
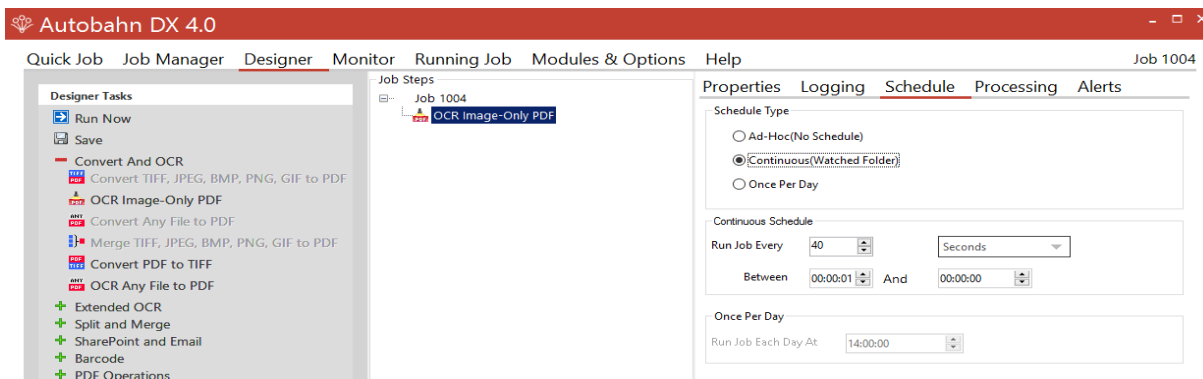
2   Processing a "hot folder"

2.1   Requirement
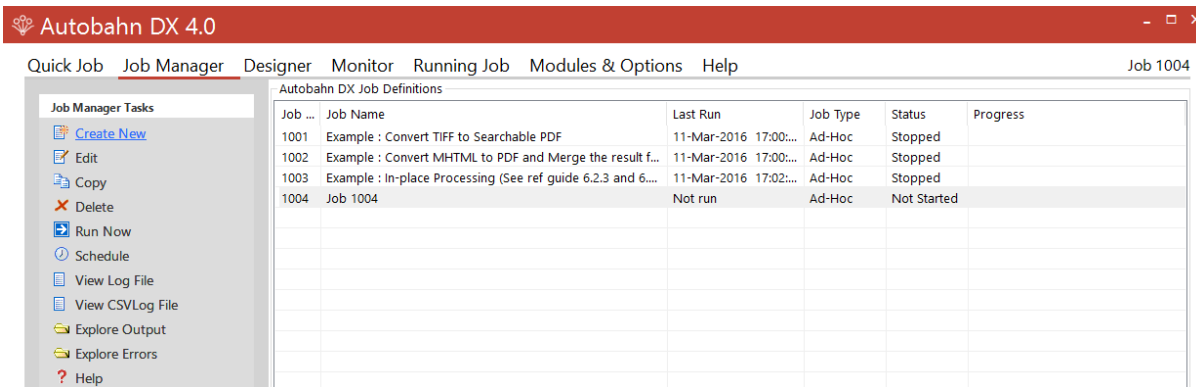
Create and configure a job to process a "hot folder" which may contain different file types that need to be converted to searchable PDF.

2.2   Solution

Select "Create New" in the "Job Manger" tab:

In the "Designer" tab select the following step:

- "OCR Any File to PDF"
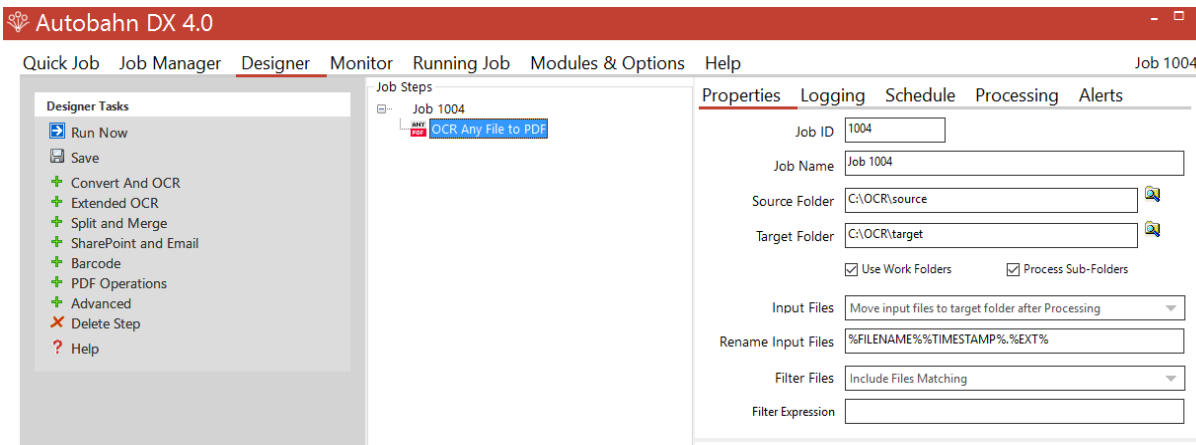
2.3   Comments

This is job which will convert all supported file types to searchable PDFs. You will need to select the value "move to archiving after processing" or "delete input files" for the "input files" option as shown in the screen capture above. This ensures the same input files are not repeatedly processed.

The step by default will add the suffix "_searchable" to the output PDF filename, if you do not require this option then you can highlight the step and modify the "Output File Name" option from: %FILENAME_searchable.pdf to %FILENAME.pdf

3 Supported File Types for "Convert Any File to PDF"

The following files types can currently be converted to PDF using the step "Convert Any File to PDF"

Word

Excel

PowerPoint

Visio

Publisher

Outlook MSG files with attachments

Outlook Express

Open Office formats

Notepad

HTML

Raster images (BMP, GIF, JPEG, PNG, TIFF, WMF)

4   Languages Supported by the Aquaforest OCR engine

English
German
French
Russian
Swedish
Spanish
Italian
Russian English
Ukrainian
Serbian
Croatian
Polish
Danish
Portuguese
Dutch
Czech
Roman
Hungarian
Bulgarian
Slovenian
Latvian
Lithuanian
Estonian
Turkish

5   Languages Supported by the Extended OCR engine

|			
English (American)	Lithuanian	Ganda	Balinese
German	Afrikaans	Hani	Bikol
French	Albanian	Ido	Ilocano
Spanish	Catalan	Interlingua	Madurese
Italian	Irish_Gaelic	Kicongo	Waray
British	Scottish_Gaelic	Kinyarwanda	None
Swedish	Basque	Malagasy	Serbian_Latin
Danish	Breton	Maori	Latin
Norwegian	Corsican	Mayan	Latvian
Dutch	Frisian	Minangkabau	Hebrew
Portuguese	Nynorsk	Nahuatl	Numeric
Brazilian	Indonesian	Nyanja	Esperanto
Galician	Malay	Rundi	Maltese
Icelandic	Swahili	Samoan	Zulu
Greek	Tagalog	Shona	Afaan
Czech	Japanese	Somali	Asturian
Hungarian	Korean	Sotho	AzeriLatin
Polish	Schinese	Sundanese	Luba
Romanian	Tchinese	Tahitian	Papamiento
Slovak	Quecha	Tonga	Tatar
Croatian	Aymara	Tswana	Turkmen
Serbian	Faroese	Wolof	Welsh
Slovenian	Friulian	Xhosa	Mexican
Luxemb	Greenlandic	Zapotec	BosnianLatin
Finnish	Haitian_Creole	Javanese	BosnianCyrillic
Turkish	Rhaeto_Roman	Pidgin_Nigeria	Moldovan
Russian	Sardinian	Occitan	SwissGerman
Byelorussian	Kurdish	Manx	Tetum
Ukrainian	Cebuano	Tok_Pisin	Kazakh
Macedonian	Bemba	Bislama	MongolianCyrillic
Bulgarian	Chamorro	Hiligaynon	UzbekLatin
Estonian	Fijan	Kapampangan	Balinese

6   OCRing Image PDFs "In-Place"
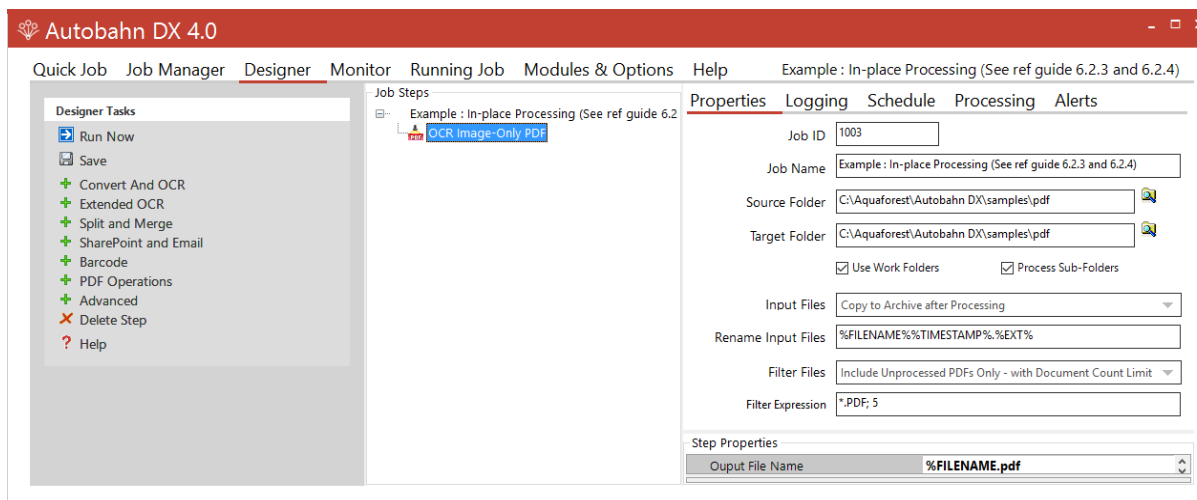
6.1   Requirement

A job to handle a folder which contains PDF files that require in-place OCRing, processing 5 unprocessed PDF files at a time.

6.2   Solution

As this is a common requirement we have pre-configured job 1003, Select job 1003 from the "Job Manger" tab:
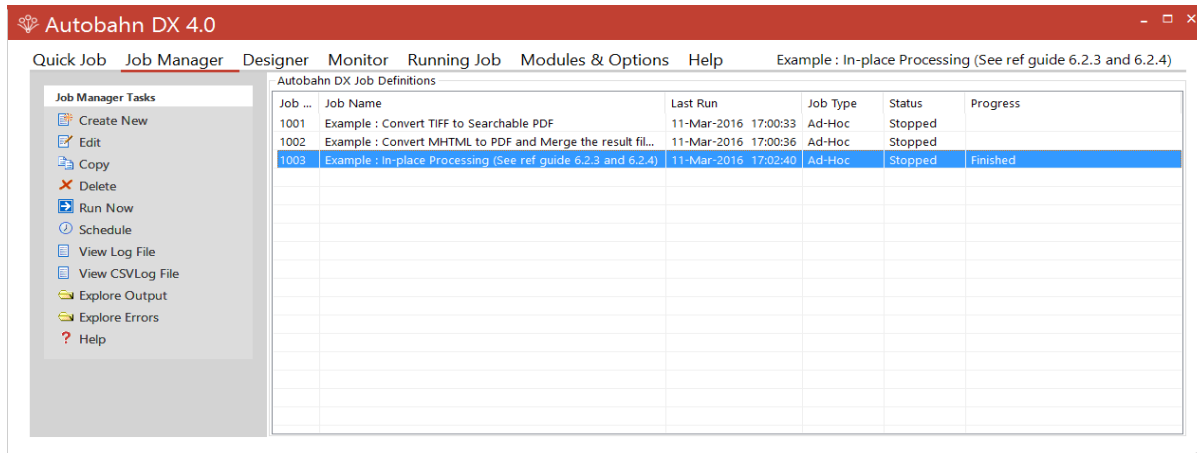
- Modify the source and target folders to that of your own.

6.3   Comments

The screen capture above shows a configuration which will process the PDF files in-place (Native) and processing 5 unprocessed PDF files at a time. The conversion will place a searchable copy of the original PDF file in the source location as well as placing a copy of the original input PDF file in the archive location.

You should also select to "Move" "Password Files" and to "Move" "Files in Error" this will stop them from being processed repeatedly.
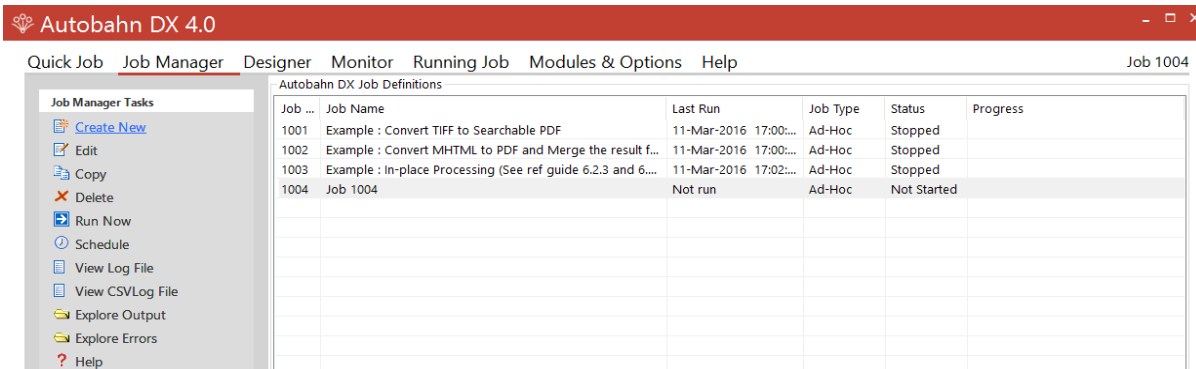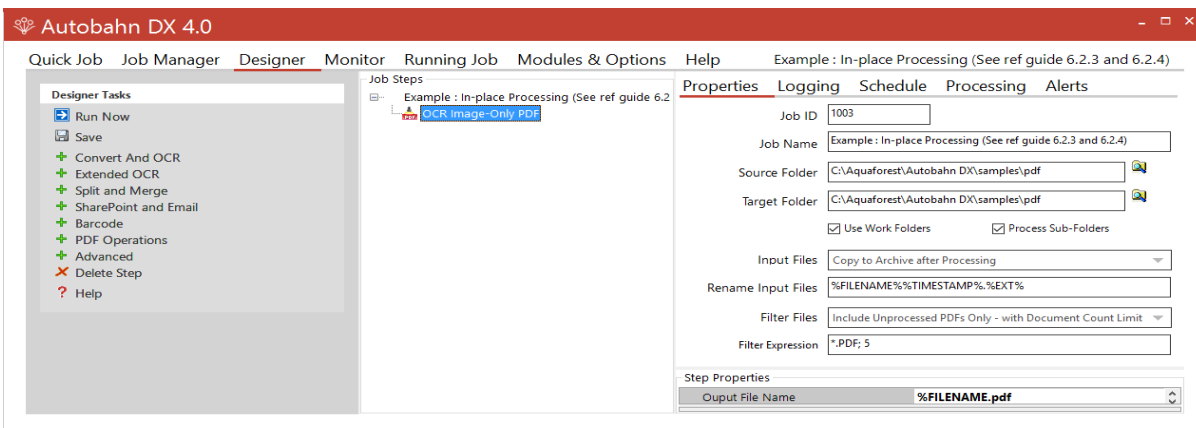
7 Multiple CPU Cores

7.1 Requirement

Configure a job to utilise Multiple CPU Cores.
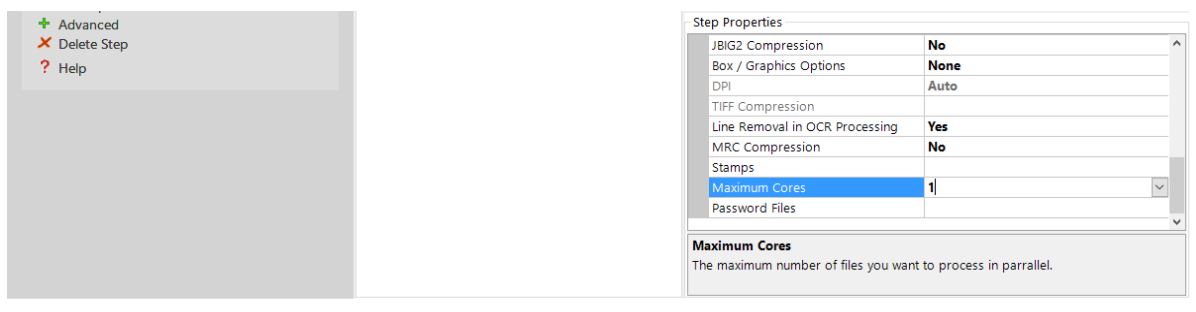
7.2 Solution

Select "Create New" in the "Job Manger" tab:

In the "Designer" tab add the step which you require, highlight the step.

Scroll to the end of the "Step Properties" section as shown in the screen capture below and use the drop down for Maximum Cores to select the number of cores you would like to utilise.

7.3 Comments

The cores will be dependent on your license and the number of physical cores on the host machine.
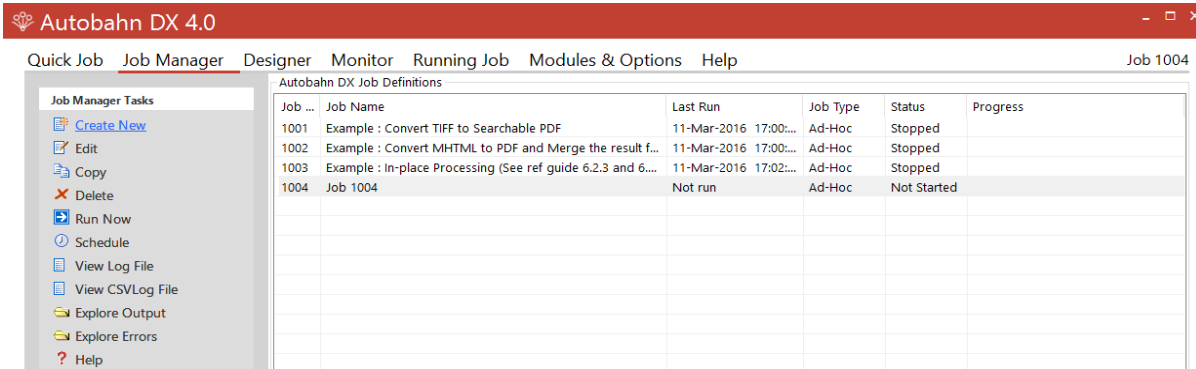
8 Using Extended OCR

8.1 Requirements

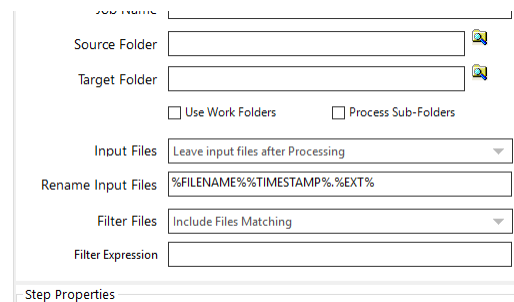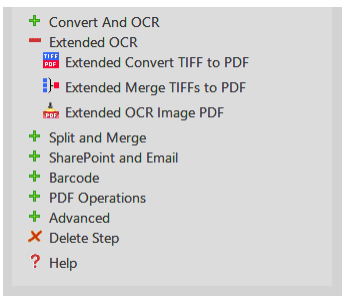Create a job to use the Extended OCR.

8.2 Solution

Select "Create New" in the "Job Manger" tab.

From the "Designer" tab select one of the following Extended Steps:

- Extended Convert TIFF to PDF
- Extended Merge TIFFs to PDF
- Extended OCR Image PDF

8.3 Comments

Ensure you have the Extended OCR license, you can check via the "Modules and licensing" tab.

9   Using the Autobahn DX API

9.1   Requirements
Programmatically execute or create jobs in Autobahn.

9.2   Solution
An API is provided which allows a .Net application to allow execution of existing job definitions and creation of "on the fly" new jobs from XML Job ticket files. The API DLL is located in the AUTOBAHN\bin folder and a sample Visual Studio 2008 project is provided in the AUTOBAHN \code examples\AutobahnAPIExample folder where AUTOBAHN is the folder where Autobahn DX has been installed – by default this is C:\Aquaforest\Autobahn DX

11.1   API Example

```
using System;
using System.Threading;
using Aquaforest.Autobahn.API;

namespace ConsoleApplication2
{
        class Class1
        {
                [STAThread]
                static void Main(string[] args)
                {
                        IJob job = null;
                        string JobFinished;
                        string Jobstatus;
                        try
                        {
                                /* Run Job and Wait to Complete */

                                job = new Job(1001);
                                job.Start();
                                job.WaitForExit();
                                JobFinished=job.JobFinished();
                                Jobstatus=job.Jobstatus();

                                if(Jobstatus=="Error")
                                {
                                        Console.WriteLine("Job Failed");
                                }
                                else if(Jobstatus=="Stopped")
                                {
                                        if(JobFinished=="True")
                                                Console.WriteLine("Job Completed");
                                        else
                                                Console.WriteLine("Job              Was
Terminated");
                                }
                                job.Dispose();
```

9.3   Comments
Further information can be found in section 11 of the Autobahn DX reference guide.

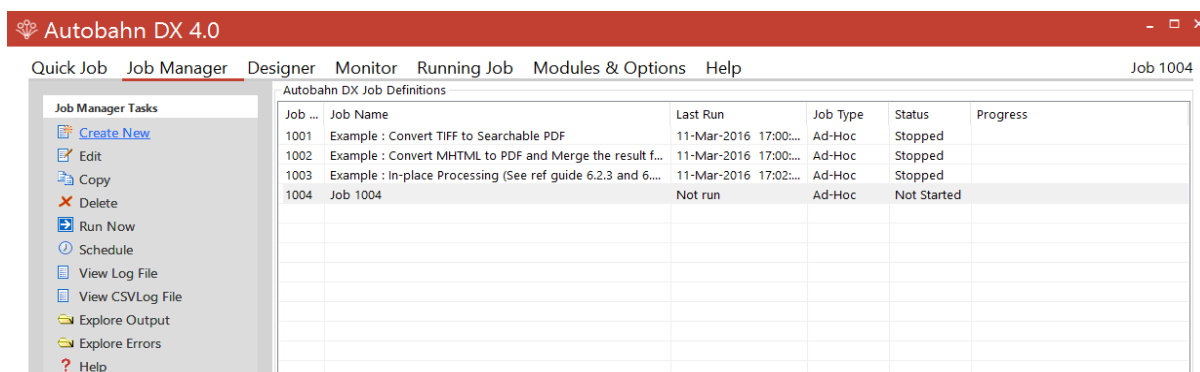# 10 Using Stamps on output PDFs

10.1 Requirements
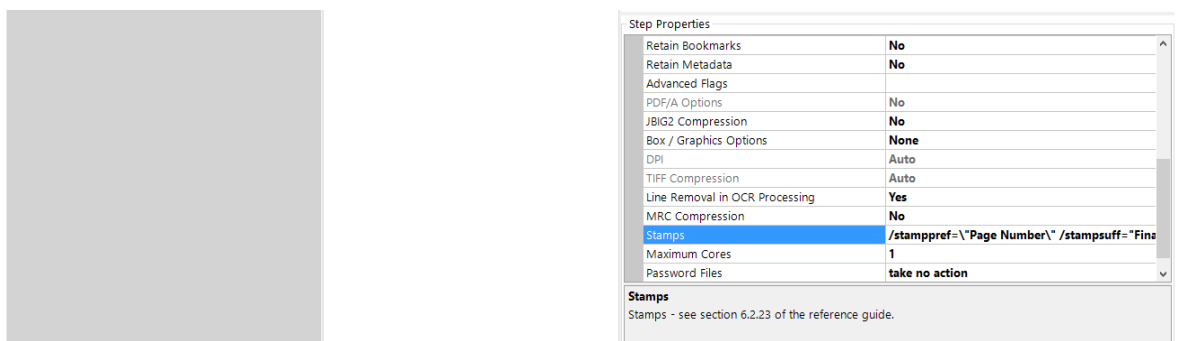
Apply a stamp to the output PDF.

10.2 Solution

This feature is only supported when selecting one of the following steps:

- "OCR Image-Only PDF"
- "TIFF to PDF"

Select "Create New" in the "Job Manger" tab and select one of the above steps.

Scroll to the end of the "Step Properties" and enter your stamp specification in the Stamp field as shown in the screen capture below.

10.3 Comments

In the above example we have entered the following stamp criteria:

/stamppref=\"Page Number\" /stampsuff="Final" /stampstart=123 /stampdigits=6 /stamppos=0 /stamptype=1

This will generate the following stamp on the top left hand side of each page in the form of an image.

`Page Number000123Final`

Further information can be found in section 6.2.23 of the Autobahn DX reference guide.

11 Using MRC or iHQC Compression during OCR

11.1 Requirements

Compress output PDF files during OCR.

11.2 Solution

Compressing output PDF files can be achieved using Mixed Raster Compression (MRC) when using standard job steps or IRIS' Intelligent High Quality Compression (iHQC) when using the Extended IRIS job steps. The following steps support the above compression types.

- "OCR Image-Only PDF"
- "TIFF to PDF"
- "Extended OCR Image PDF"
- "Extended Convert TIFF to PDF"

Select "Create New" in the "Job Manger" tab, select one of the above step which you require. Scroll down to the Step Properties section, enable "MRC" for standard steps using the drop down as shown below.
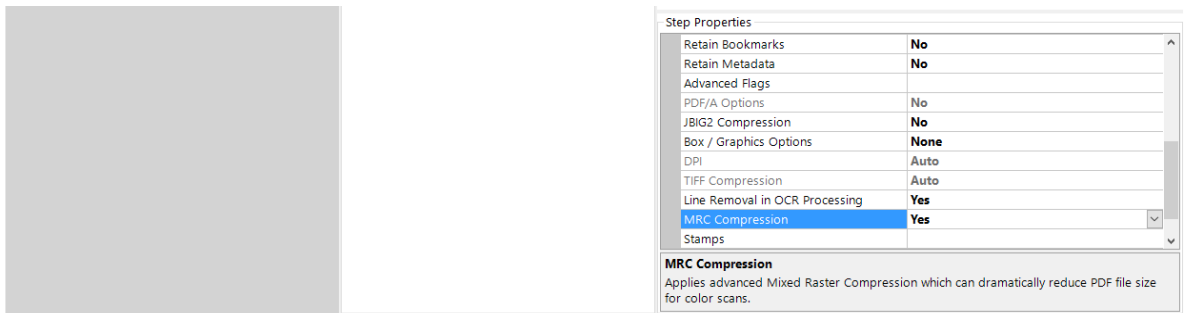
For Extended job steps scroll down to the Step Properties section and enable "iHQC"" using the drop down as shown below.
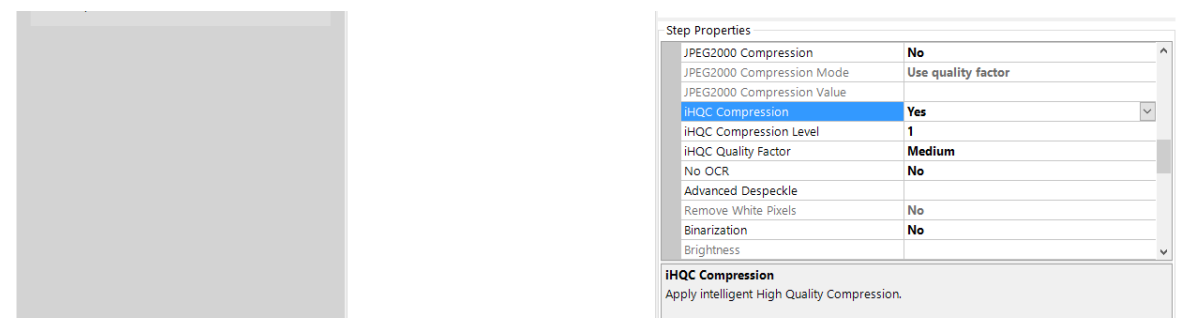
11.3 Comments

Note that the compression options cannot be used in conjunction with the "Native" Image Method for OCRing PDFs as the "Native" method does not adjust the images within the PDF.
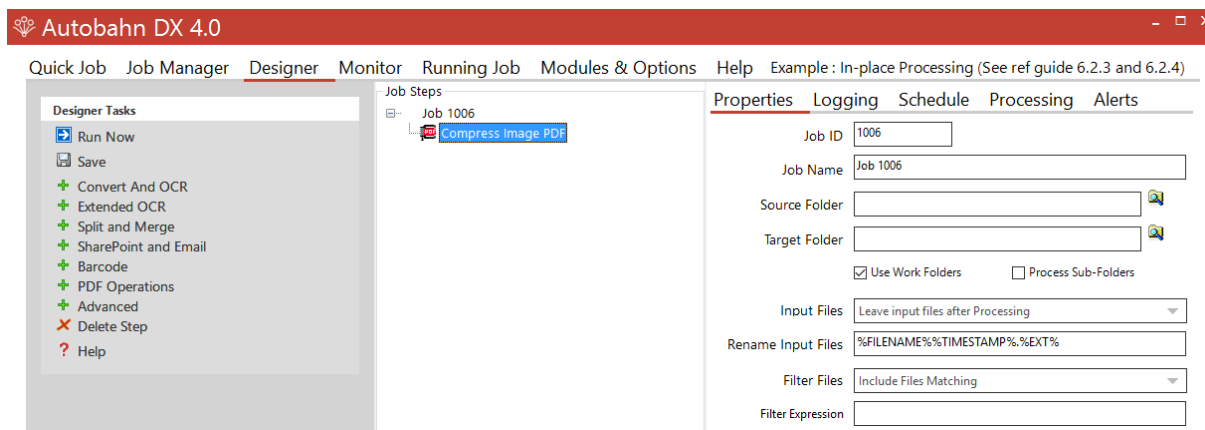
12 Compress Image PDF

12.1 Requirements

Compress Image PDF files without performing OCR.

12.2 Solution

Compressing Image PDF files without OCRing can also be achieved using Mixed Raster Compression (MRC). The following step supports this:
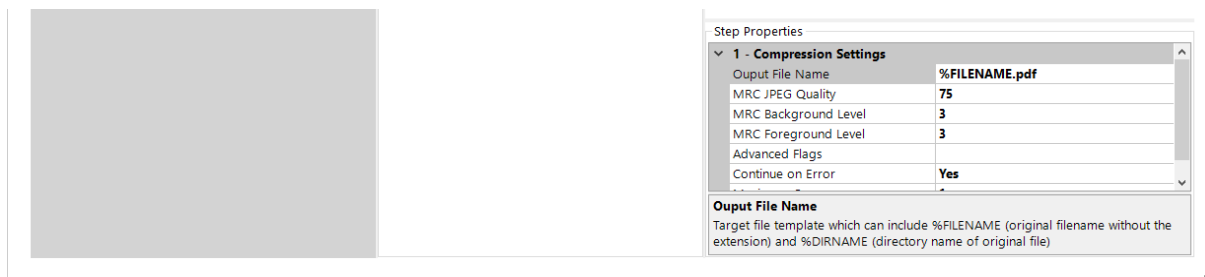
- "Compress Image PDF"

Select "Create New" in the "Job Manger" tab, and select the above step.

12.3 Comments

The default values shown in the screen capture below can be modified to achieve greater compression however this will compromise the quality of the output PDF.

Further information can be found in section 6.2.24 of the Autobahn DX reference guide.

13 Handling Job Errors

13.1 Requirements
Ensure continuity of a conversion and/or notify when job is in Error.

13.2 Solution
There are a couple of options which can be configured to handle document or job errors, these errors can occur:

 a) During a conversion.
 b) Or when a job finishes/stops prematurely and is set to Error.

To handle document errors which can occur during a conversion each step has the option "Continue on Error", by default this is enabled (set to "Yes") as show in the screen capture below. This enables the conversion to continue processing the remainder of the documents if a document(s) during the conversion generates an error. Documents which generate an error are by default routed to the error folder unless configured differently.
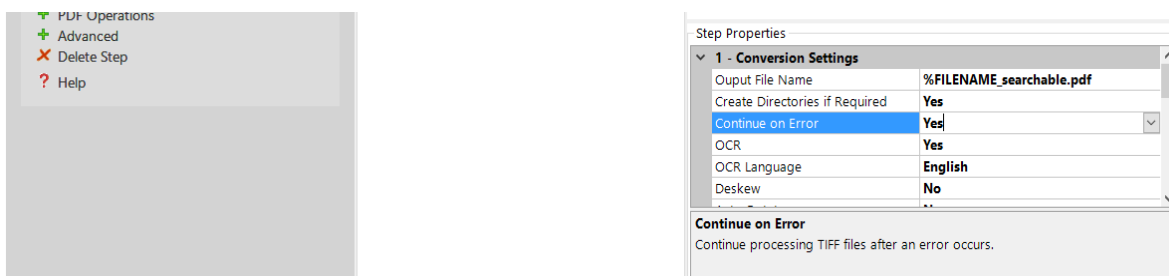
In more serious cases a job may be set to Error during a conversion or after it completes the conversion. In the Job "Alerts" tab there is an option "Stop Processing on Error" which is job specific and stops the job from running again once it is set to Error. This by default is enabled which will stop the job from re-starting when in the Error state. We recommend this to be enabled (left as default) and the "Send Email on Error" notification to be configured. In the case a job has been set to Error we recommend investigating the root cause before restarting the job.

If "Stop Processing on Error" is not enabled the job will try to restart as configured in the Schedule tab, this can be useful if the issue was intermittent i.e. server was shutdown down after updates.
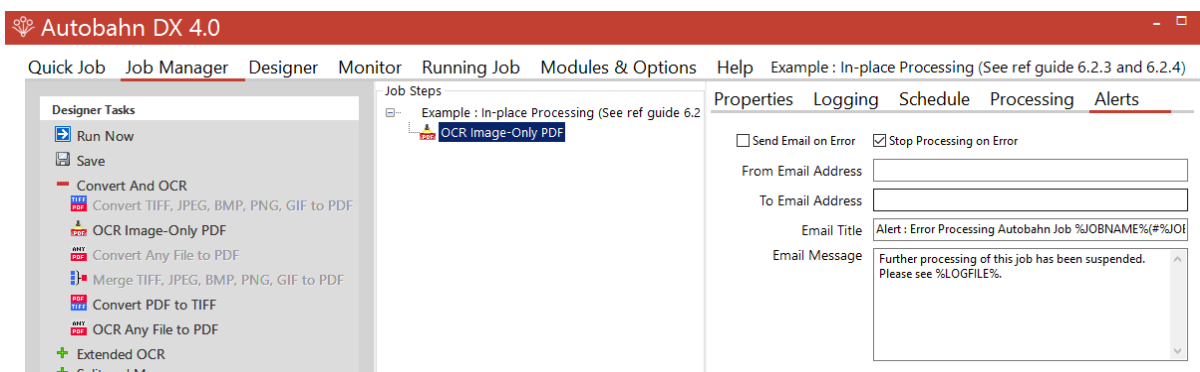
13.3 Comments
If a job has been set to Error take a look at the log file to see what has been logged and in addition take a look at the Windows Event viewer. We also recommend logging a call with the support team.