Acrobat for Legal Professionals

October 27, 2005

Batch OCR using Acrobat Professional

Have you ever received a PDF file that did not contain searchable text? You may know that you can use Acrobat's OCR (Optical Character Recognition) to add an invisible layer of searchable text on top of the file. This allows you to select, copy and search text on a paper document. Great!

What do you do when you have hundreds of TIFFs and Image-only PDFs file that you need to search for a big case? Working with these documents one at a time is not efficient.

If you have Acrobat Professional, you can batch OCR and let you computer do the work for you.

Read on to learn how ...

Batch Processing to the Rescue

There are two steps to follow:

- 1. Set up a Batch Sequence
- 2. Run a Batch Sequence

Set up a Batch Sequence

Scan your documents locally or send to a PC where Acrobat Pro is installed.

If you have the capability, scan directly to PDF or to an MTIFF (multi-page TIFF). These formats allow all of the pages of a document to be maintained as a single file.

- 1. In Acrobat Professional 7, choose Advanced—>Batch Processing
 - or —

In Acrobat Professional 8, choose Advanced—>Document Processing—>Batch Processing

2. Click the New Sequence button.

BLOGS.ADOBE.COM

ABOUT

ARCHIVES

CATEGORIES

NSS / ATOM

SUBSCRIBE TO ACROBAT FOR LEGAL PROFESSIONALS BY EMAIL

DON'T MISS OUR LEGAL EVENTS . . . SIGN UP TO BE NOTIFIED!

LEGAL

The views expressed in this blog are my own and do not necessarily reflect the views of Adobe Systems Incorporated.

Terms of Use

SEARCH

Search this site:

Search

Batch Sequences		
New Sequence Edit Sequence Rename Sequence Delete Sequence	 Embed Page Thumbnails Fast Web View Open All Print 1st Page of All Print All Remove File Attachments Save All as RTF Set Security to No Changes 	
Run Sequence	Close	2

- 3. Give the sequence a name.
- 4. Click Select Commands

Edit Batch Sequence - Batch OC	R	×
1. Select sequence of commands:	Select Commands	R

5. Choose *Recognize Text Using OCR* and click the *Add* button.

🐼 Add Watermark & Background	^	[Recognize Text Using OCR
Description			Primary OCR Languages English (US)
Embed All Page Thumbnails			PDF Output Style: Searchable Image (Exact)
Export All Images As JPEG	1		Downsample: Lowest (600 dpi)
Star Export All Images As JPEG2000		Add >>	A
See Export All Images As PNG			7
Sector All Images As 11PP		<< Remove	
Concerts Restmarks	(B)		
Make Arrescible	-	Constant of the	
Open Options		Move.up	
Print		The second s	
Recognize Text Using OCR		Move Down	
Remove Embedded Page Thumbnails			"
Security	1	Edit	
7 🛄 JavaScript			Click here to change
Execute JavaScript			OCD softings
7 🛄 Page			OCKseungs
🐼 Add Printer Marks	Y		
		1.1	

- 6. Double-click the *Recognize Text using OCR* text (right side of the window) to set OCR Options.
 - Set Downsample Images to 300 dpi. Click $\ensuremath{\mathsf{OK}}$

Recognize Text - Sett	ings	
Primary OCR Language	English (US)	
PDF Output Style	Searchable Image (Exact)	
Downsample Images	Low (300 dpi) 🗸 🗸)
	OK Cancel	<i>e</i>

- 7. Click *OK* again to get back to the main window.
- 8. Click *Output Options*

Output Options
File Naming
Same As Original(s)
Add to Original base Name(s)
Insert Before Insert After
OriginalName (.ext)
Do not overwrite existing files
Save File(s) As: Adobe PDF Files
▼ Fast Web View (PDF only)
PDF Optimizer Settings
OK Cancel

Note:

Output Options allows you specify where the OCR'd files should be written. I suggest writing them to a local drive and copying later to a network store.

- 9. Enable PDF Optimizer and Do not overwrite existing files.
- 10. Click the Settings Button.

eset Settings	
	Make compatible with: Acrobat 5.0 and later
Tourses	Image Settings
Scanned Pages	Color Images :
Fonts Transparency	Downsample: Bicubic Downsampling to 💌 150 pixels/inch. For images above 225 pixels/inch
Discard Objects Clean Up	Compression: JPEG Quality: Medium
	Grayscale Images :
	Downsample: Bicubic Downsampling to V 150 pixels/inch. For images above 225 pixels/inch
	Compression: JPEG VQuality: Medium
	Monochrome Images :
	Downsample: Bicubic Downsampling to 💟 300 pixels/inch. For images above 450 pixels/inch
	Compression: 18162 Quality: Lossy

Adjust the settings to make the smallest possible files, especially for Black and White (monochrome) files:

JBIG2 Lossless is very efficient and preserves the exact appearance of the text.

Consider trying JBIG2 Lossy which causes some visual degradation, but can be up to 70% smaller than JBIG2 Lossless.

- 11. Click OK.
- 12. Give the revised settings a name such as "B&W Lossy".

Run a Batch Sequence

Now, all you need to do is to run the batch sequence.

- 1. Place all the files you wish to process in a single folder on your hard drive.
- In Acrobat Professional 7, choose Advanced—>Batch Processing

 or –
 In Acrobat Professional 8, choose Advanced—>Document Processing—>Batch Processing
- 3. Select the sequence to run
- 4. Click OK
- 5. Select the folder to process
- 6. Click the Select button.
- 7. Select the Output Folder

That's it!

Sit back and enjoy a cup a coffee as Acrobat does the work for you.

Posted by Rick Borstein at 03:23 PM on October 27, 2005

≪ <u>Metadata and PDF</u> <u>Home</u> <u>Understanding "Flavors" of PDF</u> ≫

Comments

marydumont - 06:10 PM on March 12, 2006

I have 1,000's of pages of tiffs, each one is a single sheet. they are currently on CD's and i read them by clicking on each individually. I read your instructions, but seem to be having trouble at the first step: must I download these documents (or upload) them from CD in order to follow your instructions?

Rick's Reply:

It depends on what you want to do. If you want to convert and OCR each TIFF to matching PDFs, you should copy them from the CD to your hard drive first. If you want to create a single PDF, choose Create PDF from Multiple Files from the toolbar and use the Browse button to select the files you want to combine. I'd recommend batches of 500 or so. You will have to OCR them separately.

Rob Nicholls - 04:53 PM on April 10, 2006

This is an awesome tip which has just saved me from making a paralegal pull an all night stint and made my leading client vey happy - many thanks

Stephanie - 10:17 AM on May 17, 2006

I have a situation where I have a scanned in searchable pdf and I want to use ocr so that I can get the text into a word document to do substantial edits. I have Acrobat Professional ver. 7. The document has 168 pages, so I hate to cut and paste one page at a time. Also, I tried doing a save as to a word doc, but it gave a blank word doc.

----- Rick's Reply----

Go to Document-->Recognize text using OCR, then File-->Save As . . . and choose Microsoft Word. There will not be any text unless you OCR. Make sure you choose the All Pages option when you OCR.