Proven Product, Proven Company

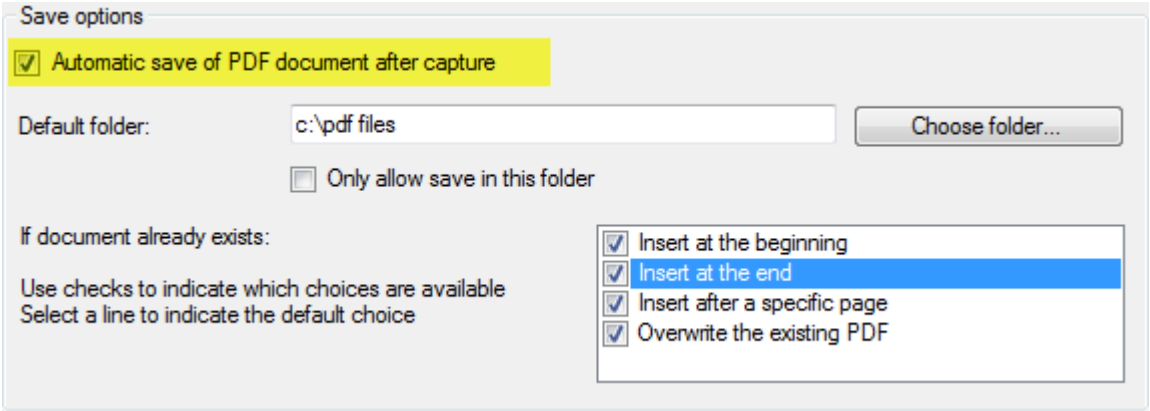
OCR Zone Data Extraction

Setup Guide

Configuring ScanToPDF Settings

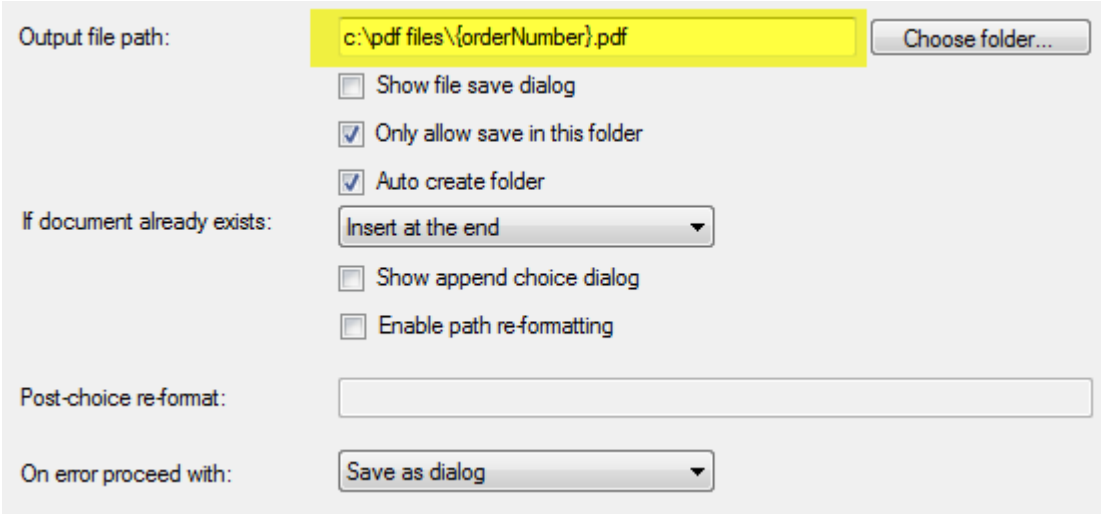
Start ScanToPDF and Click the Edit Menu, then select the “Options” choice from the menu

Click the Save Icon



Make sure the option “Automatic save of PDF document after capture” (*highlighted above in yellow)* is checked

Click the Filenamer Icon

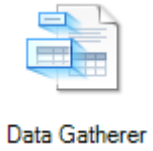


Output file path – Set the output file path (the location and file name for your PDF files). The example setting shows the file being named using the orderNumber value read from the document (ScanToPDF variables are always enclosed in curly braces so we enter {orderNumber}.pdf

On error proceed with – This will depend on the error process required. Choose from one of the three options

1. **Error case** – Name the file using the settings configured in the "Error case file name generation" TAB
2. **Save as dialog** – Display a Save dialog to enable the user to enter a file name
3. **Cancel save** – Do nothing, processing will be suspended

ALL OTHER OPTIONS CAN BE SET AS ABOVE



Making A Template

The Data Gatherer uses a template system to enable data extraction zones to be "drawn" to make zone creation far simpler. This template represents the page that you want to extract data from. Before creating/editing settings you should first create a template. Please follow these instructions.

- Click the Data Gatherer Icon
- Start ScanToPDF
- Click Edit->Options->Data Gather ICON (as shown above)
- Check the "Show 'Save as template' option on page contextual menu" option so it is ticked
- Click OK to close the options dialog
- Scan a document
- The scanned pages will be displayed as thumbnails in the left hand side of the ScanToPDF interface
- Right click the page you want to use as a template to display the context menu
- Choose "Save as template"
- A Save dialog will open to enable you to save the document as a template in the templates folder

Once save is complete you will be asked if this template should be used for current settings, answer "Yes"

Data Gatherer Settings

You must load a template before you can edit settings. If the template and settings are not displayed, select a template from the template list just below the settings grid.

The data gatherer is used to define the zones and create ScanToPDF variables from the data read from those zones. In this example we are extracting 5 zones (Order Number, Make, Model, Year and Colour) from an example Motor Vehicle bill of sale. To Edit Data Gatherer settings, click the Edit button to open up the template editor. The editor has TWO panes; the template image is shown on the left hand side, the settings for the selected zone on the right. The zones are shown on the template window as light blue boxes and the settings for each zone are shown when the zone is clicked. The zone settings for order number are shown below.

Data	
Assignment variable name	**orderNumber**
Data type	String
Zone name	**Order Number**
Extraction	
ActionScript	
CleanupScript	
Coordinates	**X: 1.370856 Y:2.7028 Width1.199499 He**
Data provenance	OCR
Default to empty	False
Exclusion list	**(Collection)**
Multi-value delimitor	
Multi-value grab	False
Regex capture	**(?<orderNumber>[0-9]{5})**
Regex grab syntax	
Regular expression case insensitivity	True
Trigger	
Condition script	
Trigger	PageAnalysis

SETTINGS NOT EXPLAINED HERE ARE NOT USED IN THIS EXAMPLE AND CAN BE DEFAULTED

Assignment variable name – The name of the variable in ScanToPDF

Data type – String, Integer, Decimal or Date (choose the one that matches your data)

Zone Name – The meaningful name for the zone

Extraction – Action Script

Extraction – Cleanup Script

Coordinates – The x, y, width and height of the zone for data extraction

Data Provenance – The data source (**OCR** or Barcode)

Default to empty – *True of False* – If the zone value is not extracted correctly should it default to empty

Exclusion list – Enter a list of values that should not be

Multi-value delimiter

Multi-value grab – *True or False* -

Regex Capture – A Regular Expression used to capture the variable. In this case a named group called orderNumber is used, which is 5 numerical digits long. It is beyond the scope of this document to explain all the features of Regular Expressions but the Internet has many sites dedicated to Regular Expressions (or Regex as it is widely known)

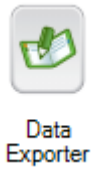
Regex Grab Syntax –

Regular expression case insensitivity – *True or False* – Boolean to indicate whether the Regular expression should be case sensitive

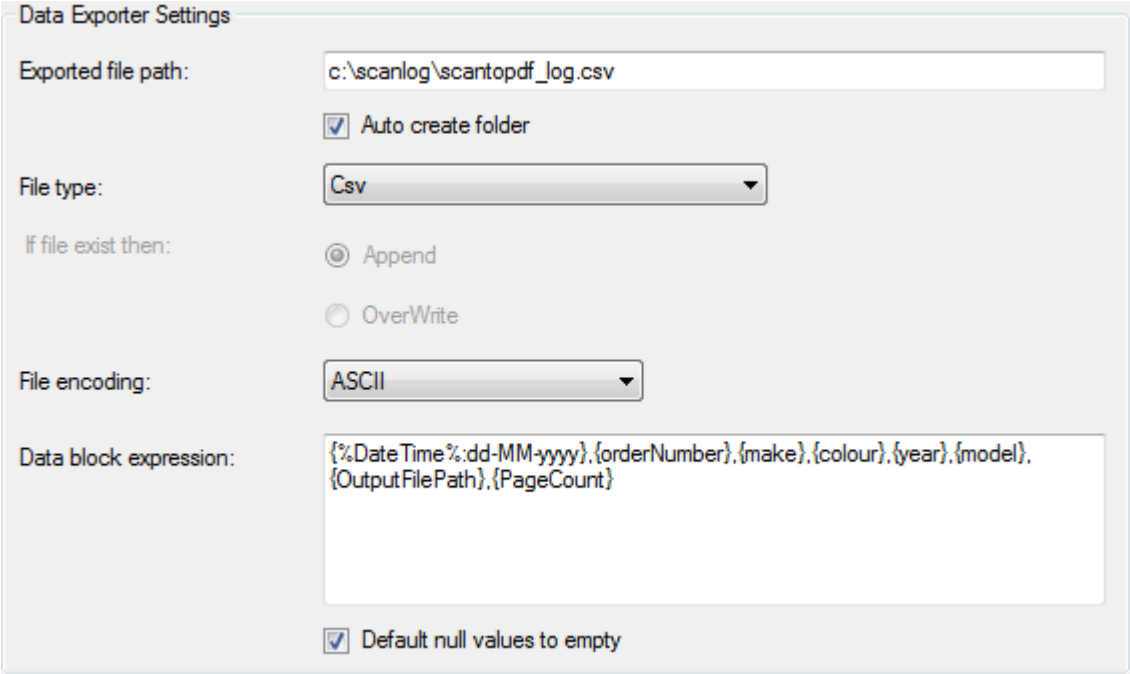
Trigger – Condition Script – Enter a condition script to control whether the zone is extracted, for example PageIdMatch=form_of_authority - Only extract the zone if the PageIdMatch (*from the ADR plug-in*) is form_of_authority

Trigger – Trigger – *Should be left as PageAnalysis (unless advised otherwise by OiC Staff)*

The other zones are configured in a similar way, these can be selected by clicking the zone in the template or choosing the zone from the drop down list at the top of the settings pane.



Click the Data Exporter Icon

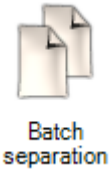


SETTINGS NOT EXPLAINED HERE ARE NOT USED IN THIS EXAMPLE AND CAN BE DEFAULTED

Exported file path – Set this to a path to the log file

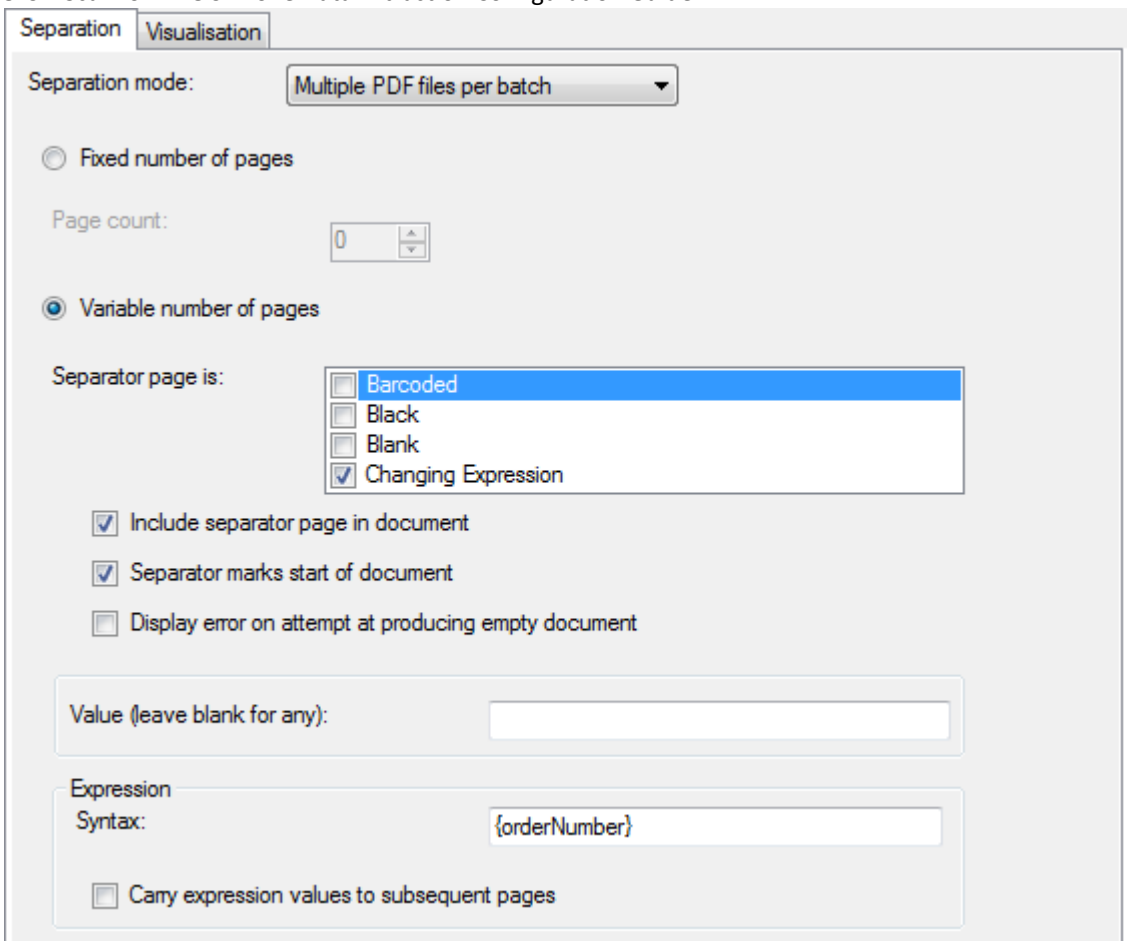
Data block expression – These are the fields which are exported to the log file (ScanToPDF variables). They are exported in this order and will produce a line as follows :

03-03-2011,54321,VOLVO,SILVER,2001,740,c:\pdf files\54321.pdf,1

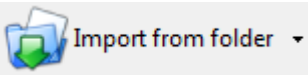


Click the Batch separation icon

In our example, each time the order number changes we want to create a separate PDF. This is done by using the “Changing Expression” setting and setting the variable {orderNumber} as the expression, as shown below.

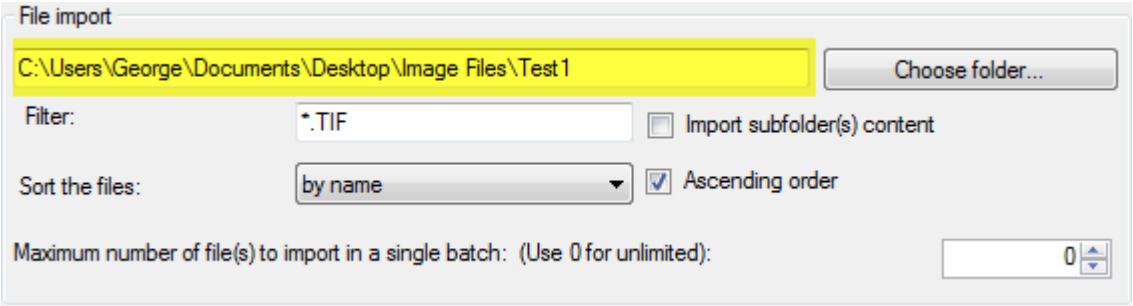


OCR Zone Data Extraction Solution Operation



In the SampleImages folder you will find 3 tiff images which can be used for testing. To use these images as a test

1. Copy the images to a new folder on your system
2. Start ScanToPDF and click Edit-Options-Import
3. Enter the folder you created into step 1 into the folder name in the Import settings



4. Click OK to close the settings screen
5. Click Import from folder on the ScanToPDF toolbar, or use File-Import-Import from folder option
6. The documents will be imported and three PDF files will be created in the output folder, named using the order numbers on the document (12345.pdf, 98765.pdf and 54321.pdf)