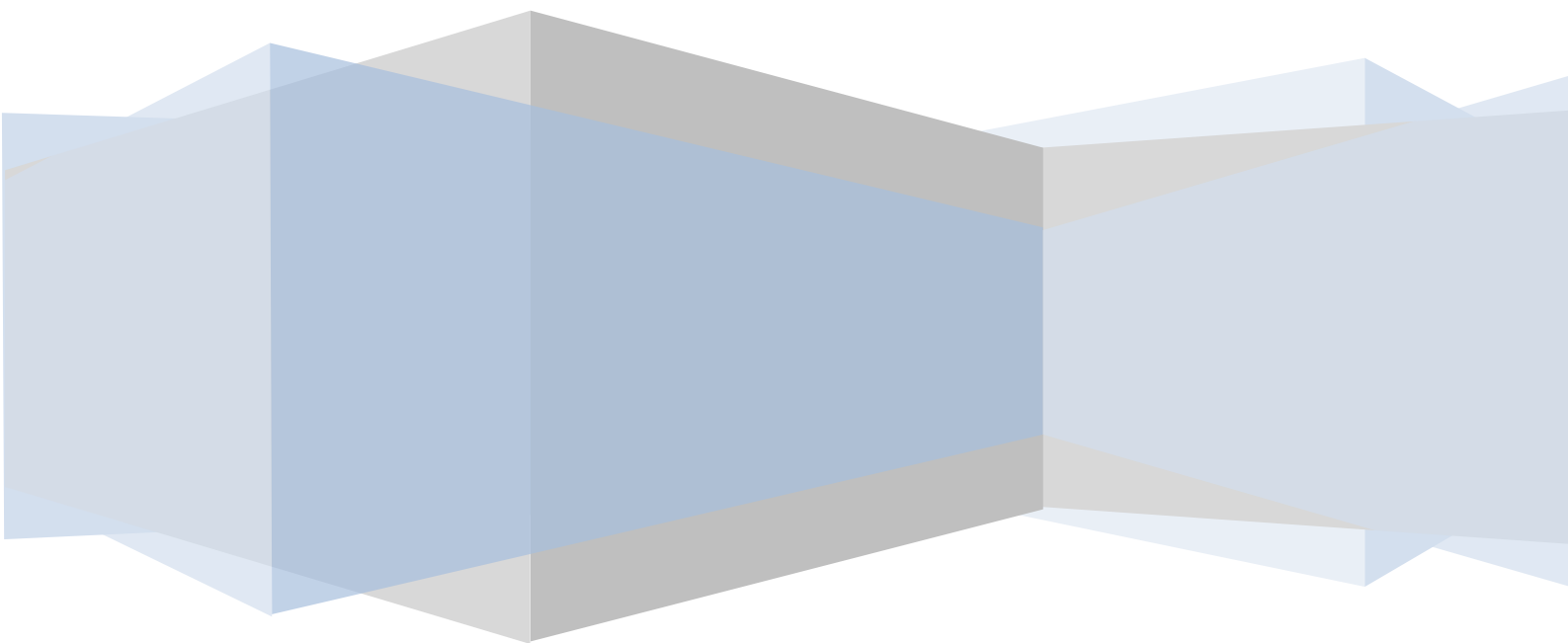




Proven Product, Proven Company

OCR Data Extraction

Setup Guide



PROVEN PRODUCT, PROVEN COMPANY

Thank you very much for choosing to evaluate a ScanToPDF solution. In 95 countries, 35,000 installations use ScanToPDF to convert paper to PDF. Together they scan 1,000,000 pages every day at home and in offices, schools and universities, factories and warehouses.

INTRODUCTION

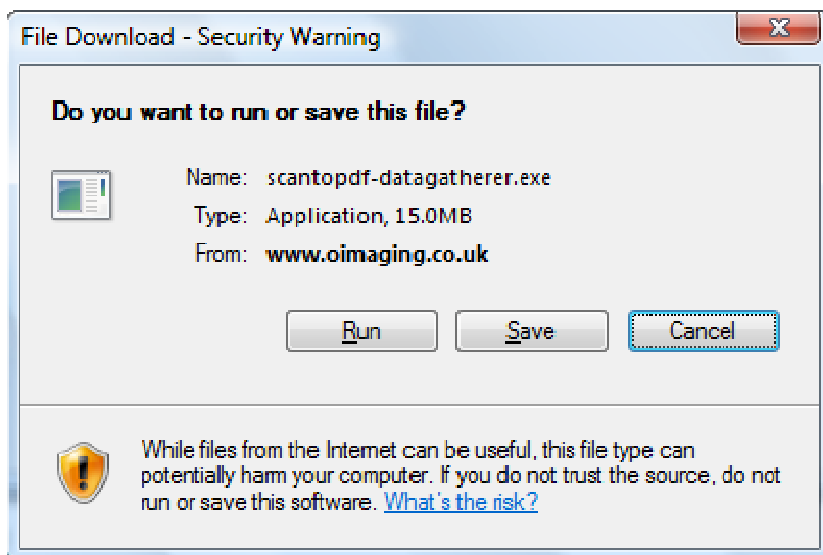
Thank you for evaluating our OCR Data Gatherer Solution. The OCR Data Gatherer solution enables data to be extracted from a scanned page based on zones. This data can then be used in ScanToPDF as variables in file naming or output to a log file for integration with a legacy system. This is a simple example showing extraction of an order number from a delivery note.

HOW TO DOWNLOAD AND INSTALL THE EVALUATION SYSTEM

Click this link

<http://www.oimaging.co.uk/downloads/clientfiles/scantopdf-datagatherer.exe>

The following dialog will appear (may differ slightly depending on browser)



Click **Save** to download the program to your machine. When download is complete locate the program and run it to extract the OCR Data Extraction Evaluation.

IMPORTANT

PLEASE INSTALL TO THE DEFAULT DIRECTORY

(C:\PROGRAM FILES\O IMAGING CORPORATION\SCANTOPDF-DATAGATHERER)

CONFIGURING DATA EXTRACTION ZONES

The data gatherer uses data extracted from the document using OCR. This data is extracted from zones and uses Regular Expression pattern matching to further validate extracted data (which also means the zones do not have to be exactly around just the data you require). ScanToPDF puts the data extracted into variables (which are user configured) for use in other plug-ins such as file naming, log file or batch separation. The zones are configured using a template which has the zones “drawn” onto it using the mouse and a template viewer.

CREATING A TEMPLATE

Scan and save the page required for the template manually as a PDF using ScanToPDF (with OCR enabled), Save the file in any folder (we will refer to this as the template and template folder). PLEASE NOTE TEMPLATES MUST BE PDF FILES CREATED WITH SCANTOPDF AND MUST HAVE OCR (OPTICAL CHARACTER RECOGNITION)

HOW TO CONFIGURE THE REQUIRED SETTINGS



Data Gatherer

All the settings are configured through a single menu

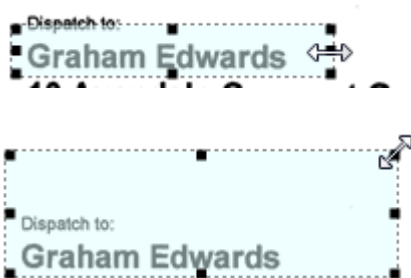
- Start ScanToPDF
- Click Edit
- Click Options
- Click The Data Gatherer Icon
- Click The “Edit” button to open the Data Gatherer configuration screen

LOADING A TEMPLATE

Click the “Load Template” button and select the template file which was created earlier.

EDITING A ZONE

Click the Zones menu and choose Edit. This will enable editing of the values in the right hand property grid and enable any of the zones to be selected and moved using the mouse. Simply click on the zone to show the Cross Hair, click the left hand mouse button and drag the zone to the new position. To change the size of the zone, move the mouse to the edge of the zone until the mouse pointer changes to a right-left arrow (as shown below) and then holding the left hand mouse button size the zone as required.



ADDING A ZONE

Click the Zones menu and choose the “Add” option. Draw a rectangle around the data you want to capture. Once the rectangle is drawn, configure the settings in the left had pane described below.

DATA GATHERER SETTINGS

Data	
Assignment variable name	orderNumber
Data type	String
Zone name	Order Number
Extraction	
Coordinates	X:0.5652477 Y:3.156127 Width:3.274505 Height:0.156127
Data provenance	OCR
Default to empty	False
Exclusion list	(Collection)
Multi-value delimiter	
Multi-value grab	False
Regex capture	\b(?:<orderNumber>[0-9]{3}-[0-9]{7}-[0-9]{7})\b
Regex grab syntax	
Regular expression case insensitivity	True
Trigger	
Condition script	

Data Field Settings

Assignment variable name – The name of the ScanToPDF variable used to store the data extracted

Data Type – The type of data (String, Integer, Decimal or Date)

Zone Name – The name which is used to refer to the zone

Extraction Settings

Co-ordinates – The rectangle that defines the extraction zone, these can be configured by using the template in the editor (described below)

Data Provenance – OCR or Barcode – Default OCR

DefaultToEmpty – Boolean to indicate if the value should default to empty if the zone value cannot be extracted or pattern matched

Exclusion List – Enter a list of values that should not be returned for this zone

Regex Capture – Enter a regular expression used to capture and match the value required

RegexCaptureIgnoreCase – Boolean indicating whether the regular expression is case sensitive

RegexGrab – Can be used to concatenate fields returned from the regex capture for example {orderNumber}{orderDate}

Trigger Settings

ConditionScript – Not Yet Implemented

REMOVING A ZONE

Either

1. Select the zone from the drop down list and click “Remove”
2. Select the zone in the left hand pane and click “Remove”
3. Select the zone and right click and choose “Delete zone”

TEST EXTRACTION

When you have made the required settings you can test the extraction by pressing the “Test Extraction” button. This opens a separate dialog and shows the result of a test extraction. The top panel (Raw Data) shows the raw data extracted from the zone. The Regex Capture panel shows the result of the Regex. applied to the raw data and the Variable Assignment shows the final result of the extraction and the value assigned to the variable. Press Next and Back to navigate through all the variables and press OK to close the dialog. Repeat the process until the variables required are extracted correctly.

USING THE VARIABLES IN OTHER PLUGINS

The variables can be used anywhere in ScanToPDF just by enclosing them in curly braces {}. In this evaluation the Filenamer plugin is using the extracted variable to name the file using the order number.

```
c:\pdf files\{orderNumber}.pdf
```

In filenamer when you press the opening curly brace a list of variables will appear for you to select the correct one. These variables can be used in the data exporter for example to output values to a log file for importing to other systems.

OTHER SETTINGS

All other settings can be set as required by the user and will not affect the data extraction operation.

It should be noted, batch separation can be added to this solution for scanning documents in batches. Please contact OiC for further details.

OPERATION

- ❖ Print the template document and place it in the scanner
- ❖ Start ScanToPDF
- ❖ Click Scan
- ❖ The document will be scanned and saved in c:\PDF Files using the Order Number (15127886) as the PDF name.