blueprism
The Operational Agility Software Company

Interfacing with PDF Documents

Introduction

There are a number of techniques available to extract text from PDF documents using Blue Prism. The techniques available are:

1. Using the Windows Clipboard to copy all the text from a pdf document
2. Using the Blue Prism 'Read Text with OCR' Read stage action to read text from a region within a PDF document
3. Using the Adobe Acrobat API to export the pdf into another format (XML or Microsoft Word) from which data is easier to extract text

If you considering interfacing with PDF documents the following advice should be considered:

- Blue Prism should only be used for extracting data from PDFs if the structure of the documents are standard and predictable with no, or a very limited number of, variations.
- Blue Prism is not a replacement for a dedicated OCR solution designed to extract data from a large volume of different format documents.
- Blue Prism has no functionality for extracting hand written text from a document
- Rather than OCR, an alternative 100% reliable option used by some Blue Prism clients is to retain a reduced number of staff to manually extract data from PDF documents into a structured format which can be used as an input to your Blue Prism solution.

Types of PDF Documents

There are two main types of PDF documents:

PDF Documents

These PDF documents are usually created using Microsoft Word or Adobe Acrobat, and saved in the read only.pdf format. You can test if your document is truly a PDF document by attempting to copy text from the document using the Windows clipboard.

For these 'true' PDF documents any of the techniques outlined in this guide can be used to extract data.

PDF Images

These are often scanned documents saves as .pdf or .tiff format images. It is impossible to copy text from these images.

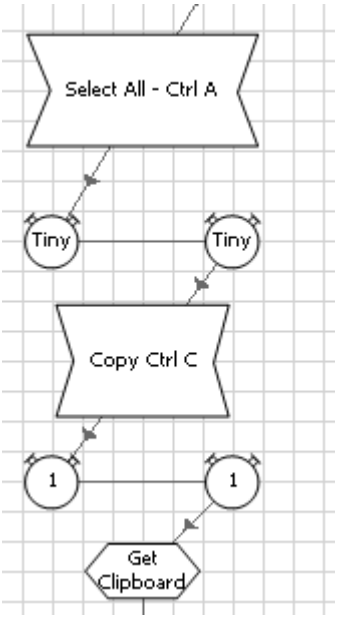
For these images only the ' Reading Text with OCR' technique can be used to extract data. OCR will only work if the image is of a high enough quality, 300dpi is recommended as a minimum. The Tesseract OCR engine used by Blue Prism cannot be used to read hand written text.

Using Windows Clipboard

It is recommend that the following training is taken prior to attempting to use this technique:

- Surface Automation Training

For this technique you need to create a simple Windows interface object in Object Studio that does the following:



- Launches or attaches to your document displayed in Adobe Reader
- Selects all the text within the document by clicking within the document and using Ctrl and a keystrokes
- Copies the selected text to the Windows clipboard by using the Ctrl and c keystrokes
- Use the GetClipboard() function in a Calculation stage to return the PDF text to Blue Prism

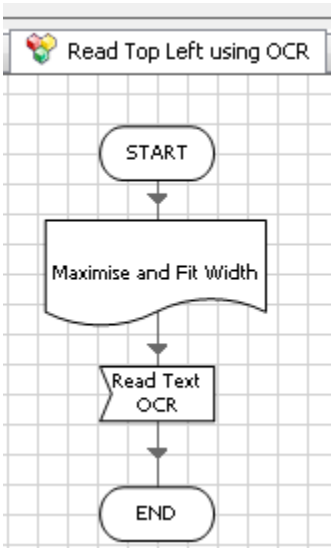
Using Read Text with OCR

It is recommend that the following training is taken prior to attempting this technique:

- Surface Automation Training
- Guide to Reading Text with OCR

For this technique you need to use Region Editor in Application Modeller to teach Blue Prism the regions of the document from which you want to read your text.

Your interface will need to contain the following steps:



- Launches or attaches to your document displayed in Adobe Reader
- Maximise and/or resize the document to make the interface more standardised.
- Uses the Read Text with OCR functionality to read text from your regions

> * Note: the OCR functionality within Blue Prism works best when used upon smaller regions rather than on large document areas.

Using the Adobe Acrobat API

It is recommend that the following data sheet is read prior to attempting this technique:

- Blue Prism Data Sheet - Extending Automations using the .NET Framework

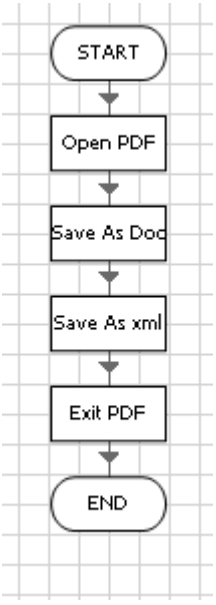
If Adobe Acrobat Standard or Professional is installed on a resource pc it is possible to use an Acrobat API to export or convert PDF documents into other formats such as XML or Microsoft Word.

Saving the document as XML or a Word document may be preferable for more complex documents where more data needs to be extracted from more complex document structures such as tables.

Once saved as XML or a Word .doc file an alternative interface can be used to interact with the document (such as the Blue Prism MS Word VBO business object).

To use the Adobe Acrobat API the following steps need to be taken:

- Adobe Acrobat Standard or Professional needs to be installed on the PC. There is a license cost for this, so if it is required it is recommended your solution is designed so that the minimum number of robots need to interface with PDF documents.
- Code stages are required that use the Acrobat API.
- The Save and SaveAs functions within the API allow you to save the PDF in different formats.



PDF created with Accessibility Features

It should be noted that PDF documents can be created using Adobe Acrobat with accessibility in mind. Documents created using features in Acrobat such as forms and/or tags can be interfaced with using the Blue Prism Active Accessibility interface.

If the PDF documents you are need to interface with are created internally within your organisation, it may be worth discussing with the owner of the documents if accessibility features could be used to make Robotic Process Automation easier.

Extracting data from text

Once you have captured the PDF document text using one of the techniques outlined above you may still need to implement some logic to extract the data you want from the within the text.

For example:

We have captured the text below from a top left region in a Purchase Order using the 'Read Text with OCR' feature:

```
The following number must appear on all
related correspondence, shipping papers,
and invoices:
P.O. NUMBER: 00012345678
TO:
Mr J Bloggs
Wigits R Us
202 Factory Street
Glasgow G5 5CD
Phone 0330 200 3000
```

From the above text taken from the PDF we just want to extract the number '00012345678' for use in the business process we are automating.

There are two methods that can be used for extract the data we want from the captured PDF text:

- Calculation stages using text expressions such as InStr, Left, Mid, etc.;

  For this example of extracting the PO number you might use InStr expressions to find the text P.O. Number: and the next newline character after it, and then a Mid expression to cut the PO Number into a data item.

  For example this expression will return the position of the P.O. NUMBER text:

  ```
  InStr([Purchase Order Text],"P.O. NUMBER:")
  ```

  For learning how to build your own calculation expressions in Blue Prism the Expression Function Builder and Evaluate Expression features within the Calculation stage properties window are recommended.

- Regular Expressions

  For this example of extracting a PO number you might use a regular expression that can retrieve the first numeric field after the text P.O. Number.

  For example, this regular expression will return the line of text containing P.O. NUMBER:

  ```
  (?:P(.?)*O(.?)*NUMBER).*
  ```

  For learning how to build your own regular expressions the website http://regexr.com/ is recommend. It contains both examples of regular expressions, and it allows you to paste in your own text and experiment with regular expression syntax to get the result you want.

> * Note: Some more technical developers might create their logic for extracting data from text in code stages. While this is a valid method the future ease of support of any bespoke code stages you create in your organisation should be considered.

Testing your PDF data extraction Logic

Your solution should be tested against large number of PDF documents.

Because the logic developed to extract text is dependent upon the predictability of the structure of PDF documents it is recommend that the logic you develop to extract data is tested using as many examples of the PDF document as possible.