RPA Design & Development

PDF Automation

Ui Path™

2 types of PDF

Native text format

Scanned image

A PDF file that is originally generated in a computer, aka "born digital", meaning that it was created from an original electronic version of a document. One quick way to tell that a PDF is native is that you can select blocks of text in the file.
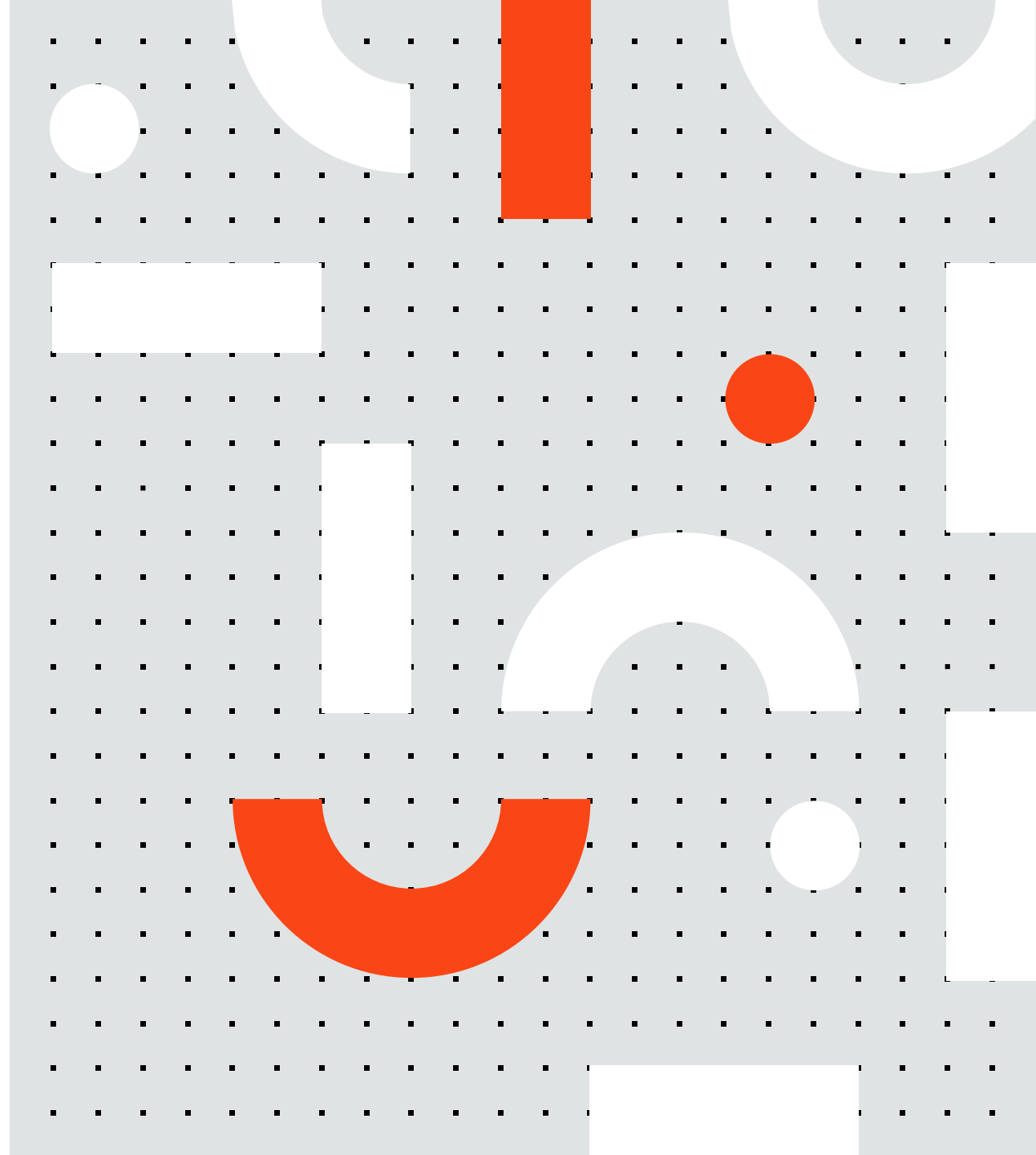
A PDF file that is made up of scanned images of a given document. With scanned PDFs you will mot be able to select text or use the search function because the PDF is a collection of images.

Extracting data from PDF

In PDF automation, data can be extracted using two separate activities:

Extract data using Read PDF Text activity
is the more accurate of the two but works only with native PDF documents.

Extract data using Read PDF with OCR activity
is less reliable but can extract text from scanned PDF documents.
also requires the selection of an OCR engine.

Both activities are available in the UiPath.PDF.Activities package.

UiPath™

Practice makes Perfect… Reading PDFs

Use some pdf documents to test the performance of the two Read PDF activities.

Extracting data from PDF through the UI

In PDF automation, data can be also extracted using the UI Activities

Get Text
Get Full Text
Screen Scraping

What technical setup do I need?

Change Adobe Reader DC Acrobat Settings

If the PDF is opened with Adobe Reader DC Acrobat, there might be a few steps to take before you can extract specific elements using UiPath Studio methods.

•Start Acrobat and press Ctrl+K. That opens the Preferences pop-up.

•Select Reading, out of the categories on the left panel.

•Verify that the drop-down Reading Order option is set to the Acrobat recommended option, 'Infer reading order from the document (recommended)'.

•'Page vs. Document': should be set to 'Read the entire document' and 'Confirm before tagging documents' should be unchecked.

•Then on the left panel, click Accessibility. In the Other Accessibility Options section, check the first two boxes if they are not already checked: 'Use document structure for tab order when no explicit tab order is specified', 'Enable assistive technology support'.

•Click OK.

Practice makes Perfect… Reading PDFs

Use some pdf documents to test the performance of the UI activities.

Questions & Answers

How can a robot read only the first page of a PDF file, using the PDF activities?

a) Set the Range property to: "1"

b) Set the Range property to: "all"

c) Set the Range property to: 1.

What is the easiest way to get the invoice number from a native PDF file?

a) Use the Read PDF with OCR activity and get the value by using string manipulation.

b) Open the PDF file with Adobe Acrobat Reader and scrape only the relevant information.

c) Use the Read PDF Text activity and get the value by using string manipulation.

If you want to extract specific information from a series of PDF files with a similar structure but the workflow only works for one file of the series, what should you investigate?

a) The TimeoutMS property.

b) The Selector property.

c) The ContinueOnError property.

If you want to extract specific information from multiple native PDF files with the same structure, what activity should you use?

a) Get Text Activity with OCR.

b) Get Text Activity

c) Read PDF Activity with OCR

If the PDF activities are not listed in your Activities Panel, how can you get them?

a) By going to the Output panel.

b) By installing them using the Manage Packages feature.

c) By finding them in the Library tab.

Next Steps

Email automation