LECTURE 10. PDF FILES PROCESSING

Robotic Process Automation
[05 December 2023]

Elective Course, 2023-2024, Fall Semester

Camelia Chisăliţă-Creţu, Lecturer PhD
Babeş-Bolyai University

Acknowledgements

This course is presented to our Faculty with the support of UiPath Romania.

UiPath

Robotic Process Automation

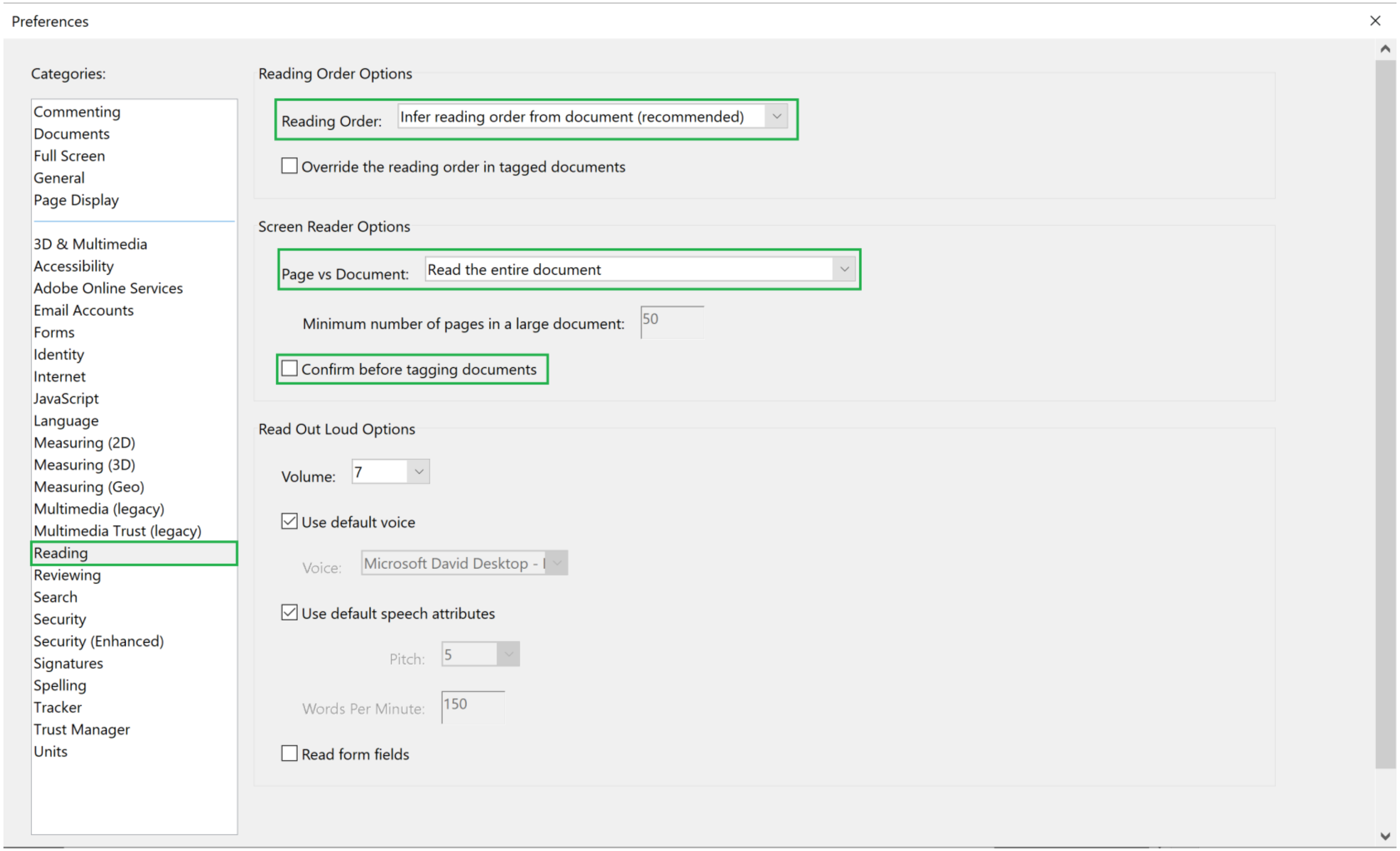
Contents

- File Processing
  - Preferences
    - Settings. Reading Option
    - Accessibility Option
    - Extracting Specific Elements
    - Details
  - Extract Text
    - Overview
    - UiPath Activities Packages
    - Read PDF Text Activity
      - Details
    - Read PDF with OCR Activity
      - Details. OCR Engine
    - Read PDF Text Activity vs Read PDF with OCR Activity
    - Demo 1. Read PDF Activities
    - Screen Scraping
      - Details
  - Extract Specific Elements
    - Anchor Base Activity
      - Details. Actions
    - Find Element Activity
    - Find Image Activity
    - Find Element Activity vs Find Image Activity
    - Demo 2. PDF and Excel Activities
- References

PDF File Processing. Preferences Settings

- Steps to be performed before working with PDF files in Acrobat Reade DC (Document Cloud) (see *next slide*):
  1. start Acrobat Reader DC (AR DC);
  2. open the **Preferences** window:
     - **Edit** menu --> **Preference**s… option or
     - shortcut **Ctrl +K**;
  3. select **Reading** section:
     - option **Page vs Document** = ‘Read the entire document’;
     - option **Confirm before tagging documents** = *unchecked*;
     - option **Reading Order** = ”Infer reading order from document (recommended)”;
  4. select **Accessibility** section:
     - option **Other Accessibility Options**:
       - first two checkboxes should be selected:
         - **Use the document structure for tab order where explicit tab order is specified**;
         - **Enable assistive technology support**.

PDF File Processing. Preferences – Reading Option



Preferences

Categories:

- Commenting
- Documents
- Full Screen
- General
- Page Display
- 3D & Multimedia
- Accessibility
- Adobe Online Services
- Email Accounts
- Forms
- Identity
- Internet
- JavaScript
- Language
- Measuring (2D)
- Measuring (3D)
- Measuring (Geo)
- Multimedia (legacy)
- Multimedia Trust (legacy)
- Reading
- Reviewing
- Search
- Security
- Security (Enhanced)
- Signatures
- Spelling
- Tracker
- Trust Manager
- Units

Reading Order Options

Reading Order: Infer reading order from document (recommended)

□ Override the reading order in tagged documents

Screen Reader Options

Page vs Document: Read the entire document

Minimum number of pages in a large document: 50

□ Confirm before tagging documents

Read Out Loud Options

Volume: 7

☑ Use default voice

Voice: Microsoft David Desktop - I

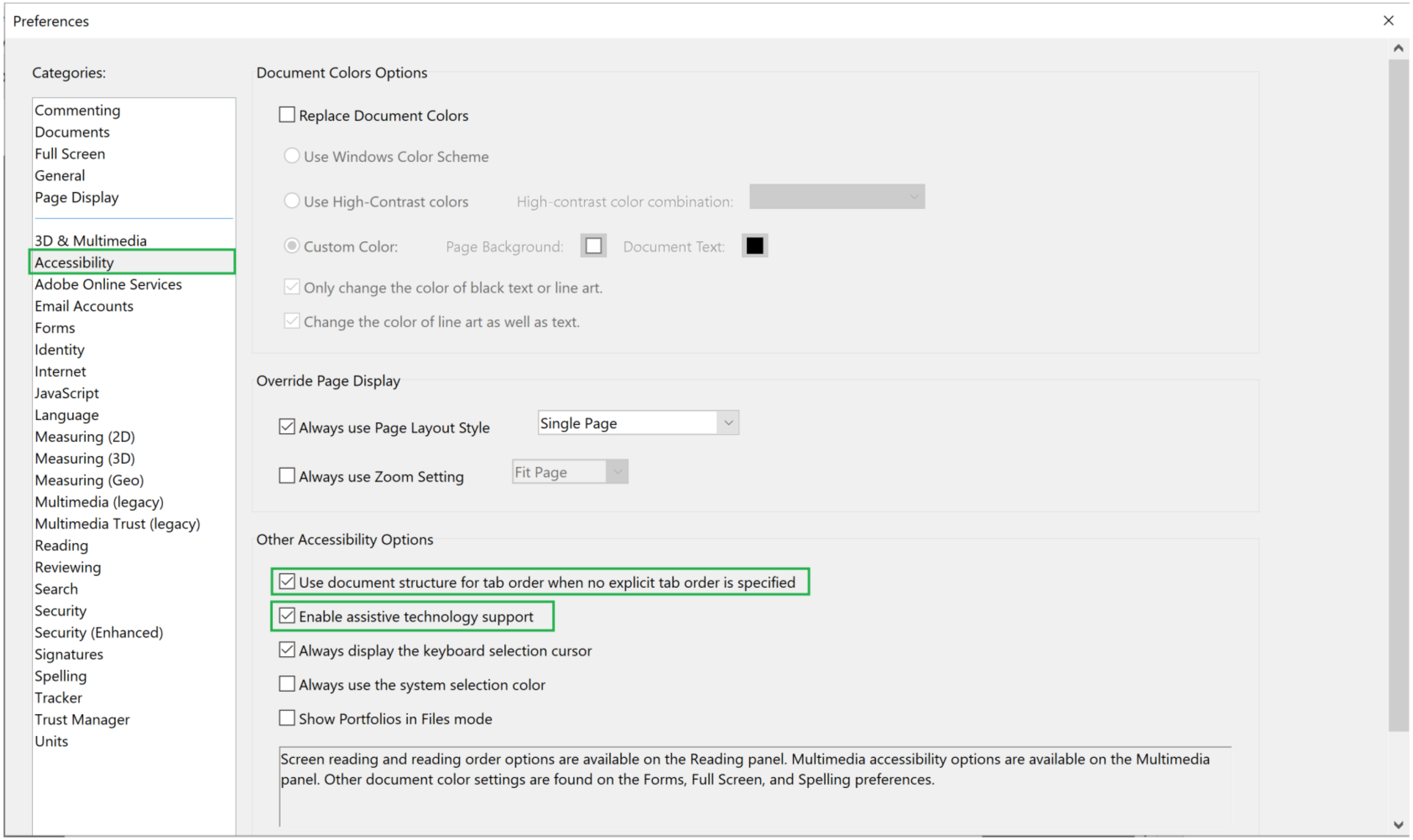
☑ Use default speech attributes

Pitch: 5

Words Per Minute: 150

□ Read form fields

PDF File Processing. Preferences – Accessibility Option



Preferences

Categories:

- Commenting
- Documents
- Full Screen
- General
- Page Display
- 3D & Multimedia
- Accessibility
- Adobe Online Services
- Email Accounts
- Forms
- Identity
- Internet
- JavaScript
- Language
- Measuring (2D)
- Measuring (3D)
- Measuring (Geo)
- Multimedia (legacy)
- Multimedia Trust (legacy)
- Reading
- Reviewing
- Search
- Security
- Security (Enhanced)
- Signatures
- Spelling
- Tracker
- Trust Manager
- Units

Document Colors Options

□ Replace Document Colors

○ Use Windows Color Scheme

○ Use High-Contrast colors   High-contrast color combination:

● Custom Color: Page Background: □   Document Text: ■

☑ Only change the color of black text or line art.

☑ Change the color of line art as well as text.

Override Page Display

☑ Always use Page Layout Style   Single Page

□ Always use Zoom Setting   Fit Page

Other Accessibility Options

☑ Use document structure for tab order when no explicit tab order is specified

☑ Enable assistive technology support

☑ Always display the keyboard selection cursor

□ Always use the system selection color

□ Show Portfolios in Files mode

Screen reading and reading order options are available on the Reading panel. Multimedia accessibility options are available on the Multimedia panel. Other document color settings are found on the Forms, Full Screen, and Spelling preferences.

PDF File Processing. Extracting Specific Elements

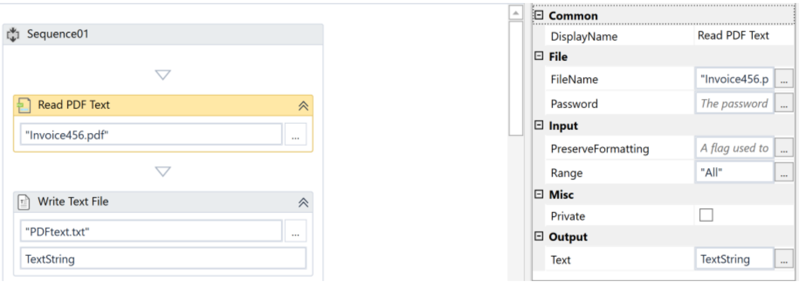
- when we have problems with extracting specific elements from PDF files opened with AR DC, an older version of ARDC may be used;
  - the web-page https://www.adobe.com/devnet-docs/acrobatetk/tools/ReleaseNotesDC/index.html consists of the list with the AR DC versions available to download;
- AR DC is updated automatically to the last available version;
- starting with version 19, there may be some problems with accessibility, as AR DC is slowly dropping support for untagged documents; steps to solve this issues:
  - uninstall the current version of AR DC;
  - install the *base* release AR DC
    - https://www.adobe.com/devnet-docs/acrobatetk/tools/ReleaseNotesDC/continuous/dccontinuous.html#dccontinuous.

PDF File Processing. Details

- PDF files consists of data presented as:
  - text;
  - images;
  - text undercover images;
- in order to identify which element is text or image, the element is selected and:
  - if it can be easily selected ==> text;
  - if an apparent block is selected ==> image;
- in UiPath there are activities and methods for:
  1. **Extracting text**, based on
     - for large chunks of native text or whole documents, based on **non-OCR** and **OCR-based** processing (**screen scraping**);
  2. **Extracting specific elements**, based on
     - **Scraping actions**:
       - **Relative scraping**;
     - **Anchor-based** actions:
       - **Find Element** or **Find Image**.

PDF File Processing. Extract Text Overview

1. **Extracting text** methods include the following activities:
   - **Read PDF Text**;
   - **Read PDF with OCR**;
   - **Screen Scraping**:
     - **Attach Window**;
     - **Get Full Text**/ **Get Visible Text**/ **Get OCR Text**.



Sequence01

Read PDF Text

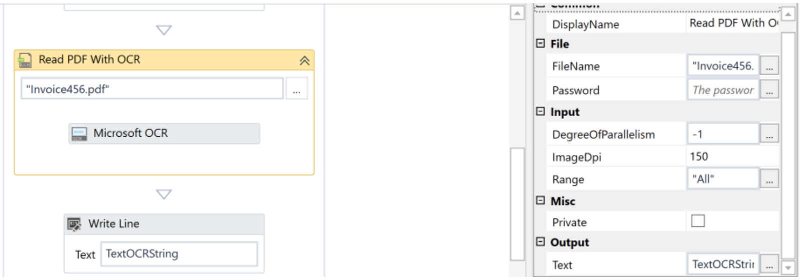
"Invoice456.pdf"

Write Text File

"PDFtext.txt"

TextString

Common	
DisplayName	Read PDF Text
File	
FileName	"Invoice456.p
Password	*The password*
Input	
PreserveFormatting	*A flag used to*
Range	"All"
Misc	
Private	□
Output	
Text	TextString



Read PDF With OCR

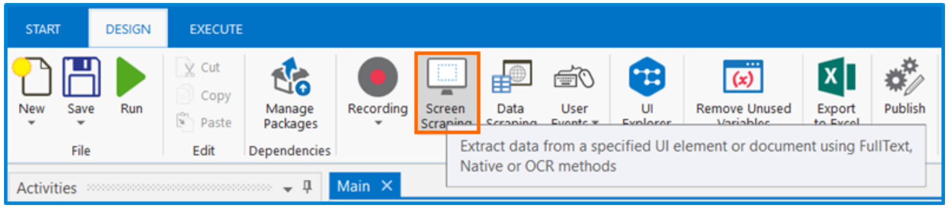
"Invoice456.pdf"

Microsoft OCR

Write Line

Text TextOCRString

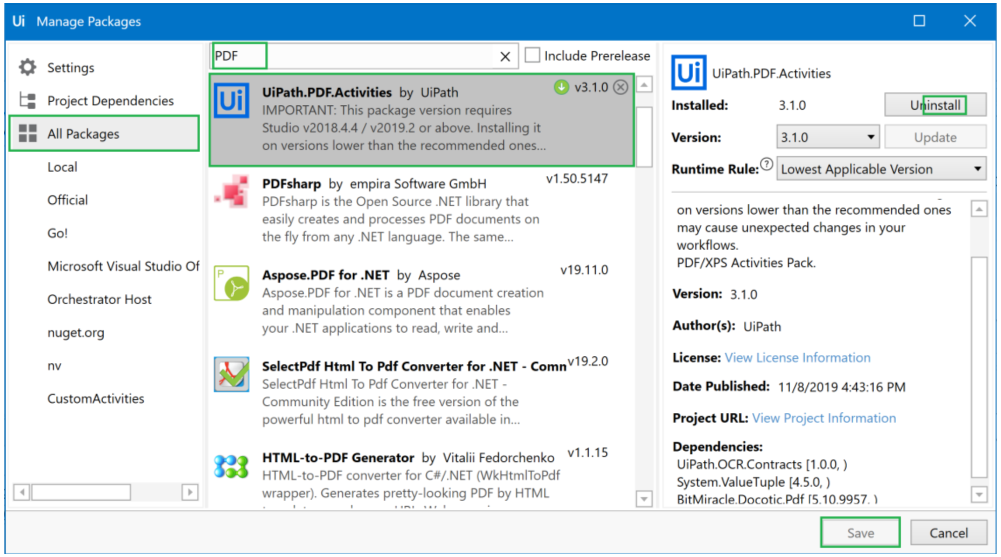
DisplayName	Read PDF With O
File	
FileName	"Invoice456.
Password	*The passwor*
Input	
DegreeOfParallelism	-1
ImageDpi	150
Range	"All"
Misc	
Private	□
Output	
Text	TextOCRStri



Extract data from a specified UI element or document using FullText, Native or OCR methods

PDF File Processing. UiPath Activity Packages

- in order to work with specific PDF activities, some packages need to be installed;
- Steps;
  1. Search in the **Activities Panel** for ‘PDF’;
  2. If the result is empty the PDF activity package needs to be installed;
  3. In **Manage Packages**, **All Packages** section, search for ‘PDF’;
  4. install **UiPath.PDF.Activities**;
  5. Install, Save.



UiPath.PDF.Activities by UiPath v3.1.0
IMPORTANT: This package version requires Studio v2018.4.4 / v2019.2 or above. Installing it on versions lower than the recommended ones...

PDFsharp by empira Software GmbH v1.50.5147
PDFsharp is the Open Source .NET library that easily creates and processes PDF documents on the fly from any .NET language. The same...

Aspose.PDF for .NET by Aspose v19.11.0
Aspose.PDF for .NET is a PDF document creation and manipulation component that enables your .NET applications to read, write and...

SelectPdf Html To Pdf Converter for .NET - Comn v19.2.0
SelectPdf Html To Pdf Converter for .NET - Community Edition is the free version of the powerful html to pdf converter available in...

HTML-to-PDF Generator by Vitalii Fedorchenko v1.1.15
HTML-to-PDF converter for C#/.NET (WkHtmlToPdf wrapper). Generates pretty-looking PDF by HTML

UiPath.PDF.Activities

Installed: 3.1.0

Version: 3.1.0

Runtime Rule: Lowest Applicable Version

on versions lower than the recommended ones may cause unexpected changes in your workflows.
PDF/XPS Activities Pack.

Version: 3.1.0

Author(s): UiPath

License: View License Information

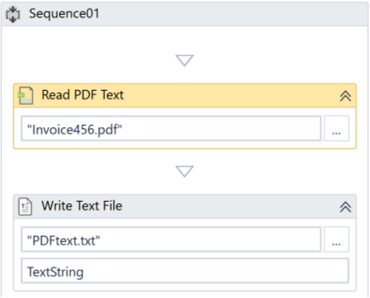
Date Published: 11/8/2019 4:43:16 PM

Project URL: View Project Information

Dependencies:
UiPath.OCR.Contracts [1.0.0,)
System.ValueTuple [4.5.0,)
BitMiracle.Docotic.Pdf [5.10.9957,)

Read PDF Text Activity. Details

- **Read PDF Text** activity
  - is used to read text from PDF files, as whole document or part of it, i.e., pages;
  - is part of the **UiPath.PDF Activities** package;
- only the text part of the document is processed;
- the image ignored; the result contains the place holder: <Text & Image PDF>;
- relevant properties:
  - **[File] FileName**= **String** variable
    - the PDF file to read from;
  - **[Input] Range** = **String** variable
    - the actual range of pages read from the PDF file;
    - this can be: “All”, “1”, “3-5” “12”;
  - **[Output] Text** = **String** variable
    - this is the result of the reading process.



Sequence01

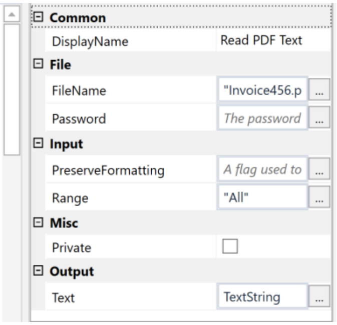
Read PDF Text

"Invoice456.pdf"

Write Text File

"PDFtext.txt"

TextString



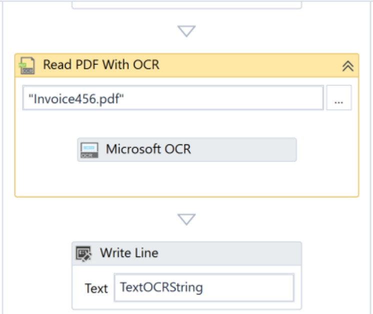
Common	
DisplayName	Read PDF Text
File	
FileName	"Invoice456.p
Password	*The password*
Input	
PreserveFormatting	*A flag used to*
Range	"All"
Misc	
Private	□
Output	
Text	TextString

see **Demo1A – PDFFiles**

UiPath™

Read PDF With OCR Activity. Details

- **Read PDF with OCR** activity
  - is used to read text that is included in the image blocks of PDF files;
  - is part of the **UiPath.PDF Activities** package;
- the image is scanned using an OCR engine and the text result is returned;
- relevant properties:
  - **[File] FileName= String** variable
    - the PDF file to read from;
  - **[Input] Range = String** variable
    - the actual range of pages read from the PDF file;
    - this can be: “All”, “1”, “3-5” “12”;
  - **[Output] Text = String** variable
    - this is the result of the reading process.



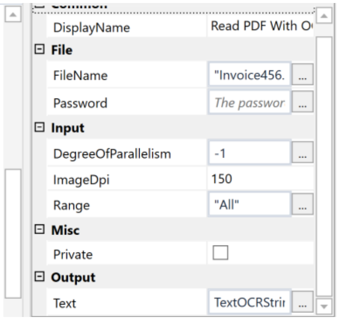
Read PDF With OCR

"Invoice456.pdf"

Microsoft OCR

Write Line

Text: TextOCRString

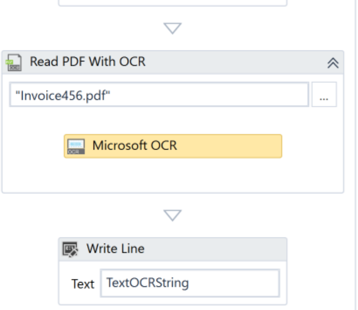


DisplayName	Read PDF With O
File	
FileName	"Invoice456.
Password	*The passwor*
Input	
DegreeOfParallelism	-1
ImageDpi	150
Range	"All"
Misc	
Private	□
Output	
Text	TextOCRStri

see **Demo1A – PDFFiles**

Read PDF With OCR Activity. OCR Engine (1)

- **OCR Engine** associated to **Read PDF with OCR** allows to customize the used OCR engine;
- various OCR engines are used:
  - **Tesseract, Microsoft, Abbyy, UiPath Screen,** etc.;
- relevant properties:
  - **[Input] Image**= the scanned image
  - **[Options] ExtractWords, Language, Profile, Scale, etc;**
    - various OCR engine properties that can be set;
    - the properties depend on the particular OCR engine;
  - **[Output] Text** = **String** variable
    - this is the result of the character recognition process;



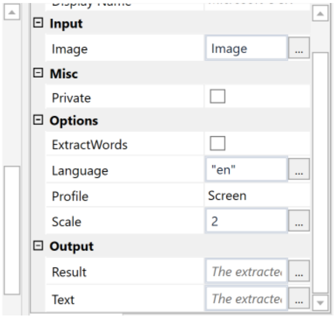
Read PDF With OCR

"Invoice456.pdf" ...

Microsoft OCR

Write Line

Text TextOCRString

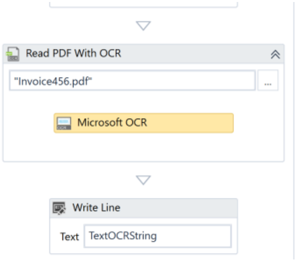


Input	
Image	Image ...
Misc	
Private	□
Options	
ExtractWords	□
Language	"en" ...
Profile	Screen
Scale	2 ...
Output	
Result	*The extracte* ...
Text	*The extracte* ...

see **Demo1A – PDFFiles**

Read PDF With OCR Activity. OCR Engine (2)

- **OCR Engine** allows to convert the entire *text* and *image* to text;
- if the text is placed in two columns the text is **intertwined** together;
    - this is because the OCR engines do not automatically recognize the 2-column layout of the document;
- **Abbyy OCR Engine** is an exception;
    - it preserves the document structure; as result it separates the read columns;
- the quality of the OCR engine results degrades quickly with the quality of the source image;
    - therefore, the recognition results depend on:
        - the font size;
        - the font face;
        - the image resolution;
    - it cannot be controlled by the user;
        - whenever is possible, *non-OCR* **Read PDF Text** activities are recommended.



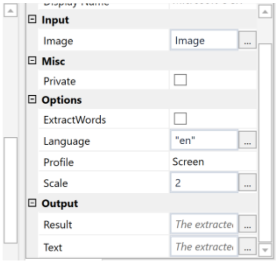
Read PDF With OCR

"Invoice456.pdf"

Microsoft OCR

Write Line

Text: TextOCRString



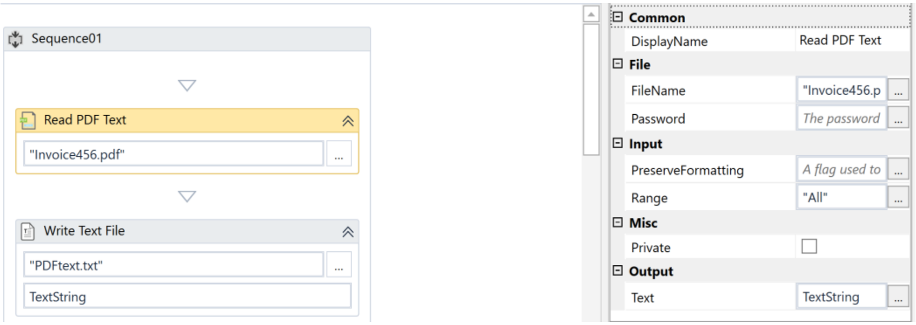
Input	
Image	Image
Misc	
Private	□
Options	
ExtractWords	□
Language	"en"
Profile	Screen
Scale	2
Output	
Result	*The extracte*
Text	*The extracte*

see **Demo1A – PDFFiles**

Read PDF Text Activity vs Read PDF With OCR Activity

Read PDF Text activity

- it reads **text only** from PDF files;
- self-contained:
  - don’t require opening the files using other applications;
  - can work in background.



Sequence01

Read PDF Text

"Invoice456.pdf"

Write Text File

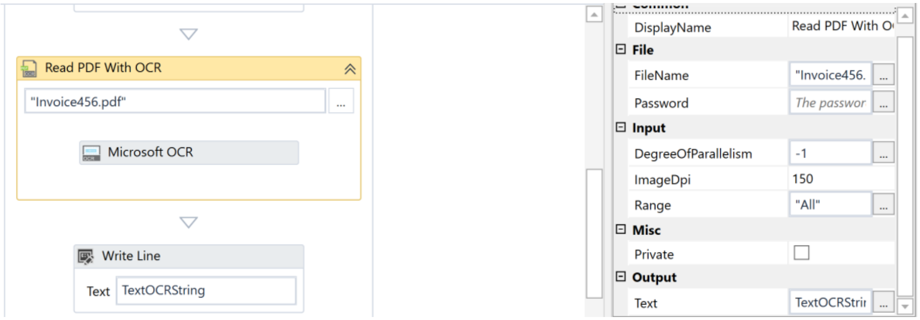
"PDFtext.txt"

TextString

⊟ **Common**	
DisplayName	Read PDF Text
⊟ **File**	
FileName	"Invoice456.p
Password	*The password*
⊟ **Input**	
PreserveFormatting	*A flag used to*
Range	"All"
⊟ **Misc**	
Private	□
⊟ **Output**	
Text	TextString

Read PDF With OCR activity

- it reads **text and image** from PDF files;
- self-contained:
  - don’t require opening the files using other applications;
  - can work in background.
- *the quality of OCR engine-based reading degrades quickly with the quality of the source image*.



Read PDF With OCR

"Invoice456.pdf"

Microsoft OCR

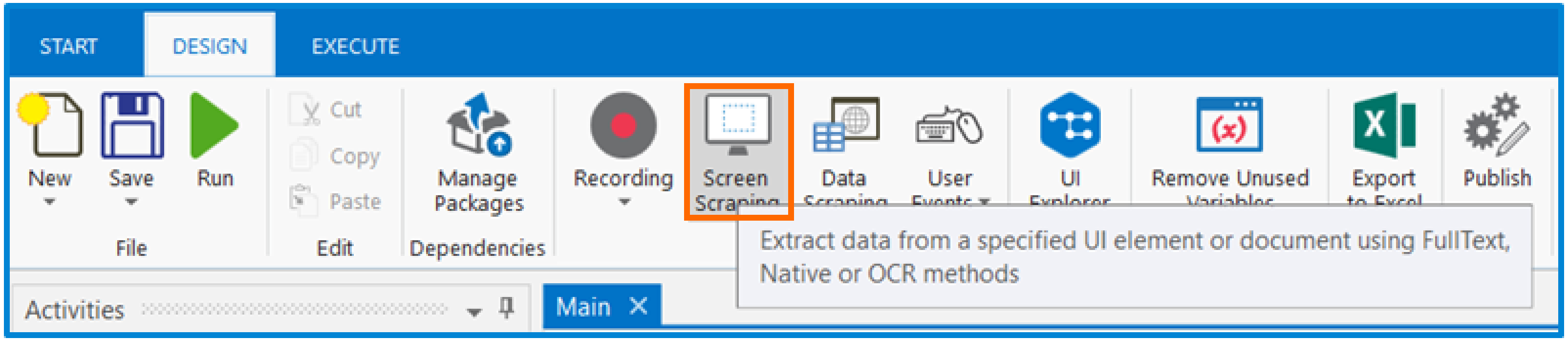
Write Line

Text TextOCRString

DisplayName	Read PDF With O
⊟ **File**	
FileName	"Invoice456.
Password	*The passwor*
⊟ **Input**	
DegreeOfParallelism	-1
ImageDpi	150
Range	"All"
⊟ **Misc**	
Private	□
⊟ **Output**	
Text	TextOCRStri

Screen Scraping. Details

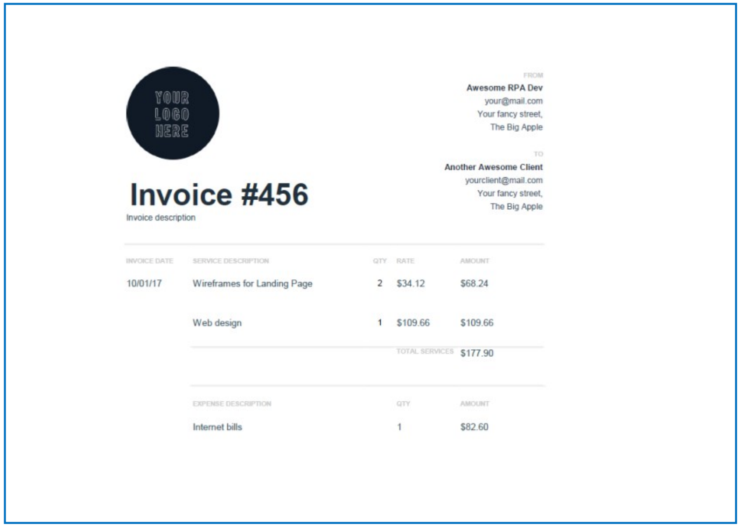
- **Screen Scraping** was mainly covered in **Lecture 06. UI Interactions**;
- the associated recorder allows to generate the following activities:
  - **Attach Window** activity;
  - **Get Full Text**/ **Get Visible Text**/ **Get OCR Text** activities based on the output method chosen.



Extract data from a specified UI element or document using FullText, Native or OCR methods

Demo 1A. Read PDF Activities

- **Indicate the UiPath.PDF.Activities package installation steps in Manage Packages;**
- **Use Read PDF Text and Read PDF with OCR activities to get data from particular .pdf files that contain text and text as image;**
  - **customize the character recognition process using different OCR engines and their corresponding properties;**
- **Write the extracted text into a .txt file;**
  - **inspect and discuss the results.**



YOUR
LOGO
HERE

FROM
Awesome RPA Dev
your@mail.com
Your fancy street,
The Big Apple

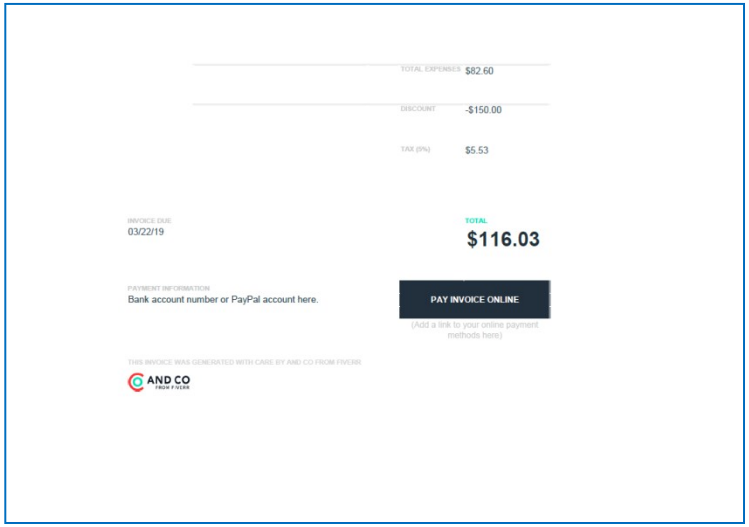
TO
Another Awesome Client
yourclient@mail.com
Your fancy street,
The Big Apple

Invoice #456

Invoice description

INVOICE DATE	SERVICE DESCRIPTION	QTY	RATE	AMOUNT
10/01/17	Wireframes for Landing Page	2	$34.12	$68.24
	Web design	1	$109.66	$109.66
			TOTAL SERVICES	$177.90

EXPENSE DESCRIPTION	QTY	AMOUNT
Internet bills	1	$82.60



TOTAL EXPENSES	$82.60
DISCOUNT	-$150.00
TAX (5%)	$5.53

INVOICE DUE
03/22/19

TOTAL
$116.03

PAYMENT INFORMATION
Bank account number or PayPal account here.

PAY INVOICE ONLINE

(Add a link to your online payment methods here)

THIS INVOICE WAS GENERATED WITH CARE BY AND CO FROM FIVERR

AND CO
FROM FIVERR

see **Demo1A – PDFFiles**

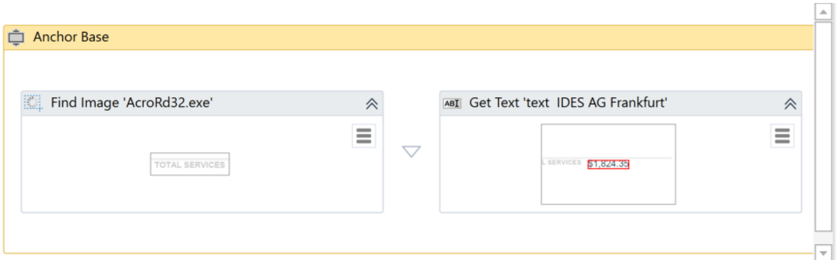
PDF File Processing. Extract Specific Elements

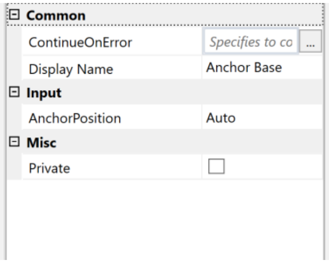
2. **Extracting specific elements** includes the following activities: based on
   - **Anchor-based** actions:
     - **Anchor Base**;
     - **Find Element** /**Find Image**;
     - **Get Full Text**/ **Get Visible Text**/ **Get OCR Text**.
   - **Relative scraping**
     - **Lecture 09. Image and text Automation**;
   - **Find Relative Element**.

see **Demo1B – PDFFiles**

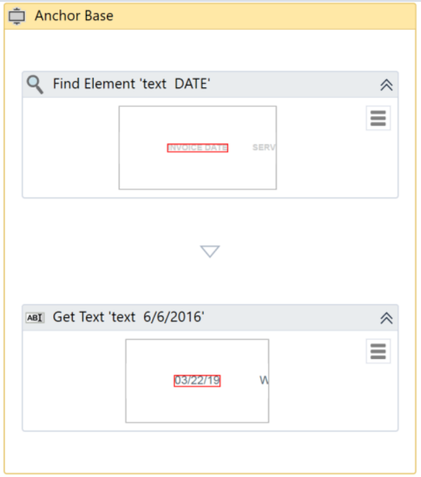
Anchor Base Activity. Details

- **Anchor Base** activity
  - is a container of two activities performed on different elements;
  - allows to perform a specific action (the second action) by identifying a *fixed* element or image (the first action) that is related to the *target* element;
  - **requires the document to be opened and the elements to interact with to be visible, otherwise it fails;**
- relevant properties:
  - **[input] Anchor Position** = **Auto, Left, Right, Bottom, Top**;
    - it indicates where the *anchor* (fixed element) is placed in relation with the *targeted* text.





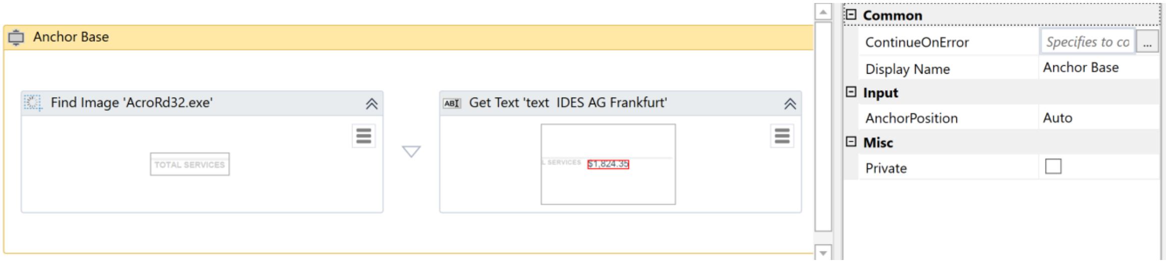
Common
ContinueOnError: Specifies to co...
Display Name: Anchor Base
Input
AnchorPosition: Auto
Misc
Private: □



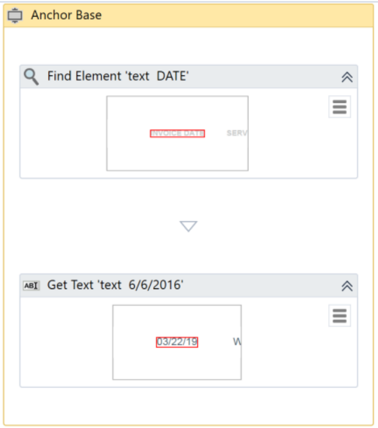
see **Demo1B – PDFFiles**

Anchor Base Activity. Actions

- **First action:**
  - usually, this is an identification activity of the anchor element: **Find Element** or **Find Image**;
  - the ***fixed element*** selector needs to be customized because it has no unique identifiers;
  - the ***target element*** selector has set only the last row of the full selector of the fixed element;
- **Second action:**
  - usually, this is a **data extraction** or a **keyboard/mouse action** activity: **Get Full Text, Get Visible Text, Get OCR Text, Click Text, Click Image, Send Hotkeys**, etc.



⊟ Common
ContinueOnError   Specifies to co
Display Name   Anchor Base
⊟ Input
AnchorPosition   Auto
⊟ Misc
Private   □

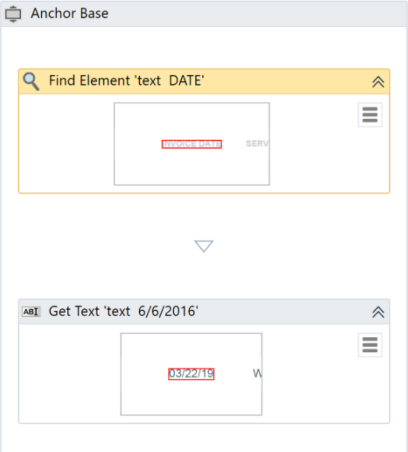


see **Demo1B – PDFFiles**

Anchor Base Activity. Find Element Activity

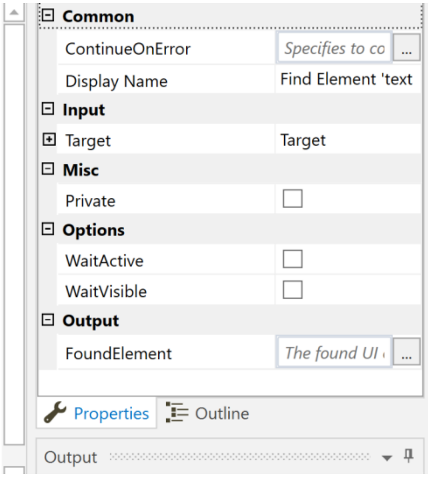
- **Find Element** activity
  - allows to identify a fixed element in the document that will be used as an anchor to perform an action on another element;
- **the structure of the document is essential, i.e., the identification process relies on the document structure;**
- the selectors consists of the relevant information to correctly identify the fixed element;
- relevant properties:
  - **[output] Found Element** = **UiElement** variable;
    - the identified element can be used in further processing.

see **Demo1B – PDFFiles**


Anchor Base

Find Element 'text DATE'

Get Text 'text 6/6/2016'


⊟ Common

ContinueOnError: *Specifies to co* ...

Display Name: Find Element 'text

⊟ Input

⊞ Target: Target

⊟ Misc

Private: □

⊟ Options

WaitActive: □

WaitVisible: □

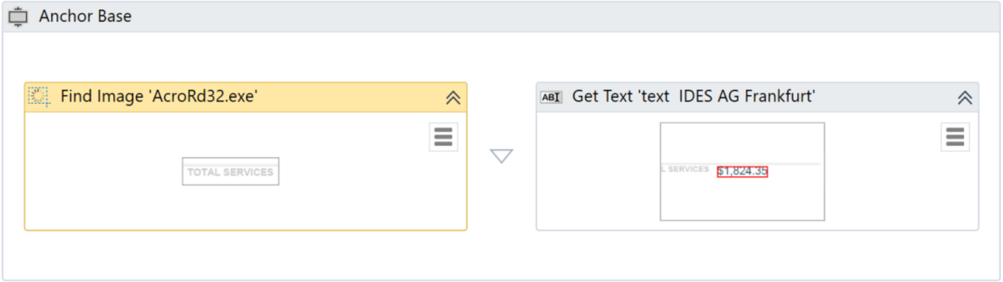
⊟ Output

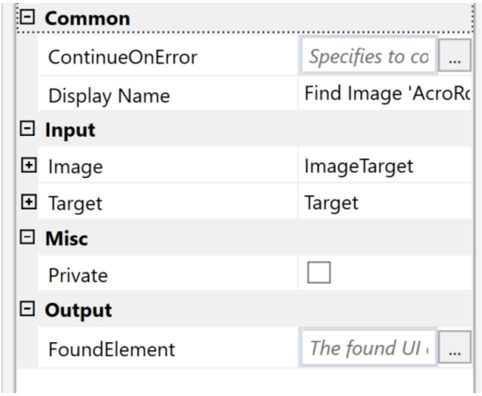
FoundElement: *The found UI* ...

Anchor Base Activity. Find Image Activity (1)

- **Find Image** activity
  - allows to identify a fixed image in the document that will be used as an anchor to perform an action on another element;
- **the structure of the document <u>is not important</u>, i.e., the identification process does not rely on the document structure;**
- **it is important to contain the indicated image anywhere in the document, i.e., the selectors are not useful anymore;**
- relevant properties:
  - **[output] Found Element** = **UiElement** variable;
    - the identified image element that can be used in further processing.

see **Demo1B – PDFFiles**

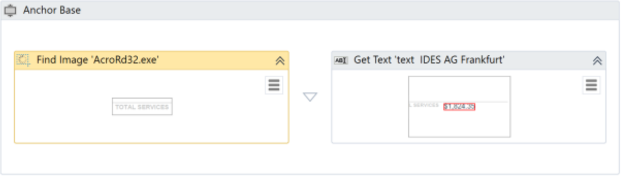




Common	
ContinueOnError	*Specifies to co* ...
Display Name	Find Image 'AcroRd
Input	
Image	ImageTarget
Target	Target
Misc	
Private	□
Output	
FoundElement	*The found UI* ...

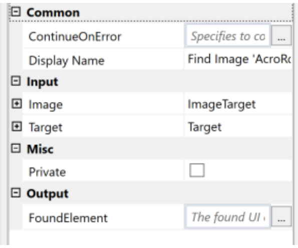
Anchor Base Activity. Find Image Activity (2)

- **Find Image** activity
  - useful for PDF documents because similar PDF documents look similar;
  - image accuracy can be affected by the size of the page of the opened document;
  - in order to have a complete and accurate image it is recommended to set a proper page size, i.e., use the **View** menu, **Zoom** option and **Actual Size** option;
  - in some cases it can be more reliable because:
    - it can handle major structural changes of the document, as long as:
      - the image looked for and the targeted data are present;
      - the relationship between the image and the targeted data is the same one to another;
    - it can handle reasonable amount of scale variation.


Anchor Base

Find Image 'AcroRd32.exe'

Get Text 'text IDES AG Frankfurt'

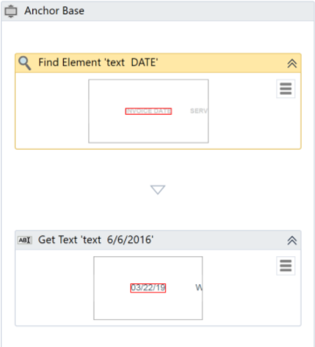

Common	
ContinueOnError	*Specifies to co* ...
Display Name	Find Image 'AcroRc
Input	
Image	ImageTarget
Target	Target
Misc	
Private	□
Output	
FoundElement	*The found UI* ...

see **Demo1B – PDFFiles**

Find Element Activity vs Find Image Activity

Find Element Activity

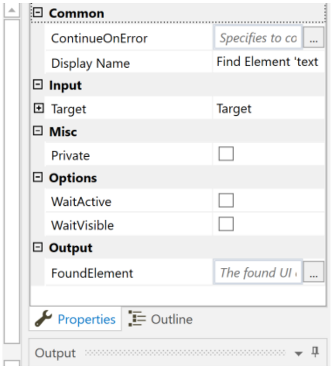
- the structure of the document is important, i.e., the identification process relies on the document structure;
- the selectors consists of the relevant information to correctly identify the fixed element.



Anchor Base

Find Element 'text DATE'

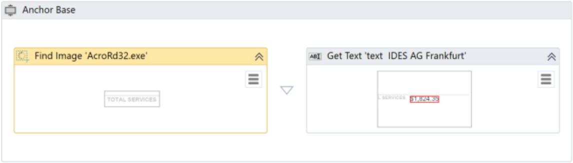
Get Text 'text 6/6/2016'



⊟ **Common**	
ContinueOnError	*Specifies to co* ...
Display Name	Find Element 'text
⊟ **Input**	
⊞ Target	Target
⊟ **Misc**	
Private	□
⊟ **Options**	
WaitActive	□
WaitVisible	□
⊟ **Output**	
FoundElement	*The found UI* ...

Find Image Activity

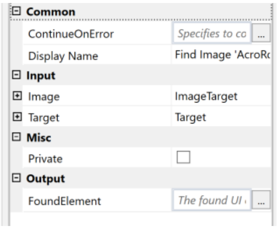
- the structure of the document is not important, i.e., the identification process does not rely on the document structure;
- it is important to contain the indicated image anywhere in the document, i.e., the selectors are not useful anymore.



Anchor Base

Find Image 'AcroRd32.exe'

Get Text 'text IDES AG Frankfurt'



⊟ **Common**	
ContinueOnError	*Specifies to co* ...
Display Name	Find Image 'AcroR
⊟ **Input**	
⊞ Image	ImageTarget
⊞ Target	Target
⊟ **Misc**	
Private	□
⊟ **Output**	
FoundElement	*The found UI* ...

Demo 1B. PDF and Excel Activities

- **Consider several PDF files that contains details on invoices;**

1. ***Extract* the following data:**
   - **invoice title, invoice date, total services, total (amount);**
     - **use activities that extract specific elements;**
2. ***Build* a data table with the extracted values;**
3. ***Sort* descending the data based on the invoice date;**
4. ***Export* the data table to an Excel file.**

see **Demo1B – PDFFiles**

References

- UiPath Docs - https://docs.uipath.com/studio
  - PDF Data Extraction - https://www.uipath.com/kb-articles/pdf-data-extraction-scrape-pdf-text
  - PDF Activities Pack - https://docs.uipath.com/activities/docs/about-the-pdf-activities-pack