

# LECTURE 09.

# PDF FILES PROCESSING

---

**Robotic Process Automation**

**[26 November 2019]**

Elective Course, 2019-2020, Fall Semester

Camelia Chisăliță-Crețu, Lecturer PhD

Babeș-Bolyai University

# Acknowledgements

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



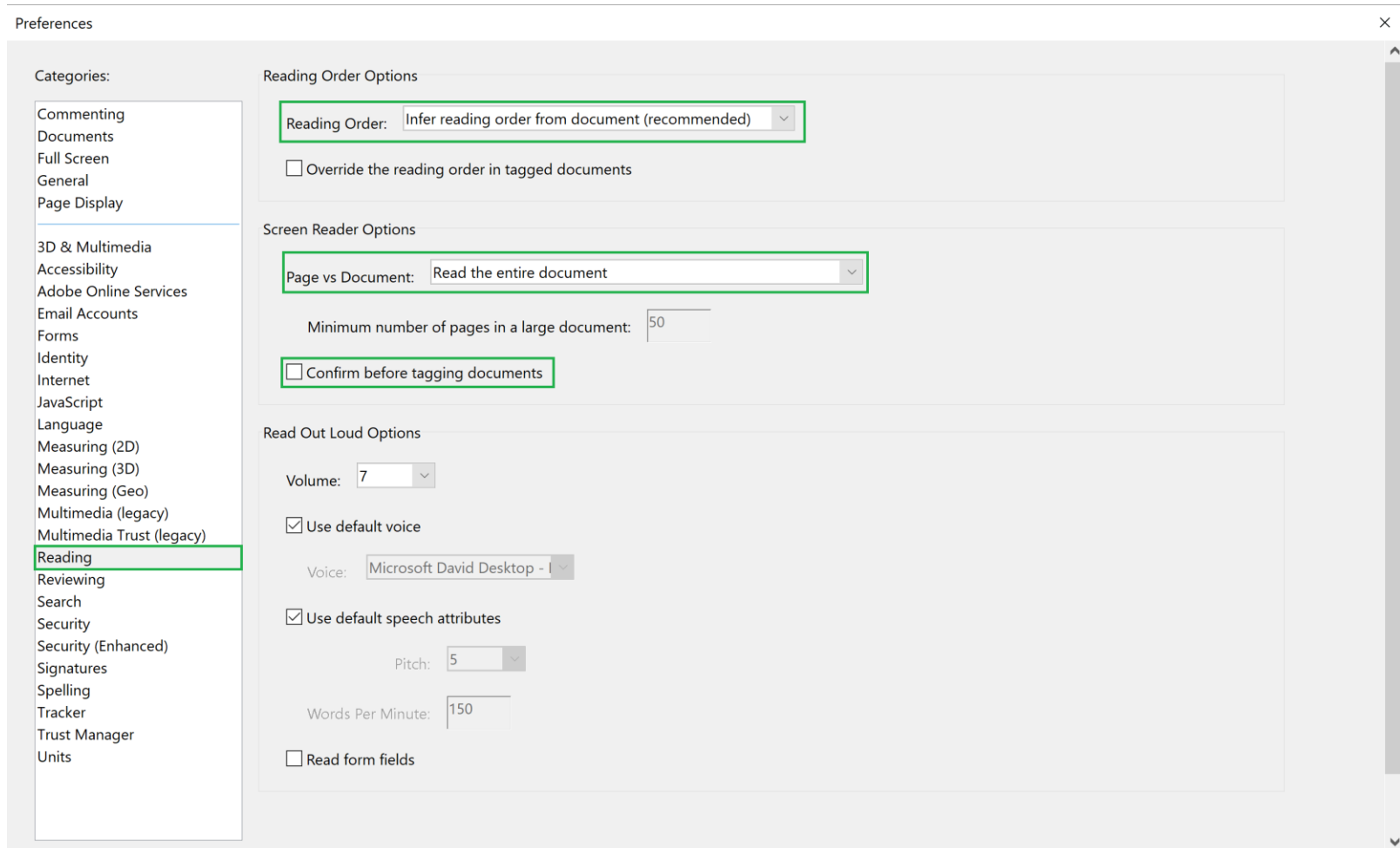
# Contents

- **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**

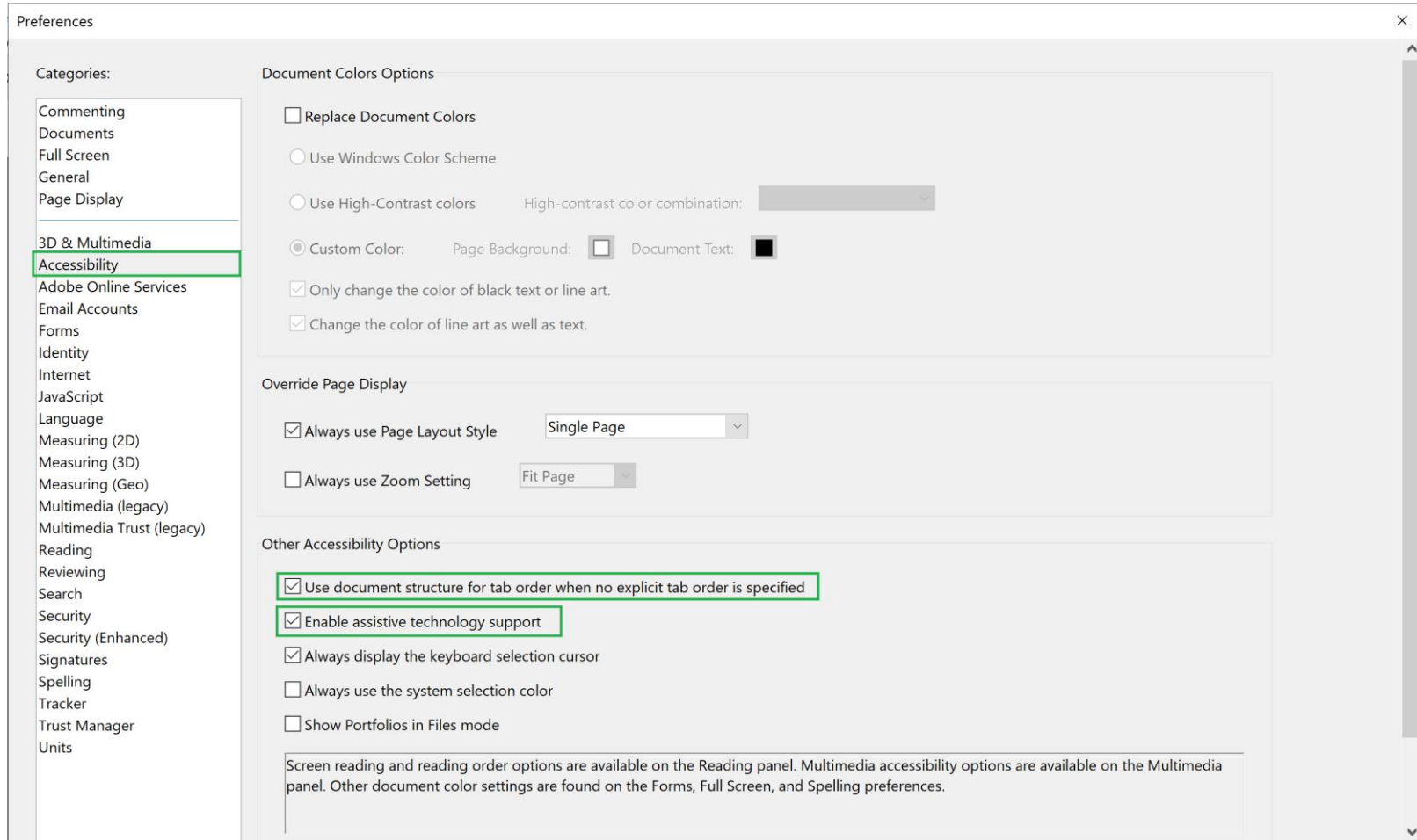
# PDF File Processing. Preferences Settings

- Steps to be performed before working with PDF files in Acrobat Reader DC (Document Cloud) (see *next slide*):
  1. start Acrobat Reader DC (AR DC);
  2. open the **Preferences** window:
    - **Edit** menu --> **Preferences...** option or
    - shortcut **Ctrl +K**;
  3. select **Reading** section:
    - option **Reading Order** = "Infer reading order from document (recommended)";
    - option **Page vs Document** = 'Read the entire document';
    - option **Confirm before tagging documents** = *unchecked*;
  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



# PDF File Processing. Preferences – Accessibility Option



# 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 version 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

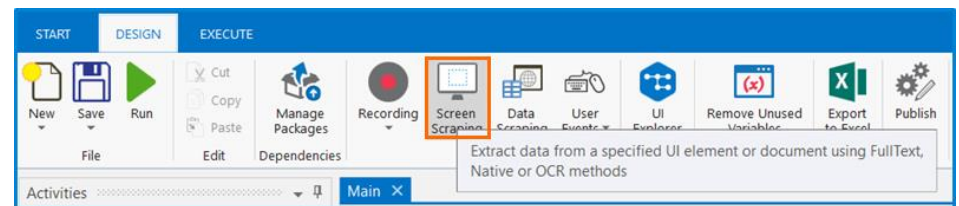
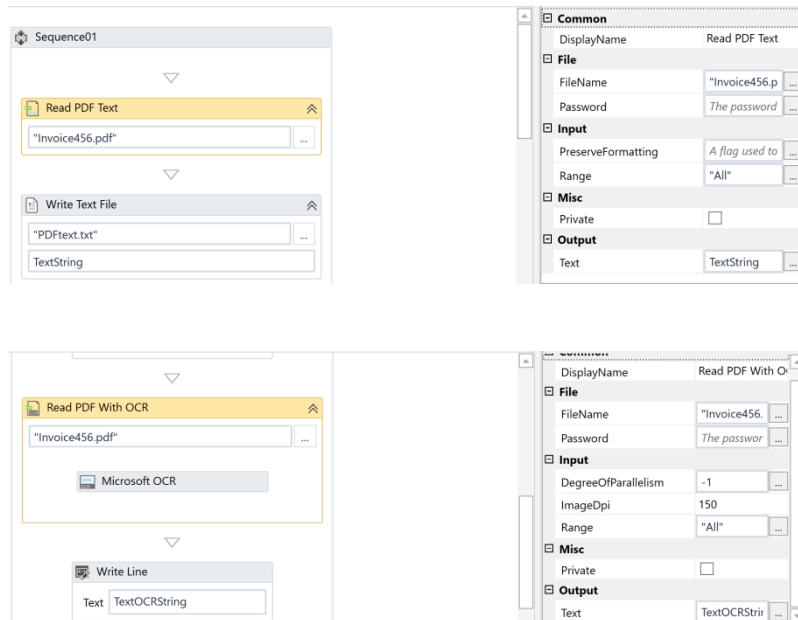
- PDF files consists of data presented as:
  - text;
  - images;
  - text undercover images;
- in order to identify which element is text or image, the element is selected and:
  - if it can be easily selected ==> text;
  - if an apparent block is selected ==> image;
- in UiPath there are activities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



# PDF File Processing. Extract Text Overview

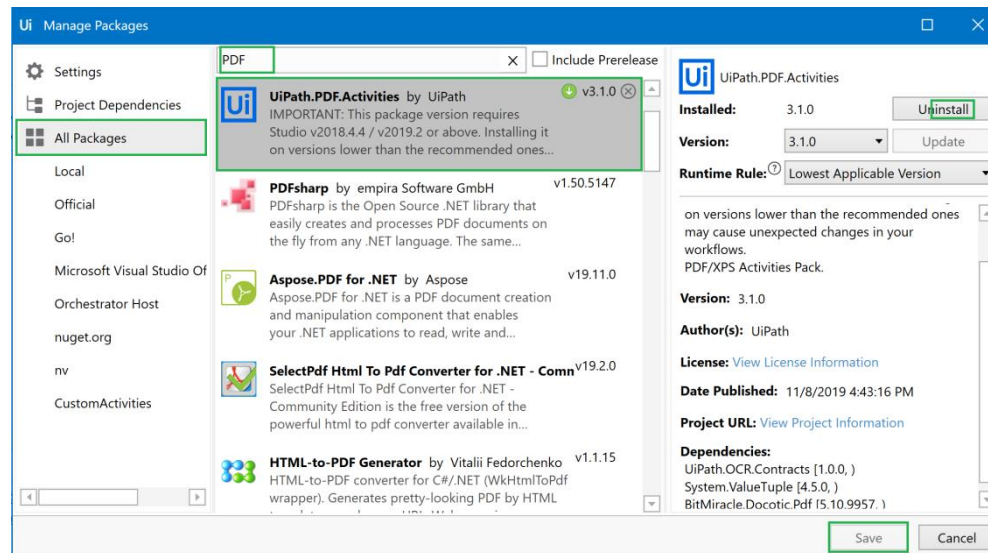
1. **Extracting text** methods include the following activities:

- **Read PDF Text;**
- **Read PDF with OCR;**
- **Screen Scraping:**
  - **Attach Window;**
  - **Get Full Text/ Get Visible Text/ Get OCR Text.**



# PDF File Processing. UiPath Activity Packages

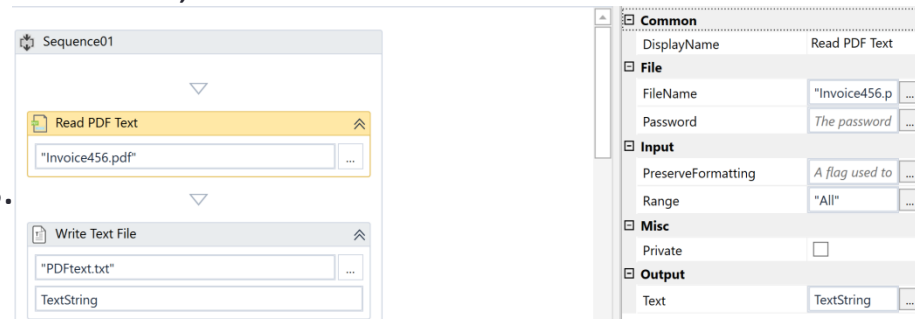
- in order to work with specific PDF activities, some packages need to be installed;
- Steps;
  1. Search in the Activities Panel for 'PDF';
  2. If the result is empty the PDF activity package need to be installed;
  3. In Manage Packages, All Packages section, search for 'PDF';
  4. install **UiPath.PDF.Activities**;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



# Read PDF Text Activity. Details

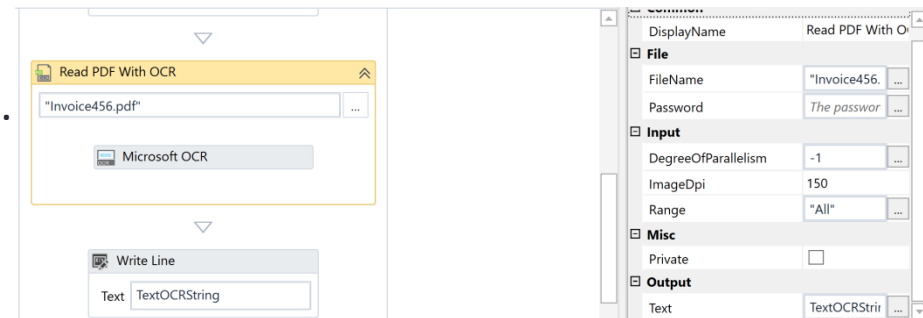
- **Read PDF Text** activity
  - is used to read text from PDF files, as whole document or part of it, i.e., pages;
  - is part of the **UiPath.PDF Activities** package;
- only the text part of the document is processed;
- the image ignored; the result contains the place holder: <Text & Image PDF>;
- 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.



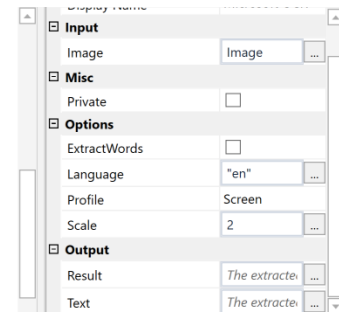
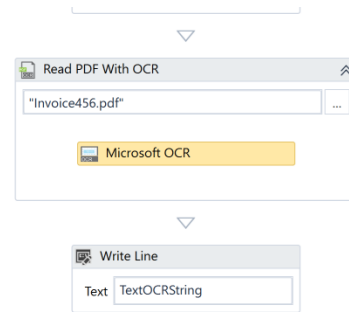
# Read PDF With OCR Activity. Details

- **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.



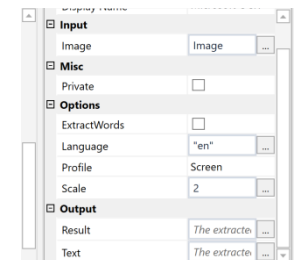
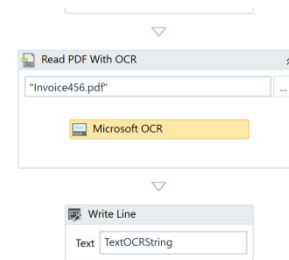
# Read PDF With OCR Activity. OCR Engine (1)

- **OCR Engine** associated to **Read PDF with OCR** allow to customize the used OCR engine;
- various OCR engines are used:
  - **Tesseract, Microsoft, Abbyy**, 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 Activity. OCR Engine (2)

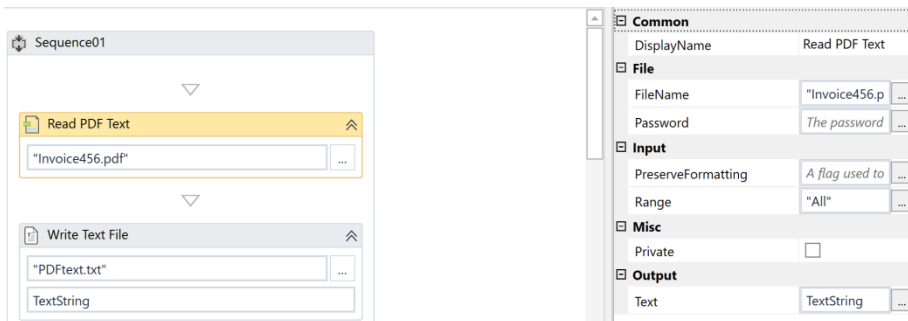
- **OCR Engine** allows to convert the entire text and image to text;
- if the text is placed in two columns the text is intertwined together;
  - this is because the OCR engines do not automatically recognize the 2-column layout of the document;
- **Abbyy OCR Engine** is an exception;
  - it preserves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



# Read PDF Text Activity vs Read PDF With OCR Activity

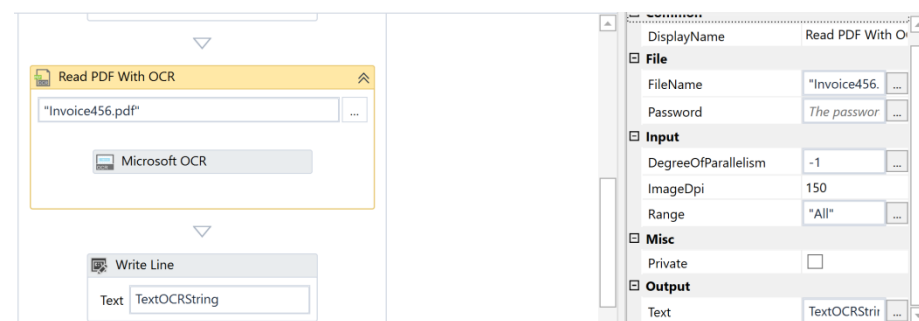
## Read PDF Text activity

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



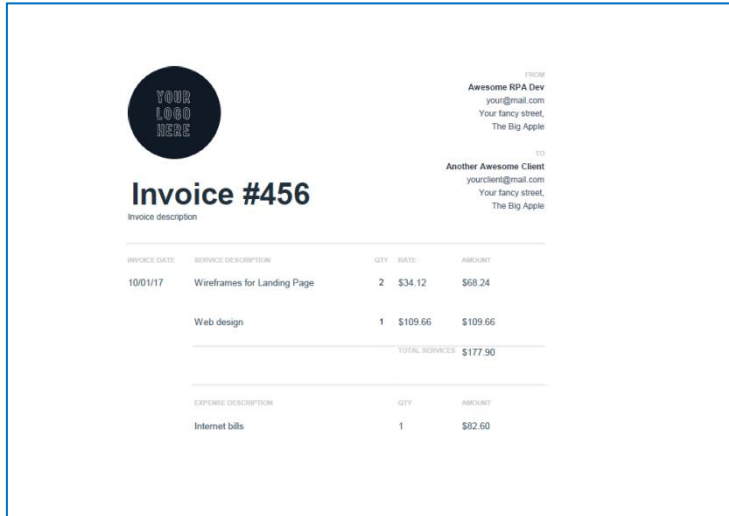
## Read PDF With OCR activity

- 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.*



# Demo 1. 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

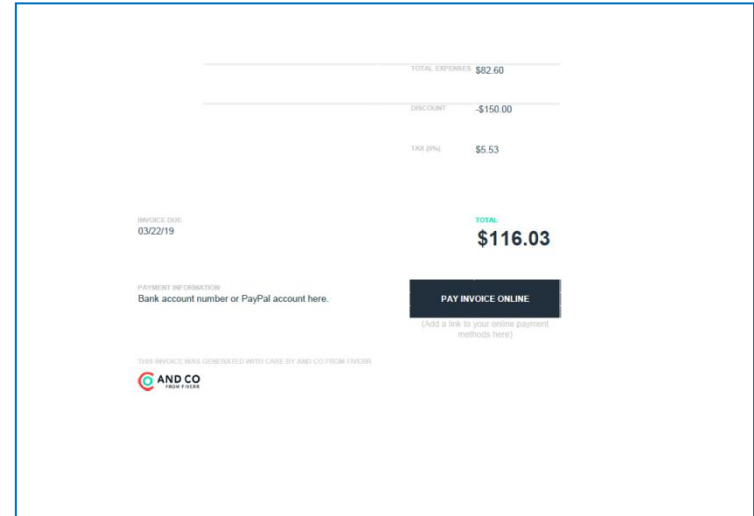
Awesome RPA Dev  
your@mail.com  
Your fancy street,  
The Big Apple

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 (9%) \$5.53

INVOICE DATE 03/22/19

**TOTAL \$116.03**

PAYMENT INFORMATION  
Bank account number or PayPal account here.

**PAY INVOICE ONLINE**  
(Click to visit the online payment gateway interface there)

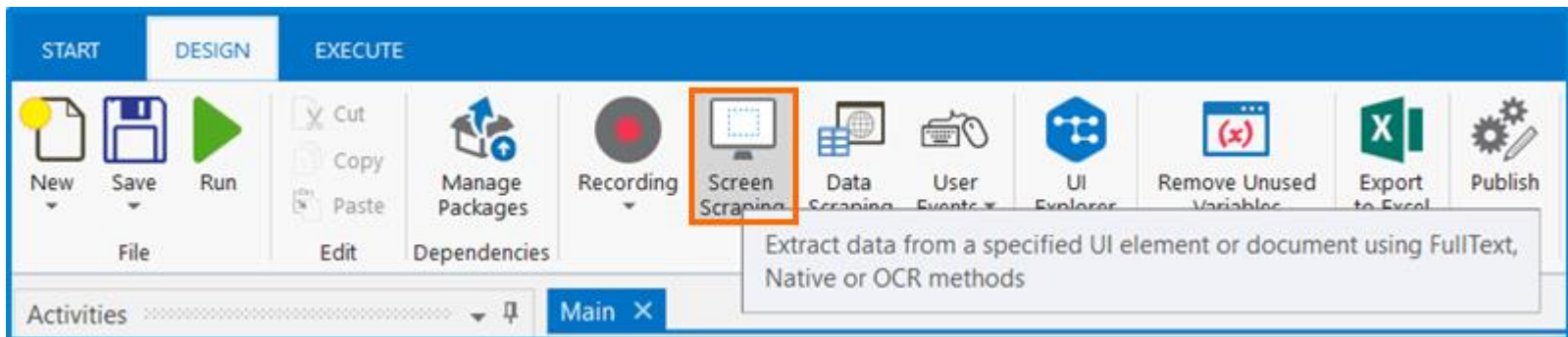
THIS INVOICE WAS GENERATED WITH CARE BY AND CO FROM FINLAND

**AND CO**  
FINLAND



# Screen Scraping. Details

- **Screen Scraping** was mainly covered in Lecture 05. 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.



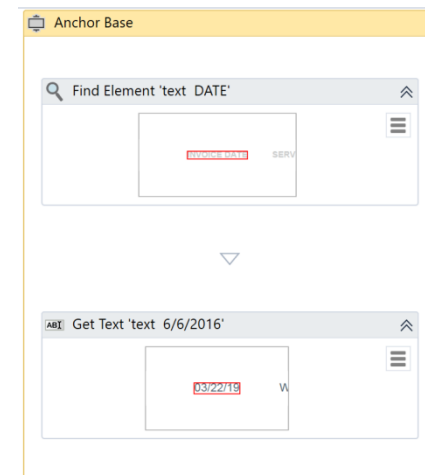
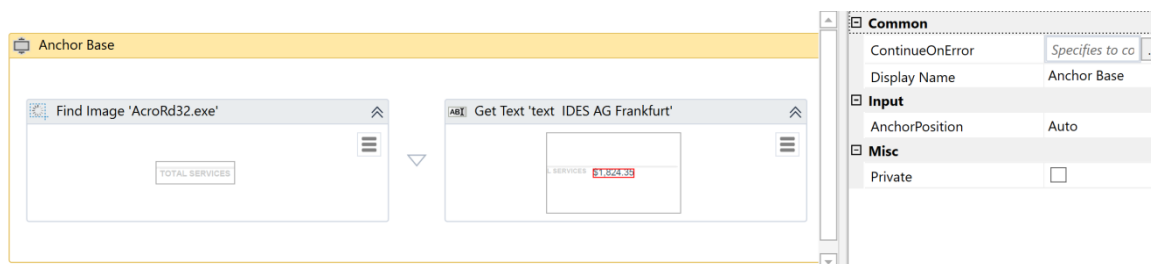
# PDF File Processing. Extract Specific Elements

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 05, Lecture 06**;
- **Find Relative Element**.

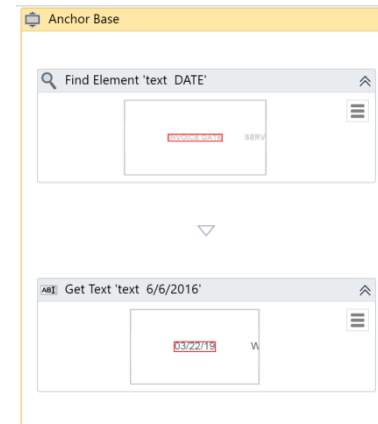
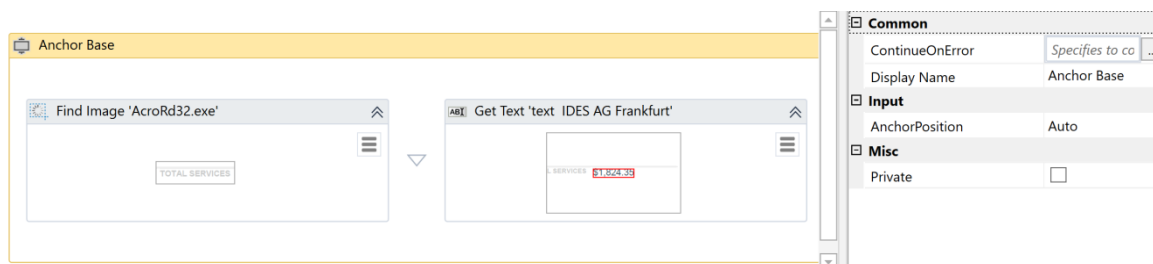
# Anchor Base Activity. Details

- **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 is placed in relation with the targeted text.



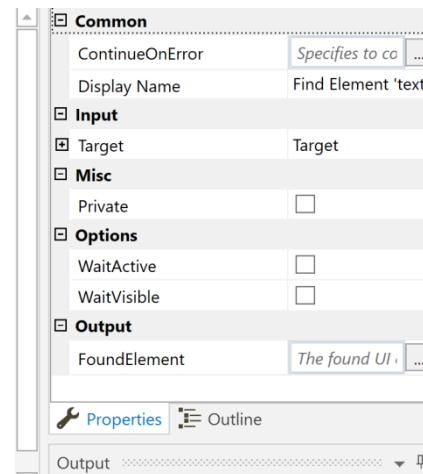
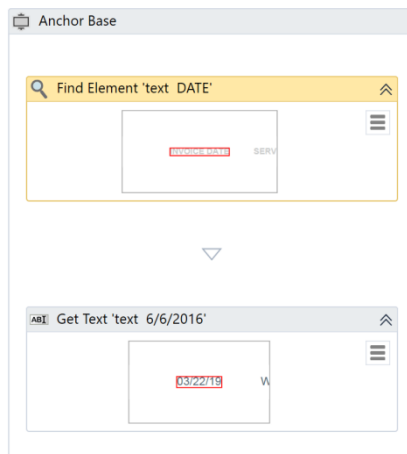
# Anchor Base Activity. Actions

- **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.



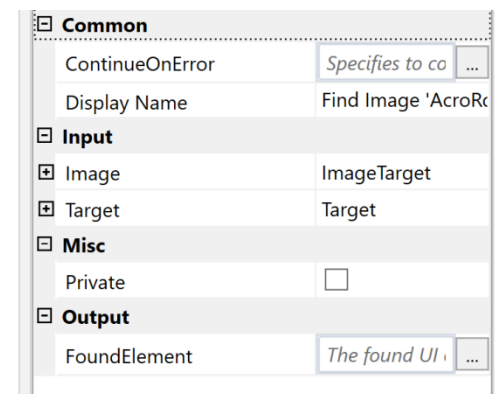
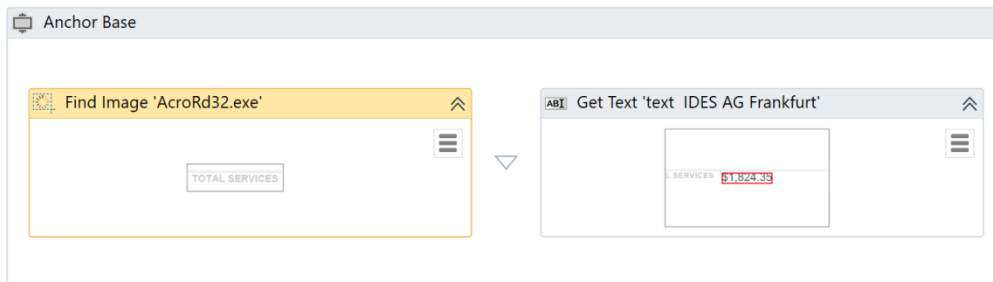
# Anchor Base Activity. Find Element Activity

- **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.



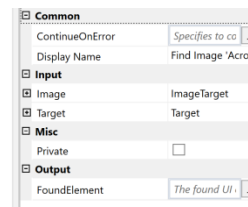
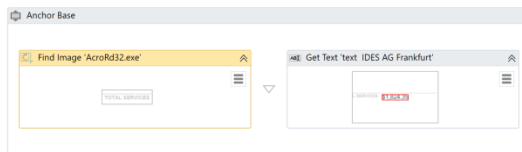
# Anchor Base Activity. Find Image Activity (1)

- **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 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;
- relevant properties:
  - **[output] Found Element** = **UiElement** variable;
    - the identified image element that can be used in further processing.



# Anchor Base Activity. Find Image Activity (2)

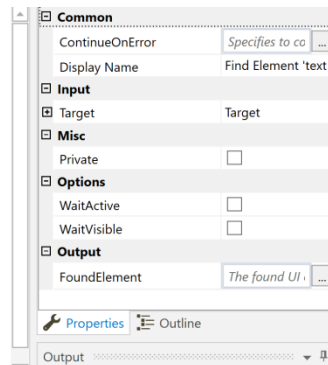
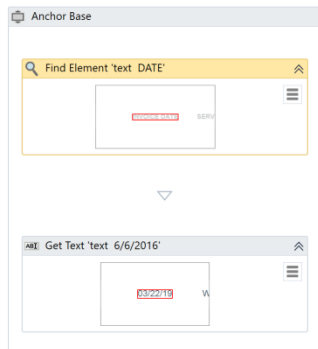
- **Find Image** activity
  - allows to identify a fixed image in the document that will be used as an anchor to perform an action on another element;
- useful for PDF documents because similar PDF documents look similar;
- image accuracy can be affected by the size of the page of the opened document;
- 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.



# Find Element Activity vs Find Image Activity

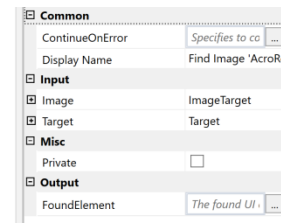
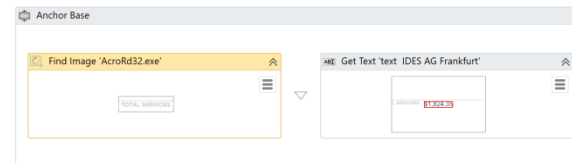
## Find Element Activity

- 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.



## 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.





# Demo 2. 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.

# References

- UiPath Academy - <https://academy.uipath.com>
  - Level 1 – Foundation Training, Lesson 10;
- 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>