

Practice Exercise: Extracting data

Extracting data from forms

A step-by-step detailed guide with instructions.

Exercise: Extracting Data



Extracting data from forms

What to expect

In this exercise, you will learn how to use the intelligent OCR activities to extract and validate data from two different forms. You will gain practical experience with:

1. **Process Design:** Taxonomy Manager
2. **Pre-Processing:** Loading the taxonomy and input files
3. **Digitization:** Digitize document settings, OCR engine selection, OCR settings
4. **Classification:** Intelligent Keyword Classifier, initial training
5. **Classification Validation:** Classification Station
6. **Classifier Training:** Train Classifiers Scope
7. **Extraction:** Form Extractor and Regex Based Extractor
8. **Validation:** Validation Station
9. **Data Export:** Export of validated results

Note

When working on the hands-on exercises, please be aware that the screenshots and instructions provided may vary slightly depending on the version of UiPath Studio you are using. To ensure a smooth experience, we recommend following the installation instructions and using the appropriate version specified in the course material.

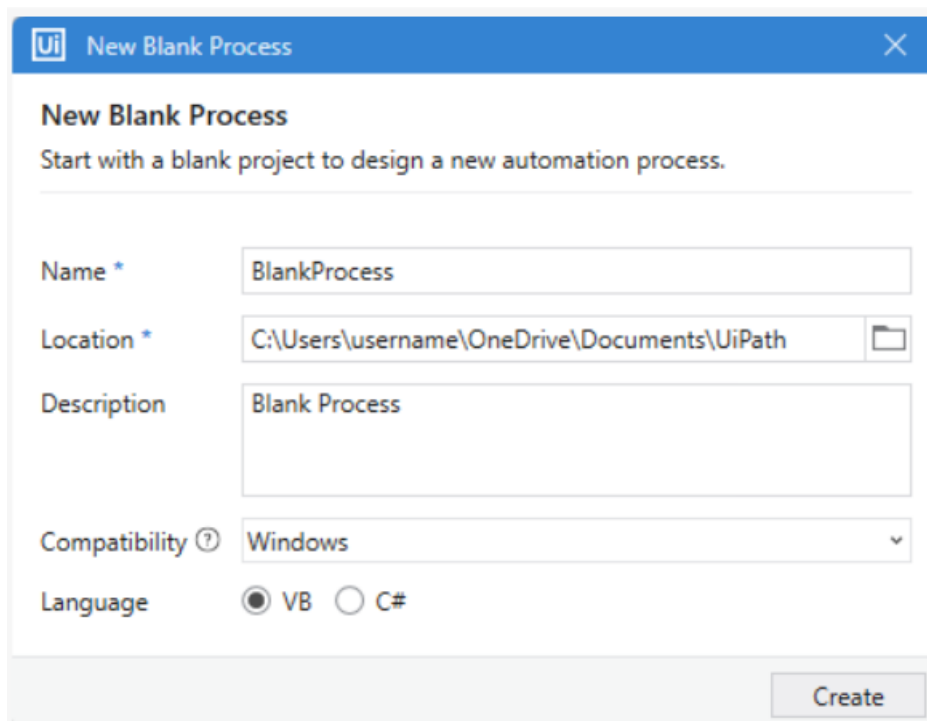
Getting started

- Create a UiPath process and extract the data requested by the business, by leveraging the Document Understanding concepts you learned today.
- Please download the Inputs.zip file containing the sample forms to get started.

Task 1: Create a blank process/project in UiPath Studio

Steps:

1. Open UiPath Studio.
2. To create a new process, on the ribbon, click **HOME** and select **Process**. The New Blank Process dialog is displayed.



3. In the Name field, type a suitable name for the new project.
4. In the Description field, type a suitable project description.

5. Click **Create**. The new project is opened in Studio.

Task 2: Install the required activities packages

Steps:

1. On the ribbon of Studio, click **DESIGN** and select **Manage Packages**.
2. Select the **Include Prerelease** check box.
3. Install the following packages:
 - UiPath.IntelligentOCR.Activities (min version 6.5.0)
 - UiPath.DocumentUnderstanding.ML.Activities (min version 1.17.0)
 - UiPath.System.Activities (min version 22.10.3)
 - UiPath.UIAutomation.Activities (min version 22.10.3)
 - UiPath.Excel.Activities (min version 2.16.0)

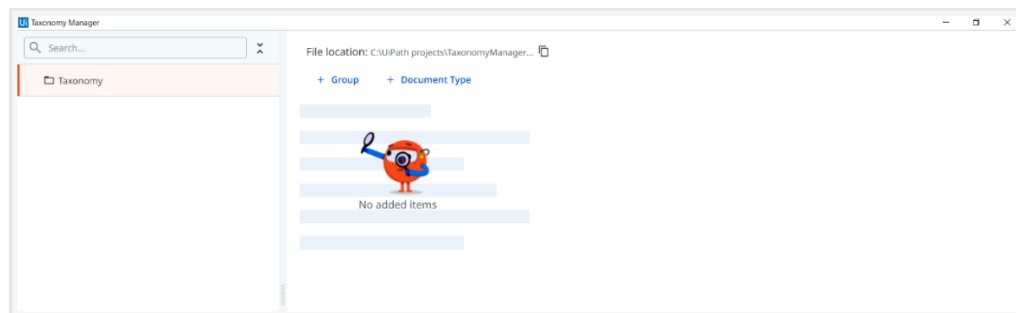
Task 3: Create taxonomy

Steps:

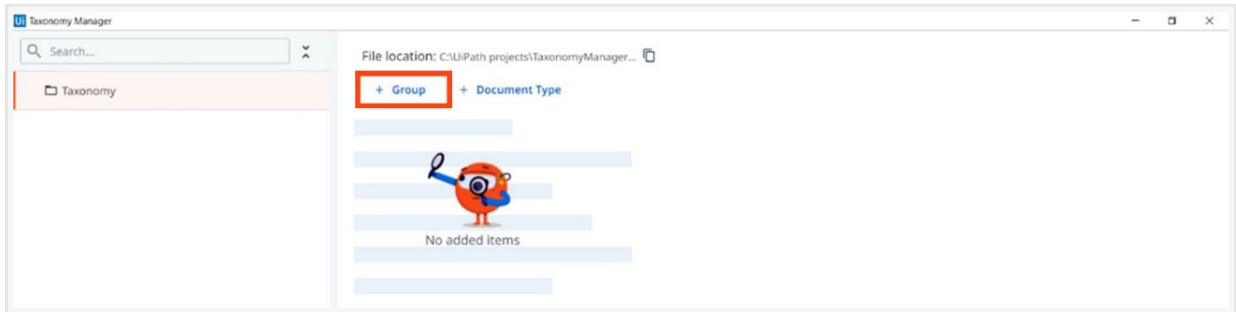
1. In the DESIGN tab, click **Taxonomy Manager**.



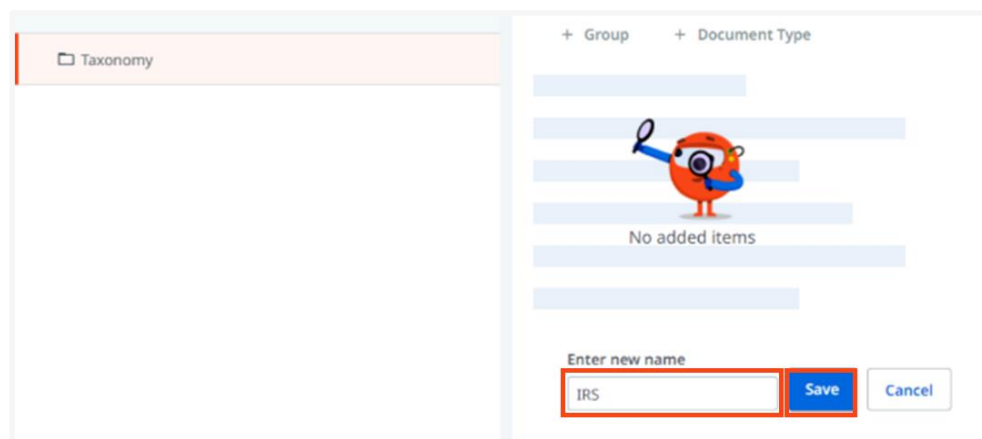
The Taxonomy Manager dialog is displayed. This option is available only when you successfully install the UiPath.IntelligentOCR.Activities package (v 6.5.0 or higher).



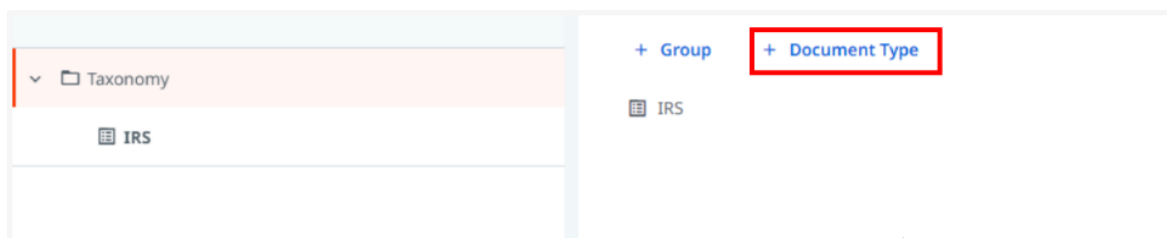
2. To create a new group named IRS, in the right panel, click **+ Group**.



3. In the Enter new name field, type the group name as **IRS** and click **Save**.

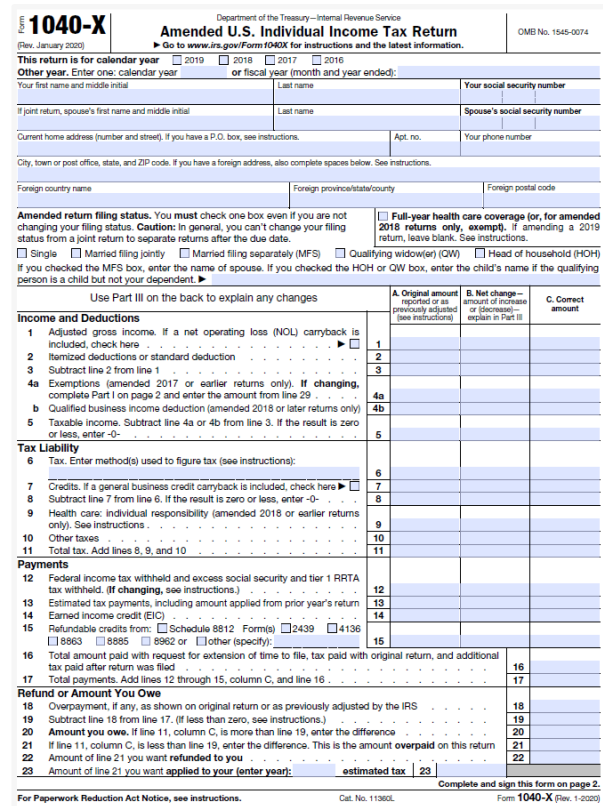


4. Within the newly created IRS group, similarly, create a new category named Forms by clicking + Category in the right panel.
5. Within the newly created category, create a new document type named Form1040x by clicking + Document Type in the right panel.



6. Within the Form1040x document type, to add the fields to be extracted, click **+ Fields** in the right panel.

- Return 2019 (Boolean)
- Return 2018 (Boolean)
- Return 2017 (Boolean)
- Return 2016 (Boolean)
- First Name (text)
- Last Name (text)
- Social Security Number (text)
- Spouses First Name (text)
- Spouses Last Name (text)
- Spouses Social Security Number (text)
- Address Line 1 (text)
- Apartment Number (text)
- Phone Number (text)
- City Town Zip (text)
- Filing Status Single (Boolean)
- Filing Status Married Filing Jointly (Boolean)
- Filing Status Married Filing Separately (Boolean)
- Filing Status Qualifying Widower (Boolean)
- Filing Status Head Of Household (Boolean)



Form 1040-X Department of the Treasury—Internal Revenue Service
Amended U.S. Individual Income Tax Return OMB No. 1545-0074
 (Rev. January 2020) ▶ Go to www.irs.gov/Form1040X for instructions and the latest information.

This return is for calendar year ☐ 2019 ☐ 2018 ☐ 2017 ☐ 2016
 Other year. Enter one: calendar year or fiscal year (month and year ended):

Your first name and middle initial Last name Your social security number
 If joint return, spouse's first name and middle initial Last name Spouse's social security number

Current home address (number and street). If you have a P.O. box, see instructions. Apt. no. Your phone number
 City, town or post office, state, and ZIP code. If you have a foreign address, also complete spaces below. See instructions.

Foreign country name Foreign province/state/country Foreign postal code

Amended return filing status. You must check one box even if you are not changing your filing status. **Caution:** In general, you can't change your filing status from a joint return to separate returns after the due date.
☐ Single ☐ Married filing jointly ☐ Married filing separately (MFS) ☐ Qualifying widow(er) (QW) ☐ Head of household (HOH)
 If you checked the MFS box, enter the name of spouse. If you checked the HOH or QW box, enter the child's name if the qualifying person is a child but not your dependent. ▶

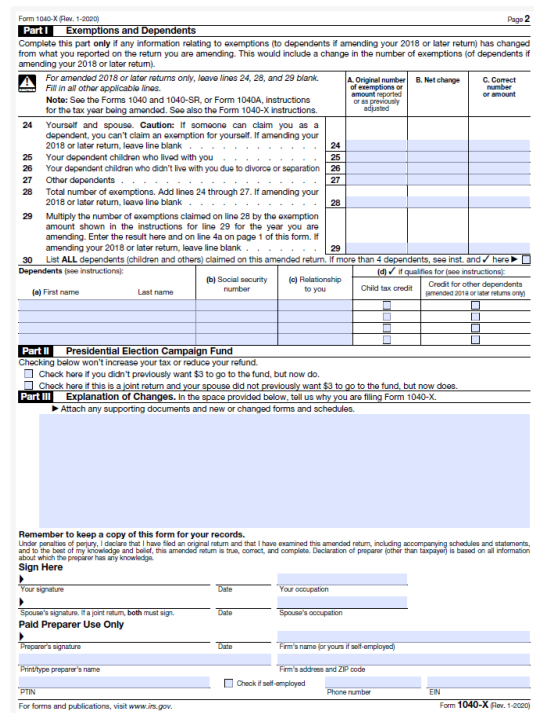
☐ **Full-year health care coverage (or, for amended 2018 returns only, exempt).** If amending a 2019 return, leave blank. See instructions.

Use Part III on the back to explain any changes

	A. Original amount reported or as previously adjusted (see instructions)	B. Net change—amount of increase or decrease—explain in Part III	C. Correct amount
Income and Deductions			
1 Adjusted gross income. If a net operating loss (NOL) carryback is included, check here ▶ <input type="checkbox"/>	1		
2 Itemized deductions or standard deduction ▶ <input type="checkbox"/>	2		
3 Subtract line 2 from line 1	3		
4a Exemptions (amended 2017 or earlier returns only). If changing, complete Part I on page 2 and enter the amount from line 29	4a		
b Qualified business income deduction (amended 2018 or later returns only)	4b		
5 Taxable income. Subtract line 4a or 4b from line 3. If the result is zero or less, enter -0-	5		
Tax Liability			
6 Tax. Enter method(s) used to figure tax (see instructions):	6		
7 Credits. If a general business credit carryback is included, check here ▶ <input type="checkbox"/>	7		
8 Subtract line 7 from line 6. If the result is zero or less, enter -0-	8		
9 Health care: individual responsibility (amended 2018 or earlier returns only). See instructions	9		
10 Other taxes	10		
11 Total tax. Add lines 8, 9, and 10	11		
Payments			
12 Federal income tax withheld and excess social security and tier 1 RRRTA tax withheld. (If changing, see instructions.)	12		
13 Estimated tax payments, including amount applied from prior year's return	13		
14 Earned income credit (EIC)	14		
15 Refundable credits from: <input type="checkbox"/> Schedule 8812 <input type="checkbox"/> Form(s) <input type="checkbox"/> 2439 <input type="checkbox"/> 4136 <input type="checkbox"/> 8863 <input type="checkbox"/> 8885 <input type="checkbox"/> 8962 or <input type="checkbox"/> other (specify):	15		
16 Total amount paid with request for extension of time to file, tax paid with original return, and additional tax paid after return was filed	16		
17 Total payments. Add lines 12 through 15, column C, and line 16	17		
Refund or Amount You Owe			
18 Overpayment, if any, as shown on original return or as previously adjusted by the IRS	18		
19 Subtract line 18 from line 17. (If less than zero, see instructions.)	19		
20 Amount you owe. If line 11, column C, is more than line 19, enter the difference	20		
21 If line 11, column C, is less than line 19, enter the difference. This is the amount overpaid on this return	21		
22 Amount of line 21 you want refunded to you	22		
23 Amount of line 21 you want applied to your (enter year): estimated tax 23	23		

For Paperwork Reduction Act Notice, see instructions. Cat. No. 11360L Form **1040-X** (Rev. 1-2020) Complete and sign this form on page 2.

- Dependents (table)
 - First Last Name (column - name)
 - Social Security Number (column - text)
 - Relationship To You (column - text)
 - Child Tax Credit (column - Boolean)
 - Credit For Other Dependents (column - Boolean)
- Signature (Boolean)
- Date (date)
- Spouses Signature (Boolean)
- Spouses Date (date)



Form 1040-X (Rev. 1-2020) Page 2

Part I Exemptions and Dependents

Complete this part only if any information relating to exemptions (to dependents if amending your 2018 or later return) has changed from what you reported on the return you are amending. This would include a change in the number of exemptions (or dependents if amending your 2018 or later return).

Note: See the Forms 1040 and 1040-SR, or Form 1040A, instructions for the tax year being amended. See also the Form 1040-X instructions.

For amended 2018 or later returns only, leave lines 24, 28, and 29 blank.

Fill in all other applicable lines.

	A. Original number of exemptions or amount reported or as previously adjusted	B. Net change	C. Correct number or amount
24 Yourself and spouse. Caution: If someone can claim you as a dependent, you can't claim an exemption for yourself. If amending your 2018 or later return, leave line blank.	24		
25 Your dependent children who lived with you.	25		
26 Your dependent children who didn't live with you due to divorce or separation.	26		
27 Other dependents.	27		
28 Total number of exemptions. Add lines 24 through 27. If amending your 2018 or later return, leave line blank.	28		
29 Multiply the number of exemptions claimed on line 28 by the exemption amount shown in the instructions for line 29 for the year you are amending. Enter the result here and on line 4a on page 1 of this form. If amending your 2018 or later return, leave line blank.	29		

30 List ALL dependents (children and others) claimed on this amended return. If more than 4 dependents, see inst. and ✓ here ▶

Dependents (see instructions):

(a) First name	Last name	(b) Social security number	(c) Relationship to you	(d) ✓ if qualifies for (see instructions):
				Child tax credit
				Credit for other dependents (amended 2018 or later returns only)

Part II Presidential Election Campaign Fund

Checking below won't increase your tax or reduce your refund.

☐ Check here if you didn't previously want \$3 to go to the fund, but now do.

☐ Check here if this is a joint return and your spouse did not previously want \$3 to go to the fund, but now does.

Part III Explanation of Changes. In the space provided below, tell us why you are filing Form 1040-X.

▶ Attach any supporting documents and new or changed forms and schedules.

Remember to keep a copy of this form for your records.

Under penalties of perjury, I declare that I have filed an original return and that I have examined this amended return, including accompanying schedules and statements, and to the best of my knowledge and belief, this amended return is true, correct, and complete. Declaration of preparer (other than taxpayer) is based on all information about which the preparer has any knowledge.

Sign Here

Your signature _____ Date _____ Your occupation _____

Spouse's signature, if a joint return, both must sign _____ Date _____ Spouse's occupation _____

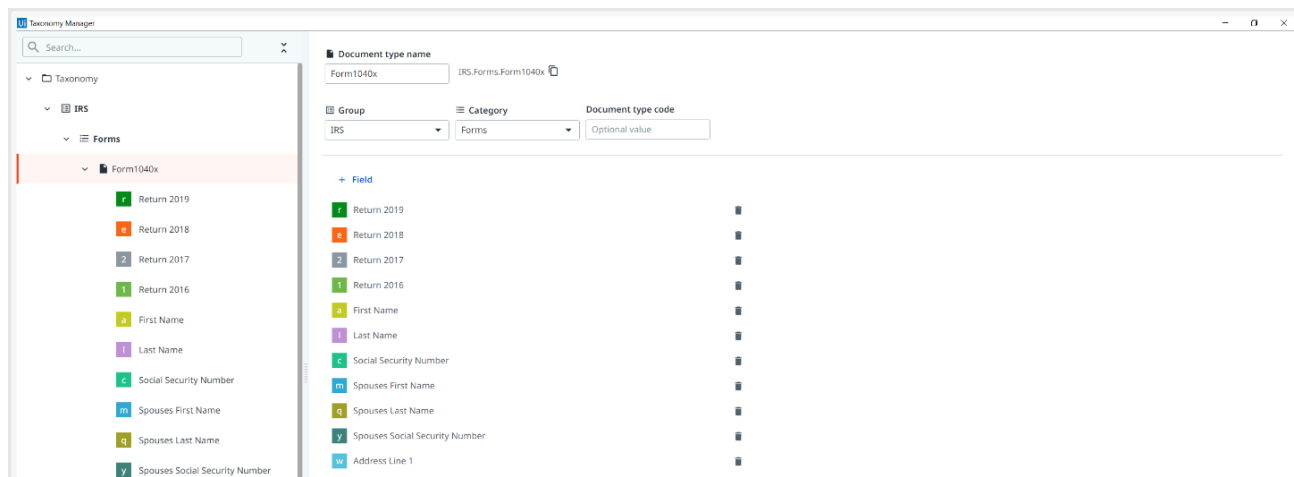
Paid Preparer Use Only

Preparer's signature _____ Date _____ Firm's name (or yours if self-employed) _____

Print/type preparer's name _____ Firm's address and ZIP code _____

PTIN _____ ☐ Check if self-employed _____ Phone number _____ EN _____

For forms and publications, visit www.irs.gov. Form **1040-X** (Rev. 1-2020)



Taxonomy Manager

Search...

Document type name
Form1040x IRS.Forms.Form1040x

Group IRS **Category** Forms **Document type code** Optional value

Field

- Return 2019
- Return 2018
- Return 2017
- Return 2016
- First Name
- Last Name
- Social Security Number
- Spouses First Name
- Spouses Last Name
- Spouses Social Security Number
- Address Line 1

7. Create another new group named Insurance, a new category named Reports, and a new document type called VehicleDamageReport.
8. Add the fields to be extracted for the new document type.
 - Registration Number (text)
 - Inspector Name (name)
 - Vehicle License Number (text)

- Driving License Number (text)
- Cost Of Damage (number)

Vehicle Damage Report Form

General Information

Title of Company/Organization: Oberbrunner Inc
Registration Number: 99-4444444
Address: 4359 Pine Tree Lane **City:** Gaithersburg **Zip Code:** 20877 **State:** MD

Report prepared by: Mary Brown
Designation: damage inspector **Phone:** 240-793-1698

Damage Report Information

Type of Vehicle Damaged: motor vehicle (car), 2012 GMC Canyon
Date of damage incident/accident: 01.13.2021
Location of damage: 4360 Pine Tree Lane
Vehicle License Number: FBK 1569
Driver Name: Mary Brown **Driving License Number:** H-630-831-483-501

Nature of Damage: -
Brief Description of Damage: Front bumper
Is the damage (Minor/Major): Minor
Any casualty of human life: No
Description of Injuries: None
Medical Treatment for Injuries: No
Person involved in damage: No
Activities of the above person in damage: None
Witnesses: None

Is the vehicle insured, if yes provide details of insurance cover

Name of Insurance Company: GEICO
Cost of Damage: \$2210

Taxonomy Manager

- ▼ Taxonomy
 - ▼ IRS
 - Forms
 - Form1040x
 - ▼ Insurance
 - Reports
 - Vehicle Damage Report

r Registration Number
 e Inspector Name
 v Vehicle License Number
 g Driving License Number
 c Cost Of Damage

Document type name

Insurance.Reports.VehicleDamageRep...

Group

Insurance ▼

Category

Reports ▼

Document type code

Optional value

+ Field

r Registration Number
 ✕

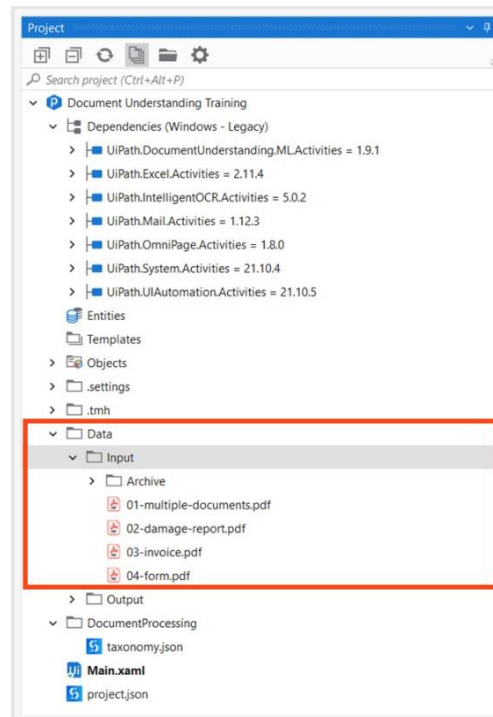
e Inspector Name
 ✕

v Vehicle License Number
 ✕

g Driving License Number
 ✕

c Cost Of Damage
 ✕

9. Extract the contents of the downloaded Inputs.zip file and move them to the Input folder under the Data folder of the project.

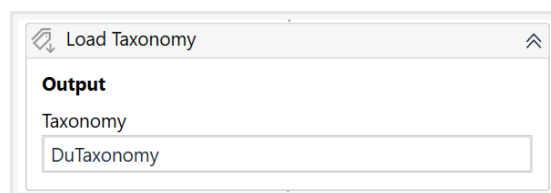


10. In the project folder that you are working in, create a folder for exporting the extracted data to Data\Output.

Task 4: Configure the process

Steps:

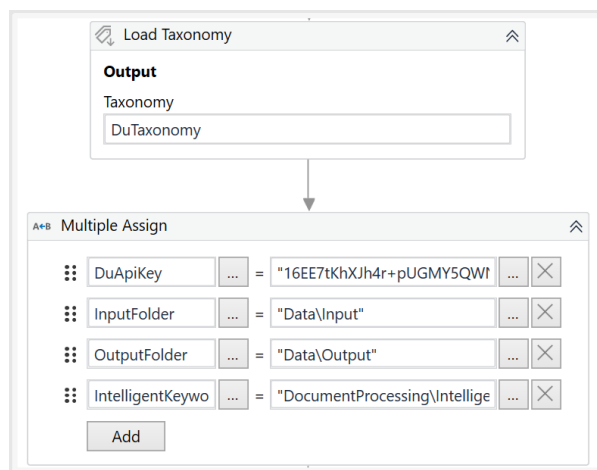
1. From the Activity panel, drag the Load Taxonomy activity and drop it into the Main workflow.
2. Set the Taxonomy property to a suitable variable to receive the activity's output.



3. Add the Multiple Assign activity to the Main workflow.

4. Inside the Multiple Assign activity:

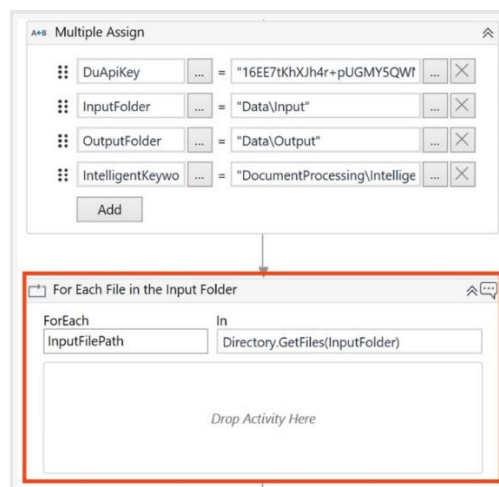
- Create a variable of type String, called DuApiKey, and assign it the value of your Document Understanding API Key (can be found in the Automation Cloud, Admin → Licenses → Robots & Services → Document Understanding).
- Create a variable of type String, called InputFolder, and assign the Data\Input value to it.
- Create a variable of type String, called OutputFolder, and assign the Data\Output value to it.
- Create a variable of type String, called IntelligentKeywordClassifierLearningFile, and assign it the path of the learning file as the value (learning file that should be an empty json).



5. Add and configure the For Each activity.

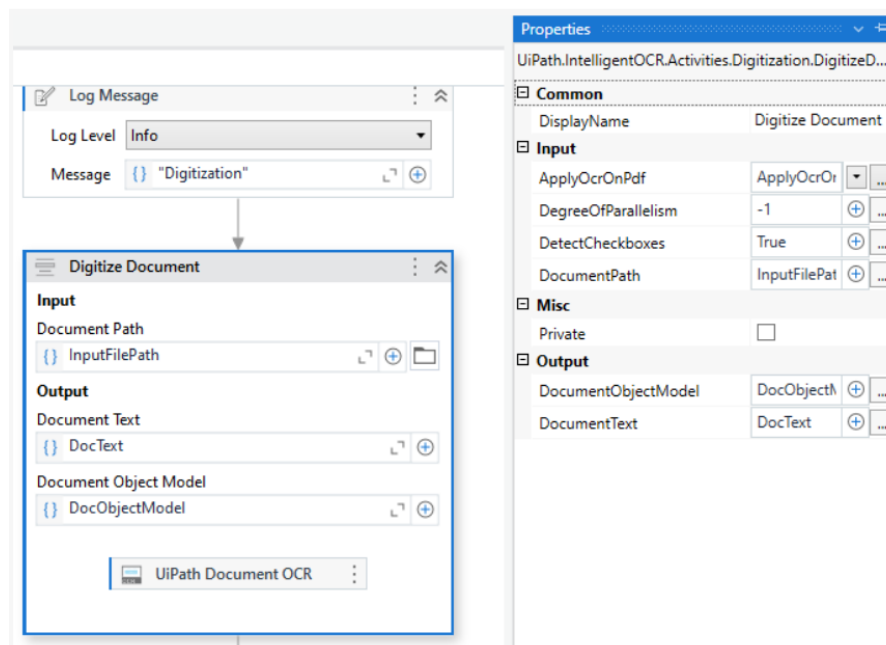
6. In the Property panel of For Each activity, set the TypeArgument property to String.

7. To configure the activity to iterate through all the paths in the Input folder, type **InputFilePath** in the ForEach field and **Directory.GetFiles(InputFolder)** in the In field.



Task 5: Digitize the documents

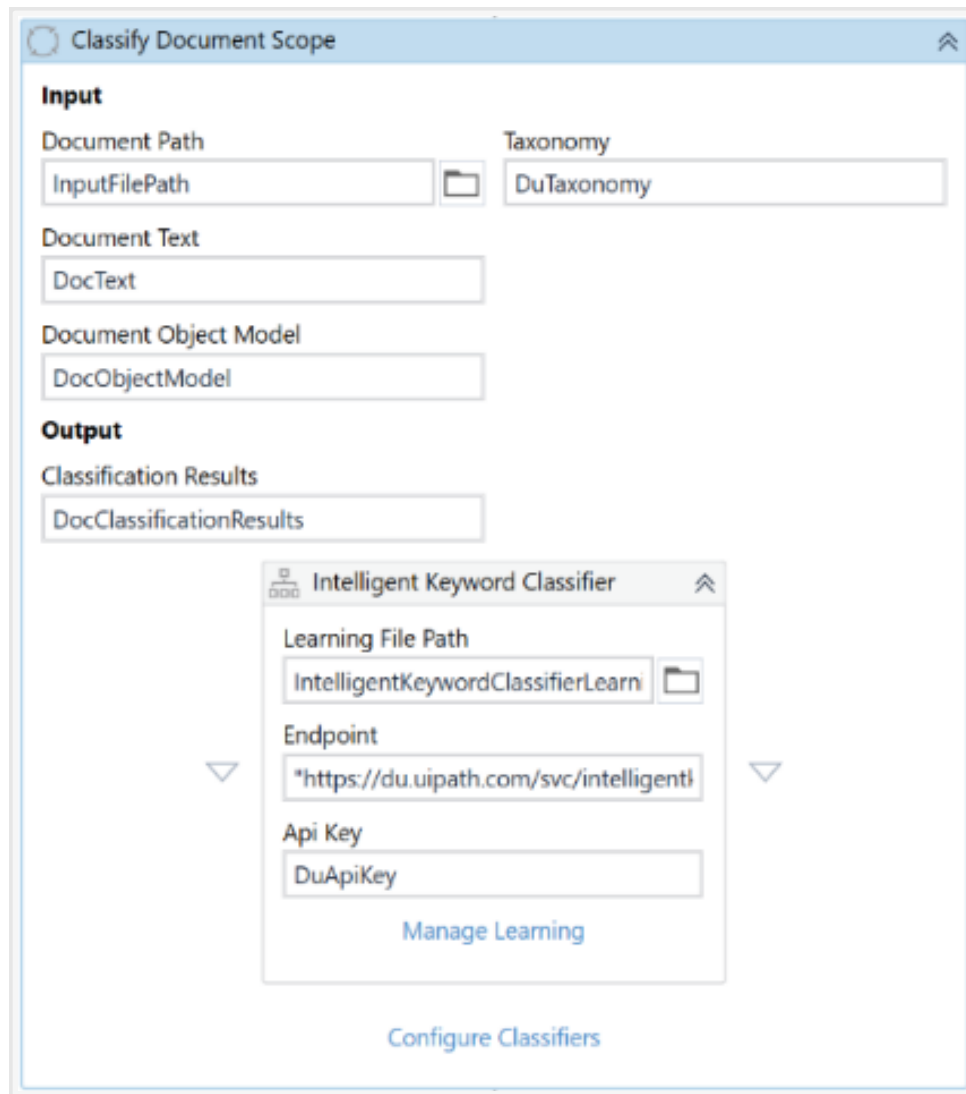
1. Add the Digitize Document activity to the For Each body.
2. To configure the Digitize Document activity:
 - i. Set the Document Path property to InputFilePath.
 - ii. Create output variables for DocumentObjectModel and DocumentText properties.
 - iii. Add the UiPath Document OCR activity to the OCR Engine section if not added by default.
 - iv. Select the UiPath Document OCR activity and configure the **API Key** and the **Endpoint** properties. Use the following OCR endpoint: <https://du.uipath.com/ocr>



Task 6: Classify the documents based on a specific training set and validate the same


Steps:

1. Add Classify Document Scope and configure Intelligent Keyword Classifier inside it.



Classify Document Scope

Input

Document Path: 

Taxonomy:


Document Text:

Document Object Model:

Output

Classification Results:

Intelligent Keyword Classifier

Learning File Path: 

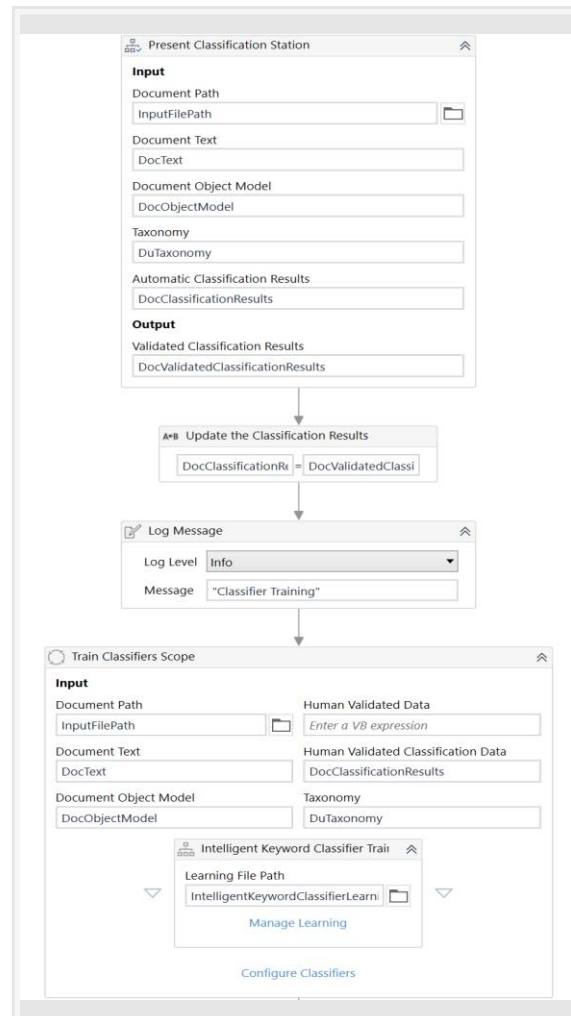
Endpoint:

Api Key:

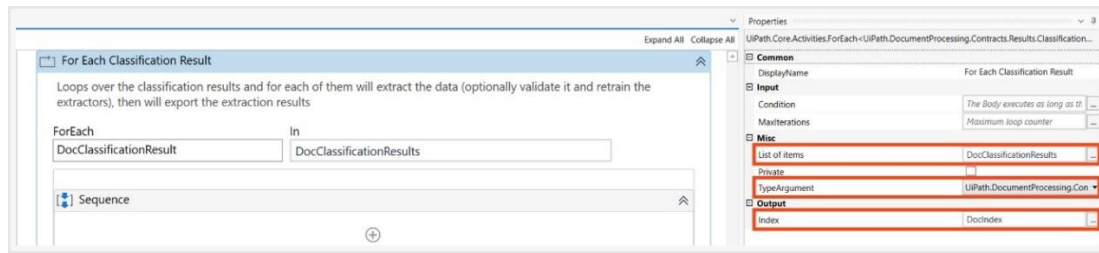
[Manage Learning](#)

[Configure Classifiers](#)

2. Validate the classification results using Present Classification Station. Use the validated data to train the classifier using Train Classifiers Scope and Intelligent Keyword Classifier Trainer.



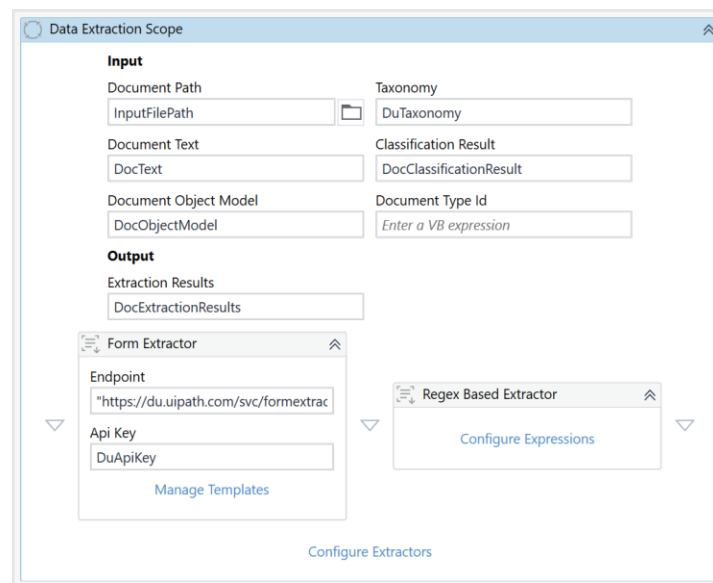
3. To loop through all the classification results, add another For Each activity.
4. To configure this For Each activity, set its:
 - i. TypeArgument property to UiPath.DocumentProcessing.Contracts.Results.ClassificationResult.
 - ii. List of items property to DocClassificationResults.
 - iii. Index property to DocIndex (creating a new Int32 variable).



Task 7: Extract the data using the Form Extractor and the Regex Based Extractor

Steps:

1. Add the Data Extraction Scope activity to the workflow.
2. Add the Form Extractor and the Regex Based Extractor to the Data Extraction Scope activity.



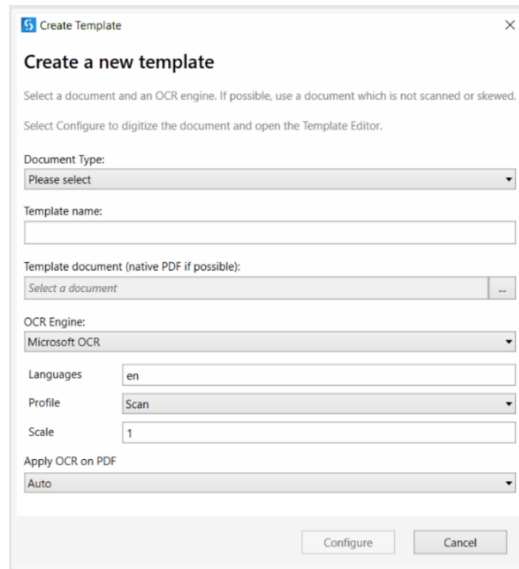
3. To open the Configure Extractors wizard, in the Data Extraction Scope activity, click **Configure Extractors**.

The Configure Extractors wizard is displayed.

4. Set the Form Extractor to handle only the Form1040x documents and the Regex Based Extractor to handle only the Vehicle Damage Report documents.
5. Within Data Extraction Scope, drag and drop the Form Extractor activity.
6. To configure the extractor for creating a new Form1040x template, in the Form Extractor activity, click **Manage Templates**.

The Template Manager window is displayed.

7. To open the Create Template dialog, in the Template Manager window, click **Create Template**. The Create Template dialog is displayed.



Create Template

Create a new template

Select a document and an OCR engine. If possible, use a document which is not scanned or skewed.

Select Configure to digitize the document and open the Template Editor.

Document Type:
Please select

Template name:
[Text Field]

Template document (native PDF if possible):
Select a document [Browse Button]

OCR Engine:
Microsoft OCR

Languages: en

Profile: Scan

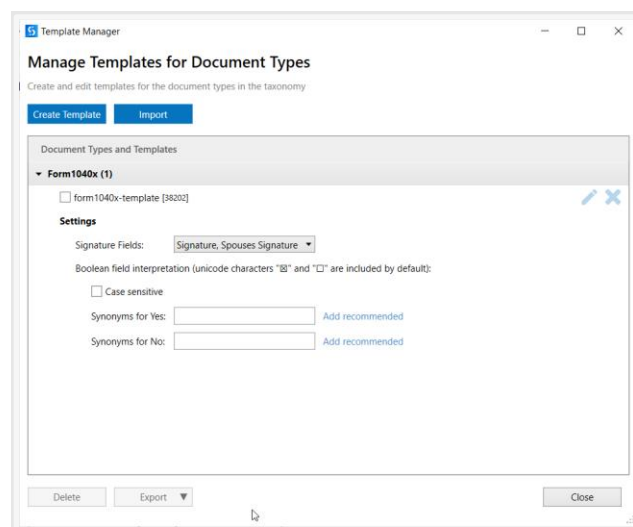
Scale: 1

Apply OCR on PDF: Auto

[Configure] [Cancel]

8. Create a new Form1040x template for the Form Extractor by setting the Document Type, Template name, Template document, and OCR Engine fields. Consider using the digital PDF of the form (NOT a scanned document).

The configured template is displayed, wherein you need to configure Signature Fields.



Template Manager

Manage Templates for Document Types

Create and edit templates for the document types in the taxonomy

[Create Template] [Import]

Document Types and Templates

▼ Form1040x (1)

☐ form1040x-template [38202]

Settings

Signature Fields: Signature, Spouses Signature

Boolean field interpretation (unicode characters "☐" and "☐" are included by default):

☐ Case sensitive

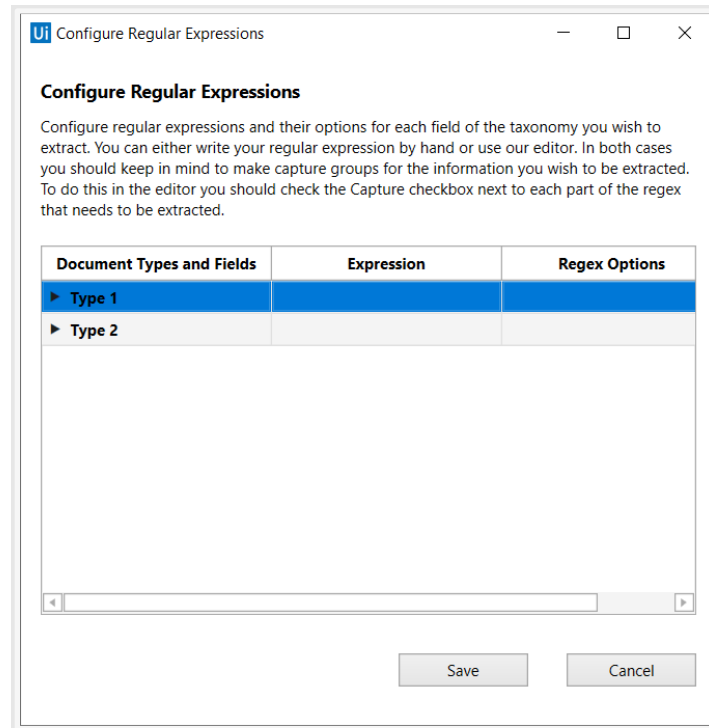
Synonyms for Yes: [Text Field] [Add recommended](#)

Synonyms for No: [Text Field] [Add recommended](#)

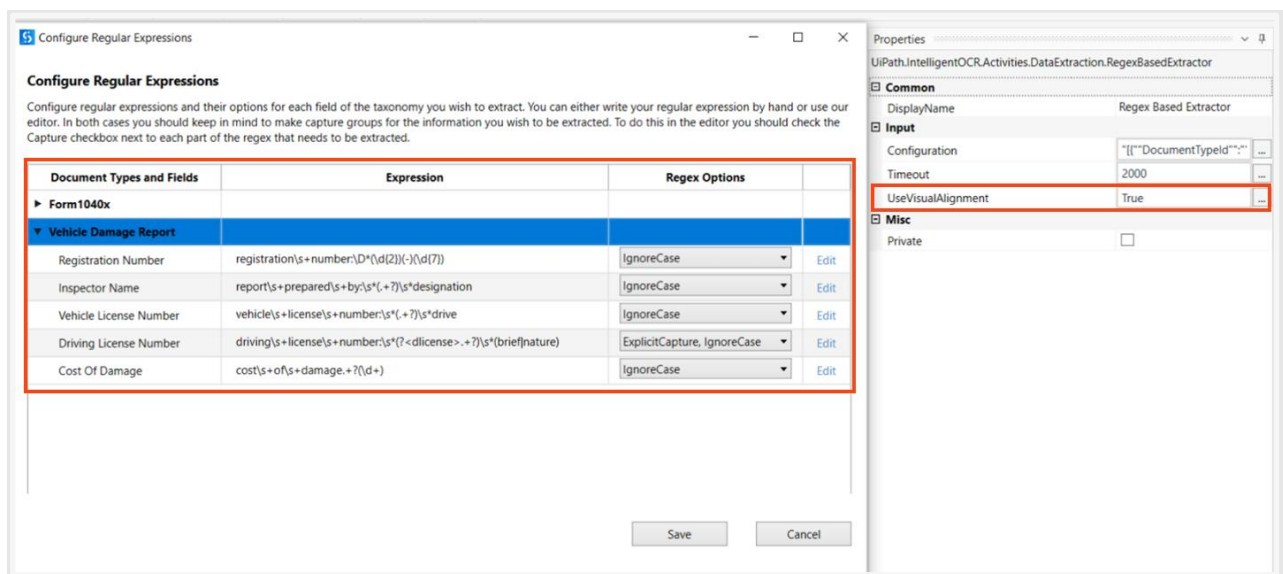
[Delete] [Export ▼] [Close]

9. Within the Data Extraction Scope activity, add the Regex Based Extractor activity.
10. To configure regular expressions, in the Regex Based Extractor activity, click **Configure Expressions**.

The Configure Regular Expressions wizard is displayed.



11. To configure the Regex Based Extractor, set the UseVisualAlignment property to True and define regular expressions.

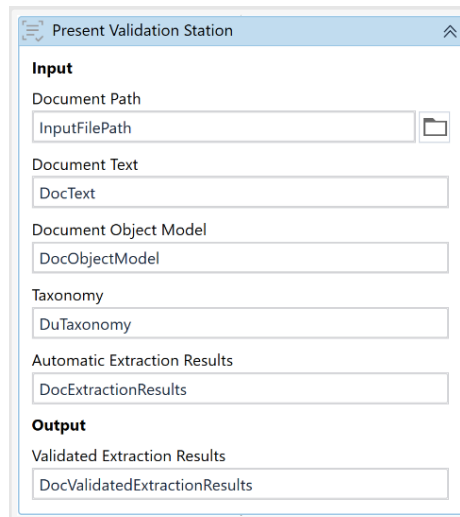


All expressions use the IgnoreCase option, indicating that the search is not case-sensitive. For Driving License Number, you also enable ExplicitCapture, indicating that the only valid captures are the ones of groups explicitly named or numbered and are defined as (?<name> subexpression). Any unnamed parentheses are to be ignored.

Task 8: Validate and export the extracted results

Steps:

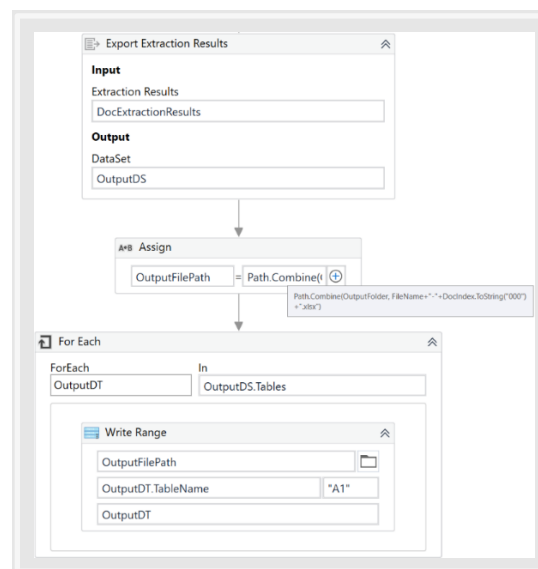
1. Validate extraction results using the Present Validation Station activity.



The 'Present Validation Station' activity configuration window is shown. It has two main sections: 'Input' and 'Output'.

- Input:**
 - Document Path: InputFilePath
 - Document Text: DocText
 - Document Object Model: DocObjectModel
 - Taxonomy: DuTaxonomy
 - Automatic Extraction Results: DocExtractionResults
- Output:**
 - Validated Extraction Results: DocValidatedExtractionResults

2. Extract the results using the Export Extraction Results activity, which you need to add to the end of your workflow.



If there are multiple documents within the input file, you must ensure a unique output path. That is why you need to add the Index variable to it: `OutputFilePath = Path.Combine(OutputFolder, FileName+"-"+DocIndex.ToString("000")+".xlsx")`.

The results are stored into a dataset containing multiple tables, which could then be written to an Excel file or used directly in a downstream process.

Full workflow solution

You can find the workflow solution in the DU_FullExercise.zip file.