

Different Format Invoice Data Extraction

This is a complete solution workflow in which you can input your documents, create a taxonomy of different formats and extract data using document understanding.

Packages to be Installed:

1)Intelligent OCR activities

2)Omni Page OCR

3)Machine Learning Extractor

Processing and extracting data from different types of documents using Document Understanding

- This workflow explains how you can use document understanding and process multiple documents of different formats and extract data.
- The UiPath Document Understanding Framework is designed to help users combine different approaches to extract information from multiple documents, not necessarily with the same structure

Steps to be followed:

1. Get Files
2. Load Taxonomy
3. Digitize Document
4. Classify Document Scope
5. Data Extraction Scope
6. Present Validation Station
7. Export Extraction Results

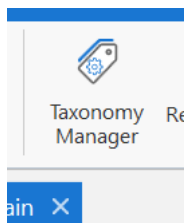
Load Taxonomy:

In this pre-processing step, you can add multiple document types and the fields you are interested in extracting. For example, you can work with Invoices, wanting to extract the Invoice Number, Invoice Date and Due Date.

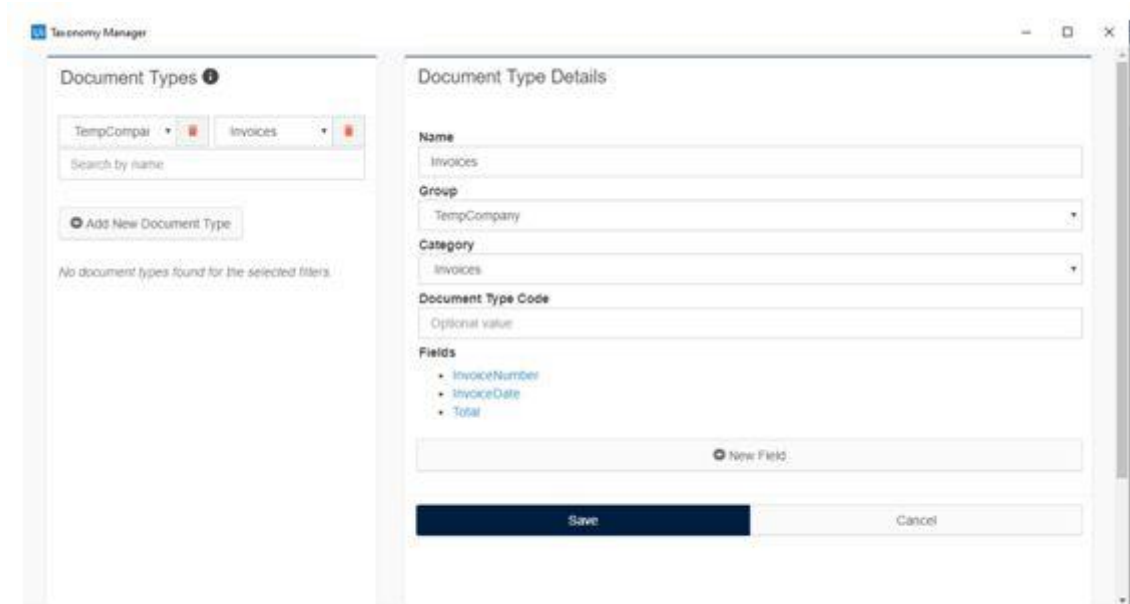


Step 2:

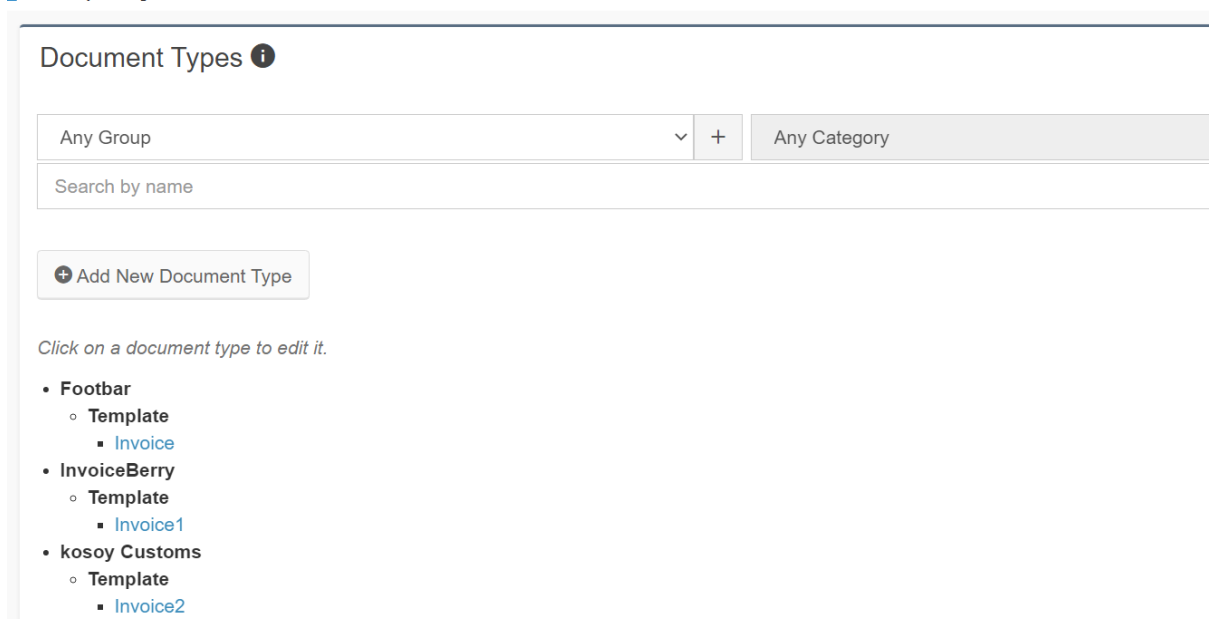
Click on Taxonomy Manager.



Step 3:

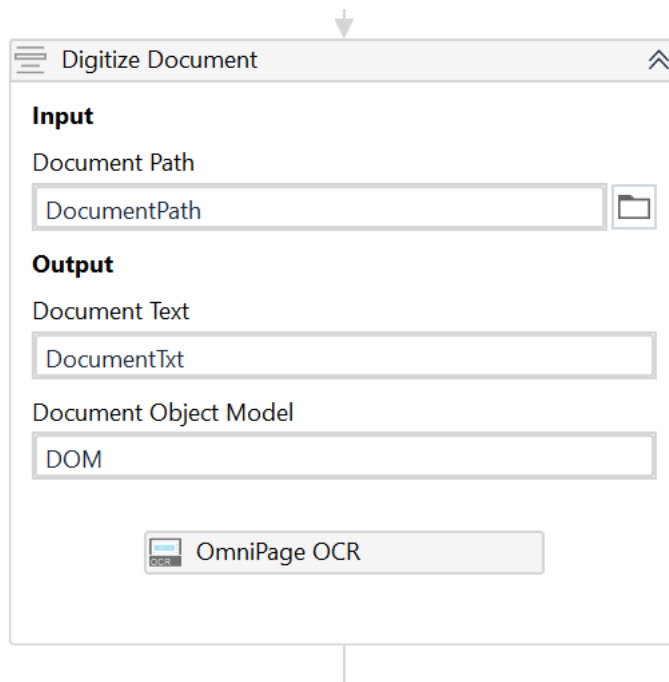


Taxonomy Manager



Digitize Document:

As the documents are processed one by one, they go through the digitization process. The difference for non-digital (scanned) documents is that you need to apply the OCR engine of your choice. The outputs of this step are the Document Object Model and a string variable containing all the document text and are passed down to the next steps.



The image shows a software window titled "Digitize Document". It has a menu icon on the left and a maximize icon on the right. The window is divided into two main sections: "Input" and "Output".

Input

Document Path

DocumentPath

Output

Document Text

DocumentTxt

Document Object Model

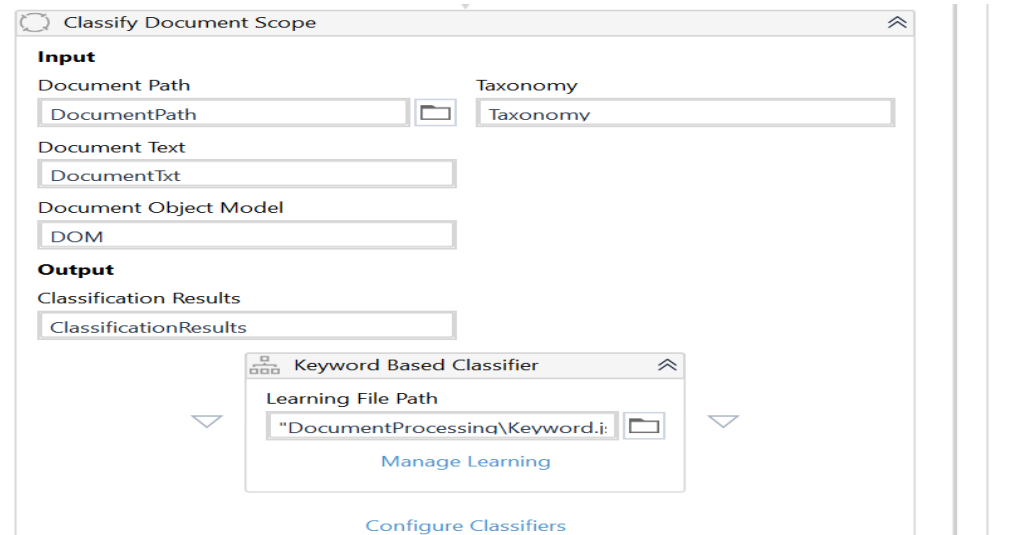
DOM

OmniPage OCR

The window contains several text input fields. The "Input" section has a "Document Path" label and a text field containing "DocumentPath", with a folder icon to its right. The "Output" section has two labels, "Document Text" and "Document Object Model", each followed by a text field containing "DocumentTxt" and "DOM" respectively. At the bottom of the window, there is a button labeled "OmniPage OCR" with a small icon to its left.

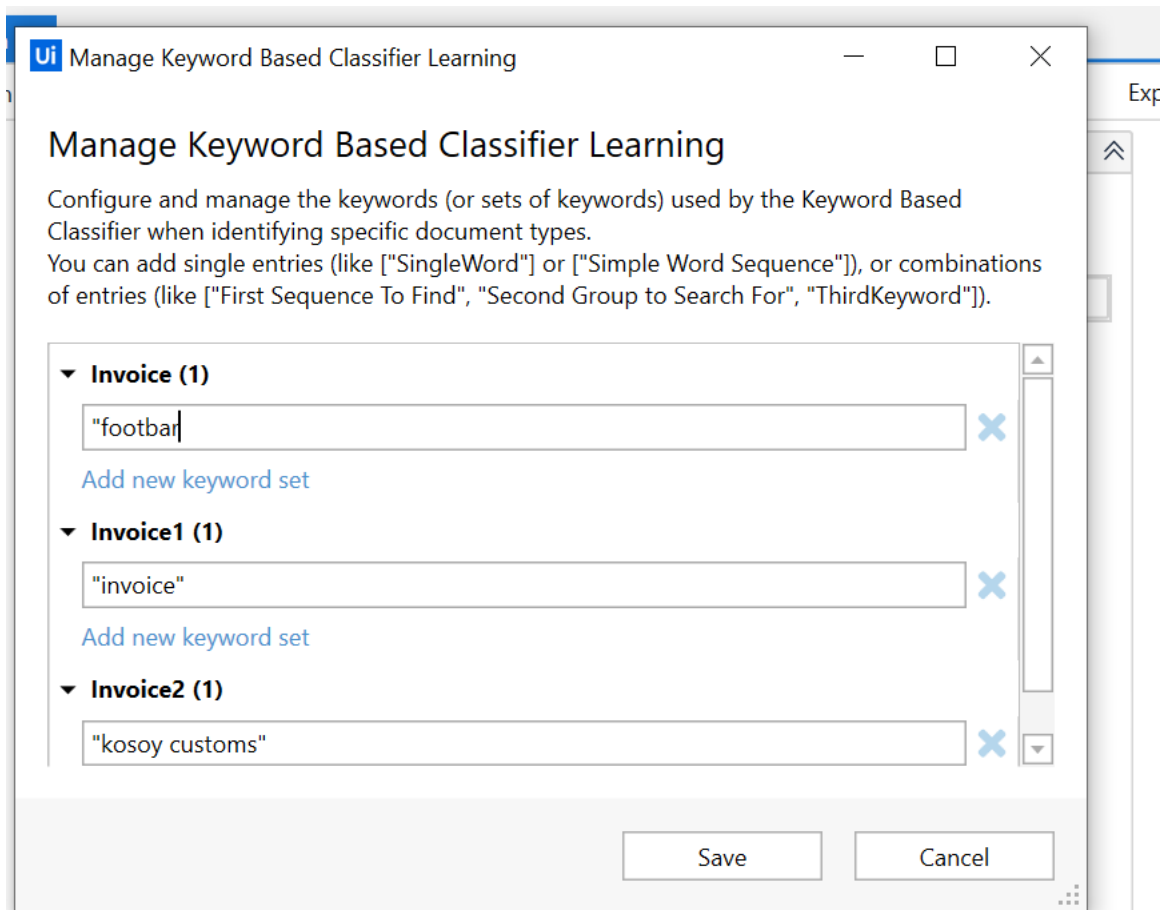
Classification:

After digitization, the document is classified. If you are working with multiple documents types in the same project, to extract data properly you need to know what type of document you're working with. The important thing is that you can use multiple classifiers in the same scope, you can configure the classifiers and, later in the framework, train them. The classification results help in applying the right strategy in extraction.



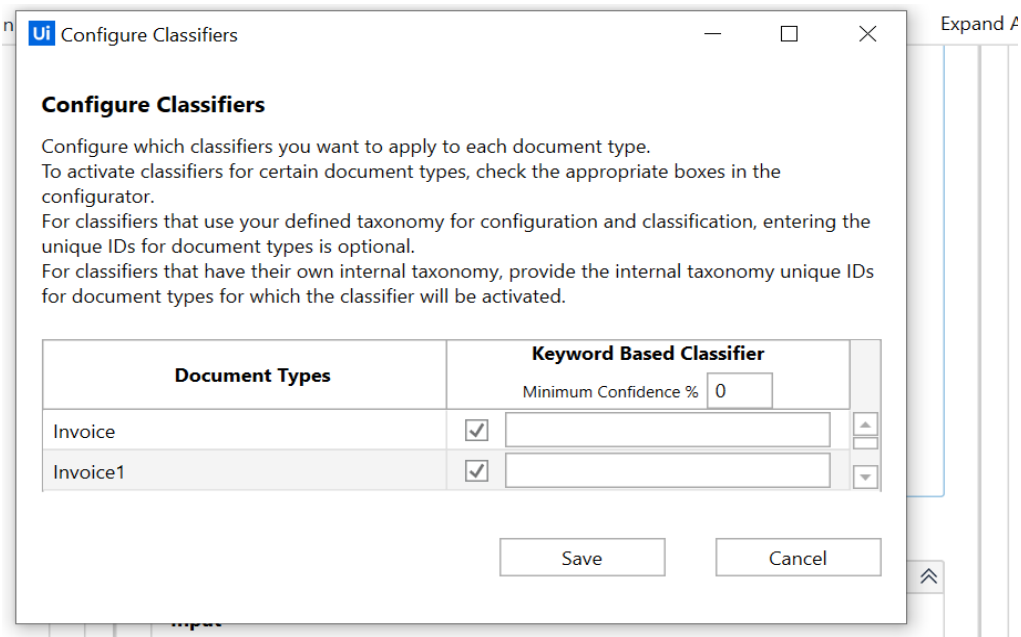
Step 2:

Click on ***Manage Learning.***



Step 3:

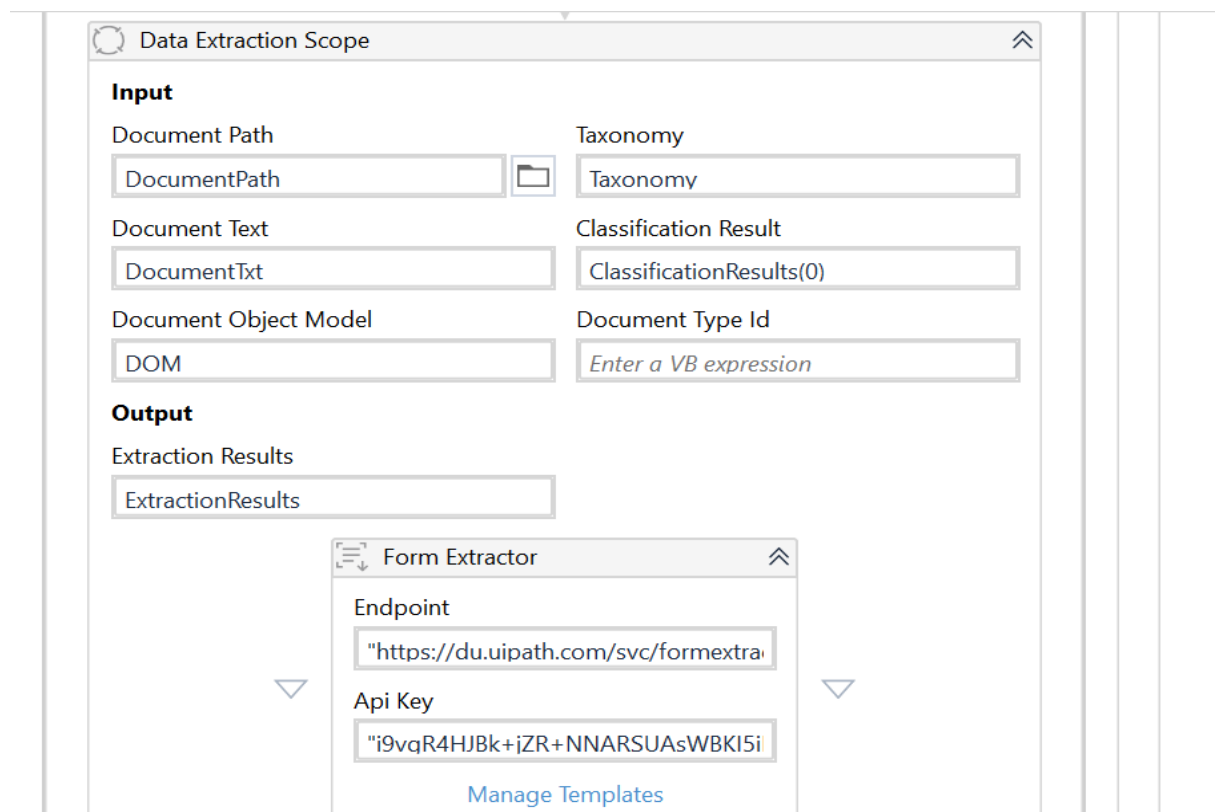
Click on **Configure Classifiers**



Click on check boxes and save.

Data Extraction Scope:

Extraction is getting just the data you are interested in. For example, extracting specific data from a 5-page document is quite troublesome if you want to do it with string manipulation. In this framework, you can use different extractors, for the different document structures, in the same scope application. The extraction results are passed further for validation.



There are different extractors,

- Form Extractor
- Machine Learning Extractor
- Regex Based Extractor

To get API Key go to Orchestrator click Admin→ Licences→Robots and Services→Generate API->Copy API

Step 2:

Click on Manage Template

Ui

Template Manager "InvoiceBerry"

o

Document Type

Invoice1

1

Page 1 Matching Info

INVOICE INVOICE TO: Company Name [Street address] [City, State]
[Country] [Postal]

v

Invoice No

Custom Selection

c

Invoice Date

Custom Selection

Discard changes

Save 4 / 4

A

Q

Q

1 / 1

INVOICE

LOGO

Your Company Name
[Street address]
[City, State]
[Country]
[Postal]

INVOICE TO:
Company Name
[Street address]
[City, State]
[Country]
[Postal]

Invoice Number
#0001
Date of Invoice
2021-01-17
Due Date
2021-01-17

DESCRIPTION	QTY	UNIT PRICE	TOTAL

SUBTOTAL0.00
DISCOUNT0.00

Step 3:

Click on Configure Classifiers

Configure Extractors


Configure which extractors you want to apply to each document type and field. To activate extractors for certain fields, check the appropriate boxes in the configurator. For extractors that use your defined taxonomy for configuration and data extraction, entering the unique IDs for document types and fields is optional. For extractors that have their own internal taxonomy, provide the internal taxonomy unique IDs for both document types and fields for which the extractor will be activated.

Document Types and Fields	Form Extractor
<div>▶ Invoice1</div>	<div>Framework Alias <input type="text"/></div> <div>Minimum Confidence % <input type="text" value="80"/></div> <div> <input checked="" type="checkbox"/> <input type="text"/> </div>

Save Cancel

Present Validation Station:

- The extracted data can be validated by a human user through the Validation Station. A best practice is to build logic around the decision of adding or not a human validation step, with rules depending on the specific use case to be implemented. Validation results can then be exported and used in further automation activities.



Present Validation Station

Input

Document Path
DocumentPath

Document Text
DocumentTxt

Document Object Model
DOM

Taxonomy
Taxonomy


Automatic Extraction Results
ExtractionResults

Output

Validated Extraction Results
ValidateResults

Export Extraction Result:

Once you have your validated information, you can use it as it is, or save it in a DataTable format that can be converted very easy into an Excel file.



Export Extraction Results

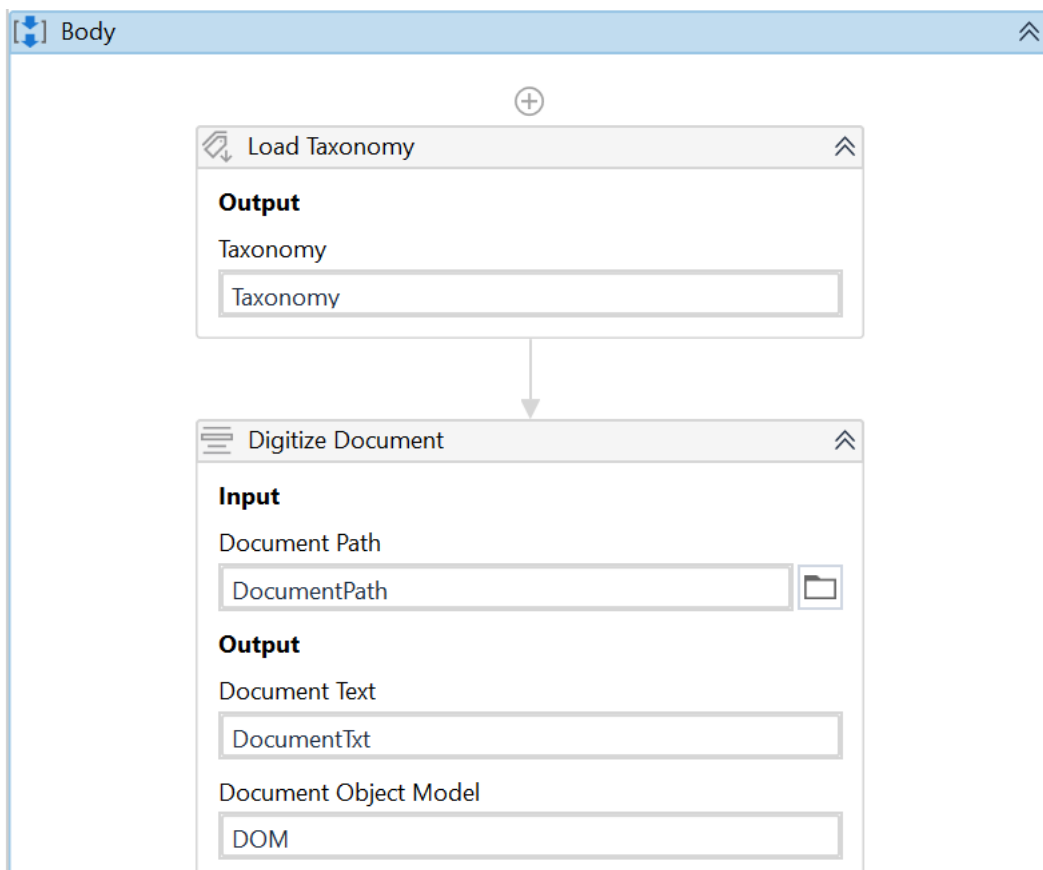
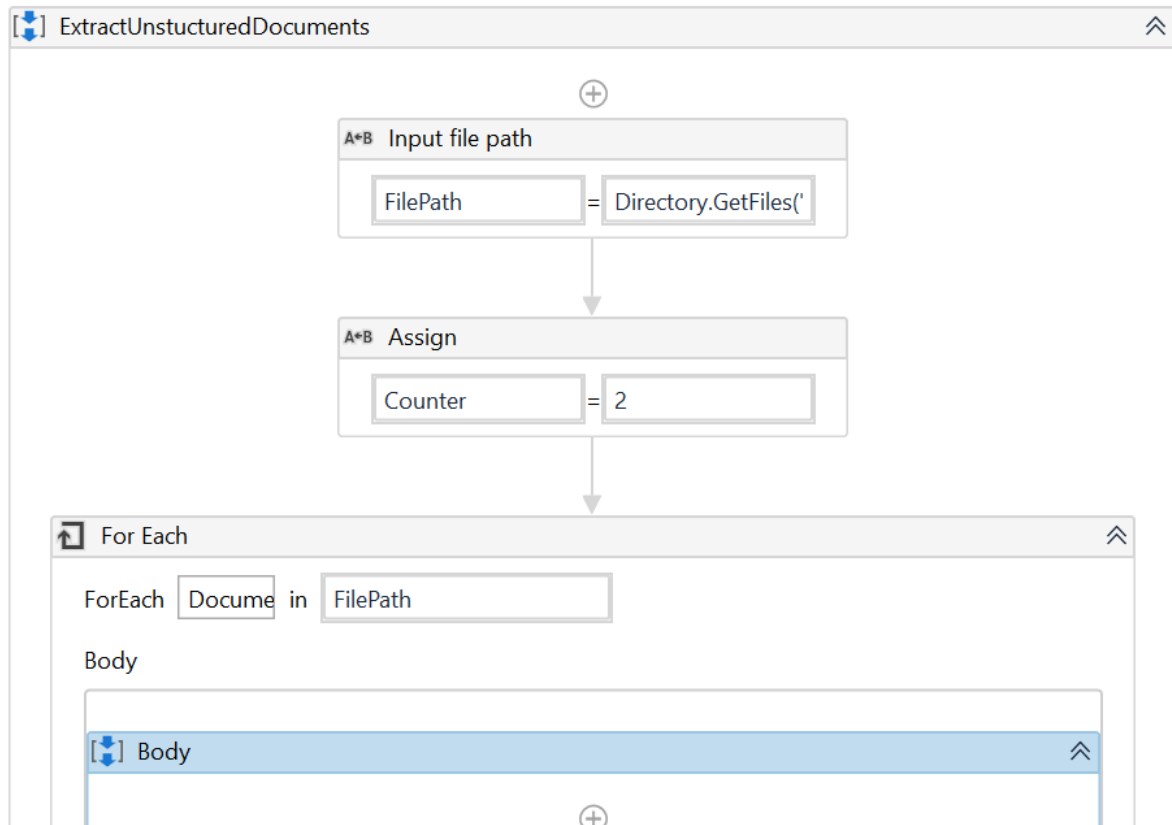
Input

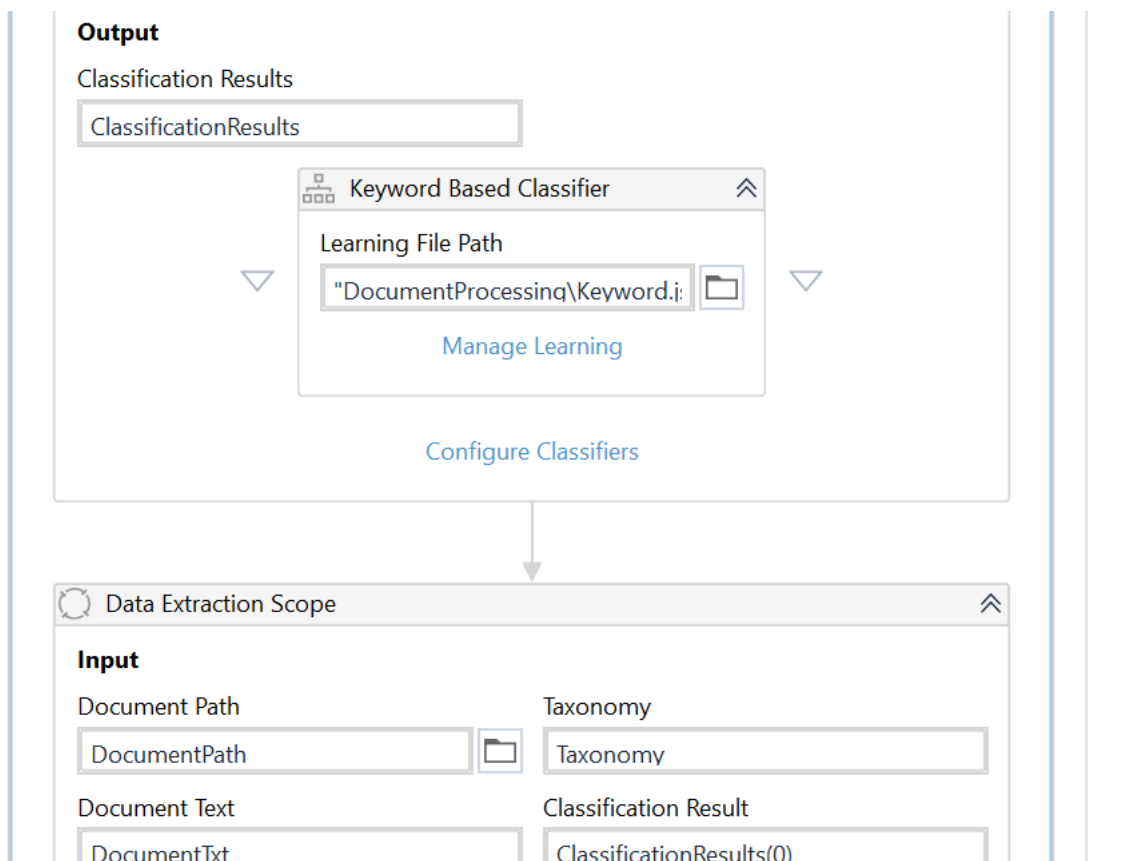
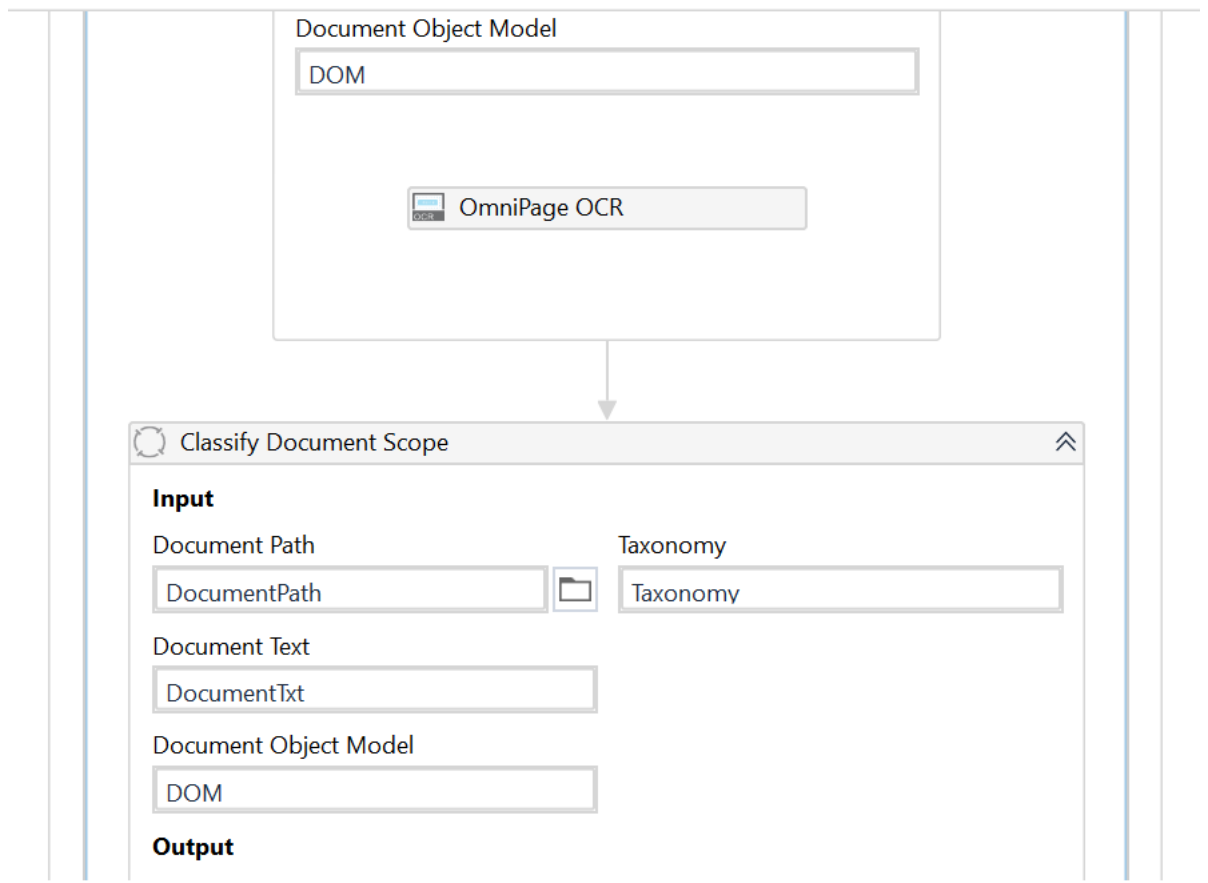
Extraction Results
ExtractionResults

Output

DataSet
DataSet

Example:





Document Text	Classification Result
<input type="text" value="DocumentTxt"/>	<input type="text" value="ClassificationResults(0)"/>
Document Object Model	Document Type Id
<input type="text" value="DOM"/>	<input type="text" value="Enter a VB expression"/>
Output	
Extraction Results	
<input type="text" value="ExtractionResults"/>	

Form Extractor


Endpoint

Api Key

Manage Templates

Configure Extractors

Present Validation Station

Input
Document Path
 

Document Text

Document Object Model

Taxonomy

Automatic Extraction Results

Output
Validated Extraction Results

validateresults

Export Extraction Results

Input

Extraction Results

ExtractionResults

Output

DataSet

DataSet

For Each

ForEach tables in DataSet.Tables

Body

Body

Body

Write Range

Var OutputFile

"Sheet1"

"A"+Cou

tables

A*B Assign

Counter

= Counter+1