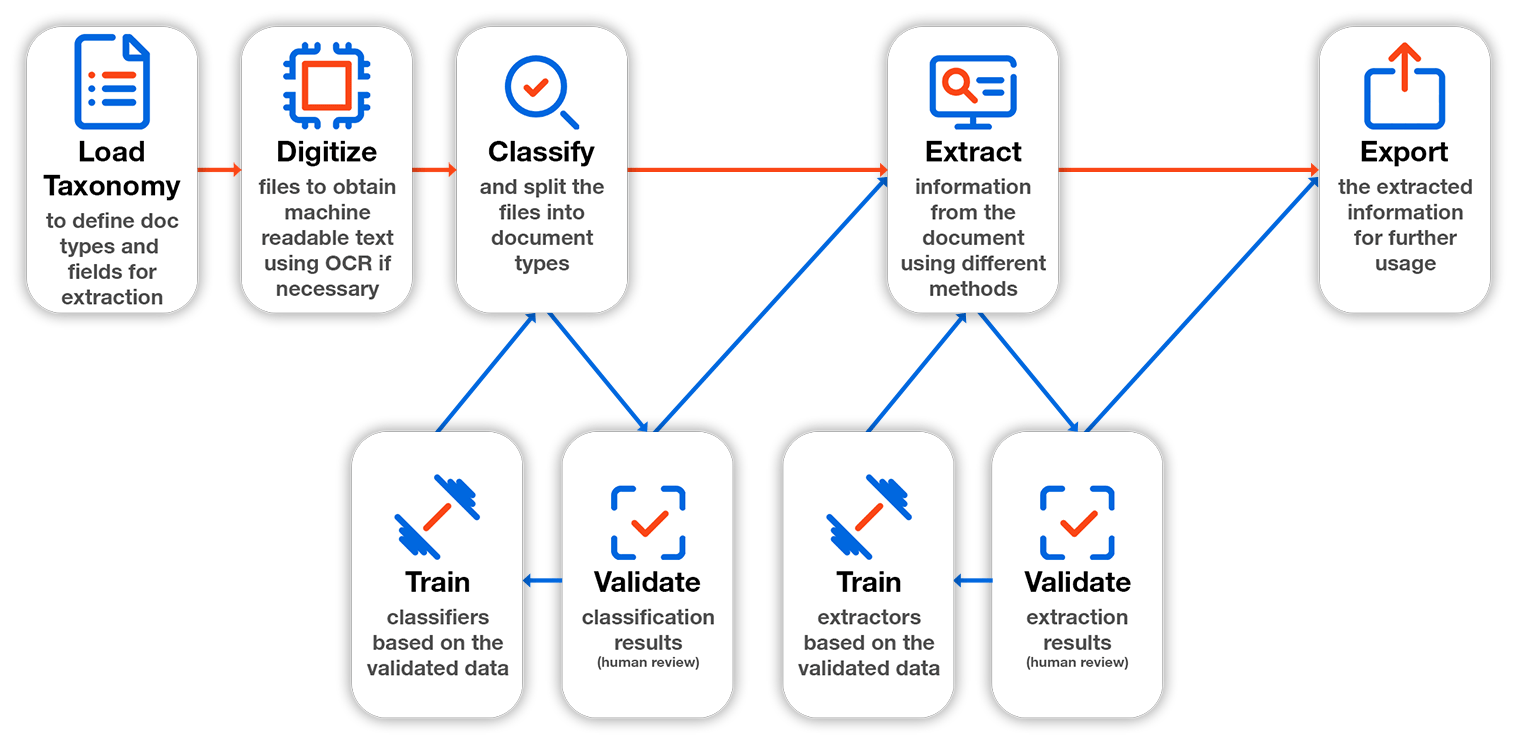
**Document Understanding**

The **UiPath Document Understanding** framework facilitates the processing of incoming files, from file digitization to extracted data validation, all in an open, extensible, and versatile environment.

**Document Understanding** is designed to help you combine different approaches to extract information from multiple document types. The main aim is to make the process of extracting data as easy as possible: creating one single workflow that will extract data from a variety of documents.

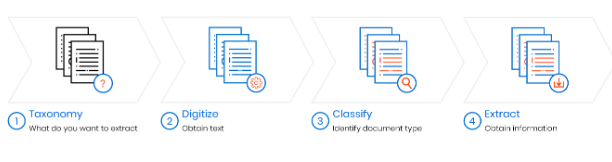
Document Understanding Framework is designed to help users combine different approaches to extract information from multiple documents, not necessarily with the same structure.

**Document Understanding Framework Components**:



**Document Understanding Framework Components**:

* [**Taxonomy**](https://docs.uipath.com/document-understanding/docs/taxonomy)  
  **What documents need to be processed and what data is required from them?**  
  Used to define the document types and the pieces of information targeted for data extraction (fields) for each document type, and formalizes this information into a dedicated Taxonomy structure. This metadata information is managed through the Taxonomy Manager.
* [**Digitization**](https://docs.uipath.com/document-understanding/docs/digitization)  
  **What does this file contain?**  
  Used to obtain the textual content and the structure of the incoming document, turning a file into machine-readable content so it can be further processed downstream.
* [**Document Classification**](https://docs.uipath.com/document-understanding/docs/classification)  
  **What types of documents from the taxonomy are found in this file?**  
  Used to automatically determine what document types are found within a digitized file.
* [**Document Classification Validation**](https://docs.uipath.com/document-understanding/docs/classification-validation)  
  **Is the predicted classification correct? This is how I can review and correct it.**  
  Used for assisting in the human validation and correction of the automatic classification and document splitting results.
* [**Classification Training**](https://docs.uipath.com/document-understanding/docs/classification-training)  
  **Did the human review the data? This is how the robot can learn from it.**  
  Used to pass the human validated information back to the classifiers, to use it to improve their future predictions.
* [**Data Extraction**](https://docs.uipath.com/document-understanding/docs/data-extraction)  
  **What data can be found in this particular document?**  
  Used to capture the information required for the identified document type, within the given input document and classification page range.
* [**Data Extraction Validation**](https://docs.uipath.com/document-understanding/docs/data-extraction-validation)  
  **Is the extracted information correct? This is how I can review and correct it.**  
  Used for assisting in the human validation and correction of the automatically extracted data results.
* [**Data Extraction Training**](https://docs.uipath.com/document-understanding/docs/data-extraction-training)  
  **Did the human review the data? This is how the robot can learn from it.**  
  Used to pass the human validated extracted data back to the extractors, to use it to improve their extraction predictions.
* [**Data Consumption**](https://docs.uipath.com/document-understanding/docs/data-consumption)  
  Used to export the validated data in order to consume it.
* [**Metering & Charging Logic**](https://docs.uipath.com/document-understanding/docs/metering-charging-logic)  
  Used to explain the consumption of units per page for each available service.





**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 vendor and the total amount, and with medical forms, wanting to extract insured ID number and patient name.

* 2

**Digitization**  
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.

* 3

**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.

* 4

**Extraction**  
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.

* 5

**Validation**  
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.

* 6

**Export**  
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.

* 7

**Training Classifiers and Extractors**  
Classification and Extraction are as efficient as the classifiers and extractors used are. If a document wasn’t classified properly, it means it was unknown to the active classifiers. The same way goes for incorrect data extraction. The Framework provides the opportunity to train the classifiers and the extractors, to improve recognition of the documents and fields.