

FAQ – Document Management System

What is a Document Type?

A Document Type acts as a template that defines how information should be extracted from documents of that type. It includes a taxonomy of attributes that are targeted for extraction and a specified layout to facilitate this process. After a Document Type is registered, a dedicated model is trained and deployed to automate the extraction.

Why can't I update registered Document Types?

Once a Document Type is registered, its underlying model is trained based on the defined taxonomy and layout. Updating a registered Document Type would alter its structure, potentially invalidating previously processed data. Thus, to maintain data integrity and consistency, updates to registered Document Types are not permitted.

What do different Document Type statuses indicate?

DRAFT: Indicates a Document Type is in preparation and not yet registered.

REGISTERED: Confirms the taxonomy and layout of the Document Type are finalized, and the extraction model is active.

ERROR: Signifies issues within the extraction workflow, requiring attention.

How is a Document Type managed after creation?

Document Types can be assigned to users or groups, granting them permissions aligned with their roles—ranging from creating new types to viewing existing ones. Management actions include viewing configurations, generating metrics, or deleting Document Types.

What layout options are available for Document Types?

Document layouts are categorized based on their structure:

Unstructured (for documents without a fixed format), Semi-Structured (for documents with partial structure such as tables or key-values), Structured

(for documents with a fully defined structure), and ID Cards (specialized for identity documents with automatic corrections like skew or rotation).

What formats are supported for document uploads?

The system supports uploads in JPEG, JPG, PNG, PDF, TIFF, and TIF formats, with a maximum file size of 50MB.

How to define a taxonomy for a Document Type?

A taxonomy is a defined list of attributes that are extracted from the document. It is set up during the creation of a Document Type and is essential for guiding the extraction process.

What are page extractions and how are they processed?

Page extractions refer to the data pulled from documents based on the defined taxonomy. This process typically involves OCR (Optical Character Recognition) to convert document text into digital form, followed by any necessary formatting adjustments.

What are the key data types used in extractions?

Data types include Alphanumeric, Barcode, Date, Freeform text, Name, and Table. Each type has specific processing rules, such as format conversions for dates or structural parsing for tables.

How to perform annotations for attribute extraction?

Annotations are made by selecting the annotation tool and drawing on the document to define the area for extraction. This action specifies where the system should focus its OCR and data extraction efforts.

What does the confidence percentage tell us?

The confidence percentage is an estimate of the accuracy of the extracted data. It helps users gauge the reliability of the extracted information and determine if manual review might be necessary.
