

Split Configuration

Last Modified on 05/02/2025 6:25 am EDT

Overview

You can split a document to multiple documents based on the following logics:

- Blank page between the documents.
- Blank page with strong label.
- Page numbering.
- Barcode with Static label.

Split is available for AP Flow adapter only.

To use the split functionality, you need to configure the following two yaml files:

- Split-configuration.yaml
- Tenant-workspace.yaml

The split-configuration.yaml configuration file defines the criteria based on which a document will be split into multiple documents.

In the tenant-worksapce.yaml file, you configure three parameters under split handling. For details, see [this article](#).

Template

```
kind: document
metadata:
  name: extraction/v1/documents/split-configuration
spec:
  outputFileName: '{FILE_NAME}_{FILE_INDEX}.{FILE_TYPE}'
  ocr:
    - ocrProvider: DocumentIntelligence
      endpoint: https://open-ai-form-recognizer.cognitiveservices.azure.com
      apiKey: 8eb04706b5e844a49184h764b82698f1
      modelName: 'prebuilt-layout'
      enabled: true

  splitters:
    - fileTypes: PDF
      splitProvider: PdfSplitter
      enabled: true

    - fileTypes: TIFF,TIF
      splitProvider: TiffSplitter
      enabled: true

  strategies:
    - type: EmptyPage
      enabled: true
      splitPosition: Discard

    - type: PlaceholderPage
      enabled: true
      splitKeywords: OCR_INVOICE_SEPARATOR,BLANK PAGE
      splitRegex: ""
      splitPosition: Discard

    - type: BarcodeLabel
      enabled: true
      splitKeywords: ABDCC-70402,BUT2324149987
      splitRegex: ""
      splitPosition: AddToDocumentStart

    - type: InvoiceNumber
      enabled: true
      splitKeywords: Invoice \#:,Invoice \#,Invoice Number:,Invoice Number,Invoice Num:,Invoice Num,Invoice No.
      splitRegex: '##KEYWORD##\s*(\S+)'
      splitPosition: AddToDocumentStart

    - type: PagesCount
      enabled: true
      splitKeywords: Page 1 Of,Page 1/,1 of,Page \#1,1/,Page
      splitRegex: '##KEYWORD##\s*(\S+)'
      splitPosition: AddToDocumentEnd
```

Parameter	Description
Type	The split condition.
Enabled	Enables the split.
Split Keywords	The keywords based on which the document is split.

Parameter	Description
Split Regex	The regular expression string to decide where in the documents the split happens.
Split Position	The position of the split.