User guide

Nuxeo OCR addon

Document tracking

Change history

Version	Date	Authors	Description
1.0.0	18/09/2020	Sébastien Guillaume (OCDM)	Initial version

Distribution history

Version	Date	Distributed to	For
1.0.0	18/09/2020	Sébastien Guillaume (OCDM)	Provided to Nuxeo

Documentary references

Name	Version	Description

Table of Contents

1. Main concept

Nuxeo OCR addon is designed to provide a fulltext extraction solution embedded in Nuxeo using a secured OCR service accessible through REST API.

Your Nuxeo instance must have access to this endpoint https://dm-ocr.oceaneconsulting.com

This guide describes provided functionalities and how to use this features in Nuxeo. This guide will not describe :

- How to install and configure the Nuxeo OCR Addon (check install guide documentaion)
- How to configure a Nuxeo custom operation, service or workflow

2. Technical features

Nuxeo OCR addon provides functionalities through Nuxeo operations listed and described in this document.

Some operations manage data (facet/schema) and will be detailled in the document.

2.1. Restrictions

Access to OCR API service is restricted by a configuration and end-users can have access to OCR feature only with the following elements :

- The document must have a content with one of the following mime-type (application/pdf will be supported in next release) :
  - image/png
  - image/jpeg
  - image/tiff

2.2. Operations

Nuxeo operations allow to use Nuxeo OCR API where you need it :

- Automation chains (workflows, events, ...)
- Services
- WebUI components
- ...

Each operation has :

- Unique identifier
- Input/Output
- Parameters

Operations can produce :

- Specific errors / exceptions

2.2.1. Document.OCRTextExtractor

Identifier	Description
Document.OCRTextExtractor	Call OCR REST API service passing the current document to extract fulltext content and assign it to configured metadata

Input	Output
Document	Document *Same document with OCR content available in ocr schema*

Parameter	Description	Type	Mandatory
accuracy	Name of the OCR accuracy to use (Fast by default)	String	x
language	Name of the OCR language to use (English by default)	String	x
provider	Name of the OCR provider to use (Tesseract by default)	String	x
xpath	Metadata (schema:metadata) used as content for OCR (file:content by default)	String	x

2.2.2. Document.OCRTextSubmit

Identifier	Description
Document.OCRTextSubmit	Launch an OCR process call REST API endpoint to submit content from a document in order to process OCR in an asynchronous process

Input	Output
Document	String *Identifier of your OCR job usable with Document.OCRTextFetch operation*

Parameter	Description	Type	Mandatory
accuracy	Name of the OCR accuracy to use (Fast by default)	String	x
language	Name of the OCR language to use (English by default)	String	x
provider	Name of the OCR provider to use (Tesseract by default)	String	x
xpath	Metadata (schema:metadata) used as content for OCR (file:content by default)	String	x

2.2.3. Document.OCRTextFetch

Identifier	Description
Document.OCRTextFetch	Launch an OCR process call REST API endpoint to fetch fulltext content from a ticketId and assign it to a document metadata

Input	Output
Document	Document *Same document with OCR content available in ocr schema*

Parameter	Description	Type	Mandatory
ticket	Job identifier used to check if OCR content is available	String	x

2.3. Enricher

To check if a document can be eligible to OCR, two new enrichers (document and blob) are available. These enrichers can be invoked through REST API with the following endpoint :

- http://NUXEO_URL/nuxeo/api/v1/id/DOCUMENT_ID?enrichers.document=ocr
- http://NUXEO_URL/nuxeo/api/v1/id/DOCUMENT_ID?enrichers.blob=ocr

2.3.1. OCR document enricher (enrichers.document=ocr)

This enricher is calculating if the current document contains at least a blob in *file:content* and if this blob matches the list of supported mimetypes defined by the Nuxeo OCR addon. This list is embedded into the Nuxeo configuration and currently set to image/jpeg, image/png and image/tiff.

- If all criterias match requirements, the JSON replied by endpoind will compatible = true and this document is eligible to OCR service

```
"contextParameters": {
  "ocr": {
    "compatible": true
  }
}
```

- If any criteria does not match requirements, the JSON replied by endpoind will be compatible = true and this document is NOT eligible to OCR service

```
"contextParameters": {
  "ocr": {
    "compatible": false
  }
}
```

2.3.2. OCR blob enricher (enrichers.blob=ocr)

This enricher is calculating if the current blob matches the list of supported mimetypes defined by the Nuxeo OCR addon. This list is embedded into the Nuxeo configuration and currently set to image/jpeg, image/png and image/tiff.

- If all criterias match requirements, the JSON replied by endpoind will compatible = true and this blob is eligible to OCR service

```
"file:content": {
  "name": "ocdm_test_008.jpg",
  "mime-type": "image/jpeg",
  ...
  "ocr": {
    "compatible": true
  }
}
```

- If any criteria does not match requirements, the JSON replied by endpoind will be compatible = true and this blob is NOT eligible to OCR service

```
"file:content": {
  "name": "ocdm_test_009.jpg",
  "mime-type": "image/jpeg",
  ...
  "ocr": {
    "compatible": false
  }
}
```

2.4. Nuxeo ocr schema

In order to store OCR informations, a new schema is available called **ocr** with 2 metadatas :

- **fulltext** as a simple String contains last extraction of OCR
- **advanced** as a complextype list, that means the complete history list of all extraction executed, contains the following sub-metadatas :
  - *fulltext* as a simple String contains extraction of OCR
  - langage as a simple String contains the langage used for this OCR extraction
  - field as a simple String contains the xpath of the metadata where content has been extracted
  - locale as a simple String contains the locale used for OCR (not yet used)
  - accuracy as a simple String contains technical information on extraction
  - provider as a simple String contains the name of the OCR provider (currently, only Tesseract is available)

2.5. Nuxeo ocr facet

The **ocr** facet contains the **ocr** schema. The Nuxeo OCR addon is able to dynamically add the **ocr** facet on a document if this document does not contain this facet yet. This feature is embedded in two operations *Document.OCRTextExtractor* and *Document.OCRTextFetch*.

2.6. Data structure

Structured data managed by OCR API service and Nuxeo addon are defined as two data structure : input data and ouput data.

This two structures are defined in `org.nuxeo.ocdm.ecm.ocr.model` package with two Java classes providing classic getter/setter.

2.6.1. OCRRequest

Parameter	Description	Type
accuracy	Name of the OCR accuracy to use	String
language	Name of the OCR language to use	String
provider	Name of the OCR provider to use	String
xpath	Metadata (schema:metadata) used as content for OCR	String

2.6.2. OCRResponse extends OCRRequest

Parameter	Description	Type
accuracy	Name of the OCR accuracy to use	String
language	Name of the OCR language to use	String
provider	Name of the OCR provider to use	String
xpath	Metadata (schema:metadata) used as content for OCR	String
fulltext	Content extracted from binary as fulltext OCR	String
ticket	Identifier of the job OCR extraction	String
error	Error message for any problem occuring during OCR process	String

3. Functionalities and end-user features

3.1. Fulltext button availability

By default, for any document matching the requirements (check 2.3Enricher), a new button Fulltext will be available in blob slots.

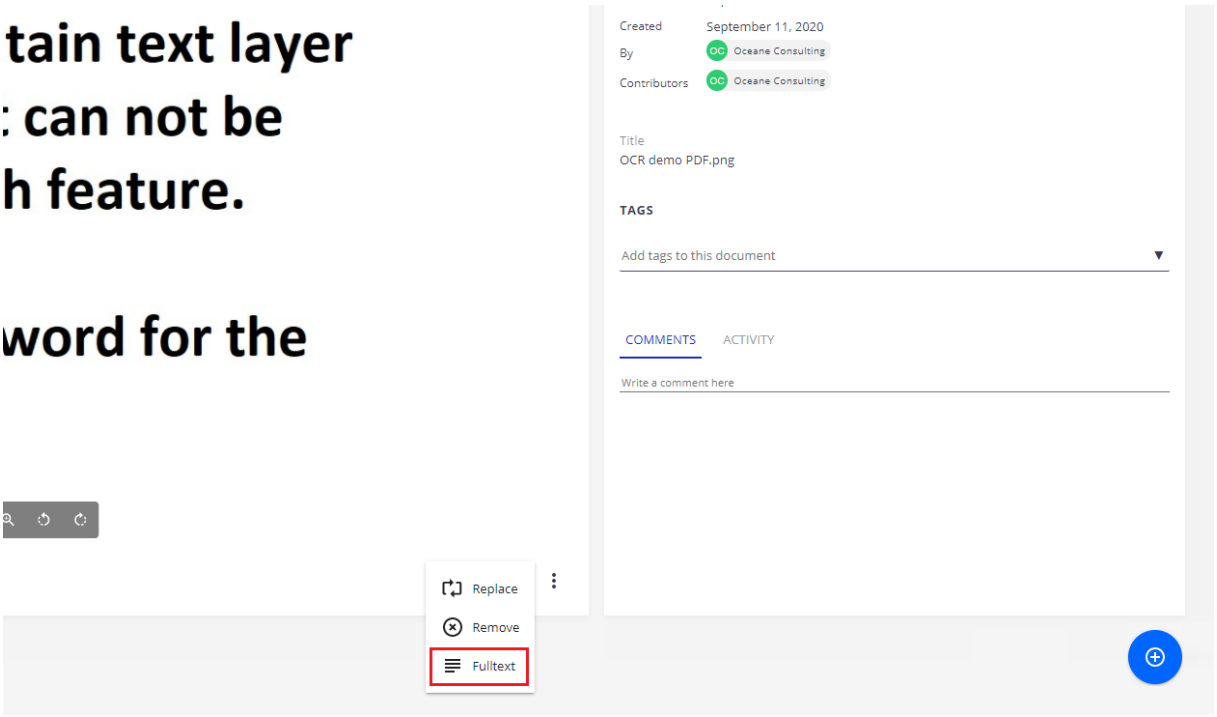

Illustration 1: Default blob action with Fulltext button available

As all blob slots are concerned (default, attachments, …), you can also run an OCR extraction on specific blobs :

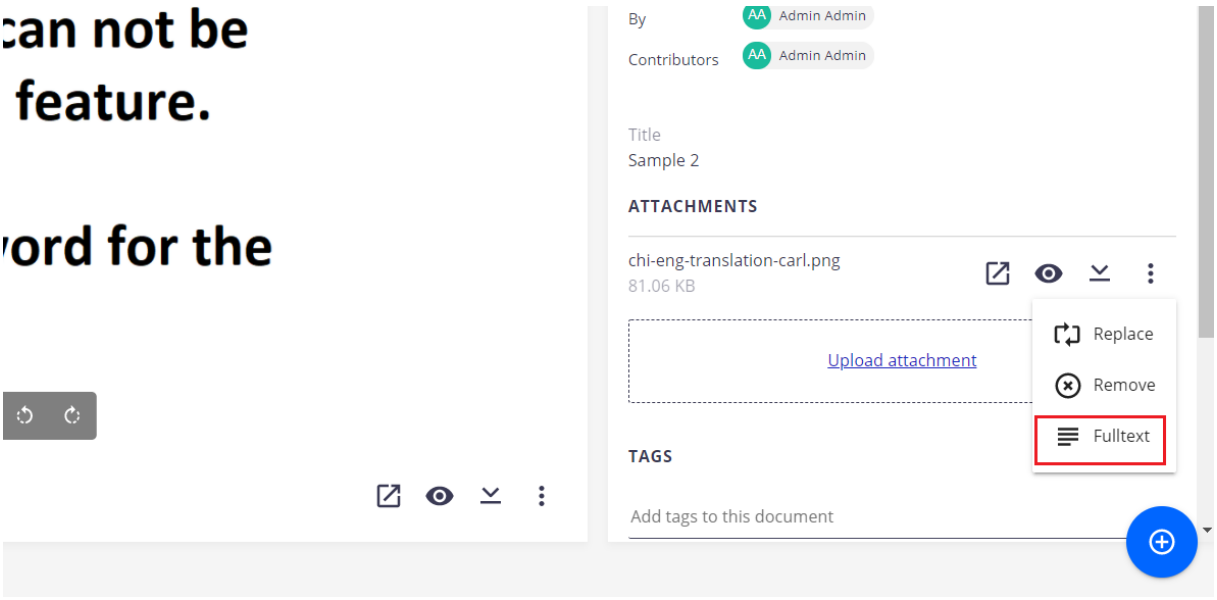

Illustration 2: Attachment blob action with Fulltext button available

3.2. Fulltext layout

Once clicked on Fulltext button, a new layout appears asking the end-user to select the language of extraction and the provider. Both informations are mandatory.

The list of available languages is provided by Oceane Consulting DM and is corresponding to the default provider Tesseract.

Currently, only Tesseract is supported as OCR provider.

After selecting language and provider, click on OK.

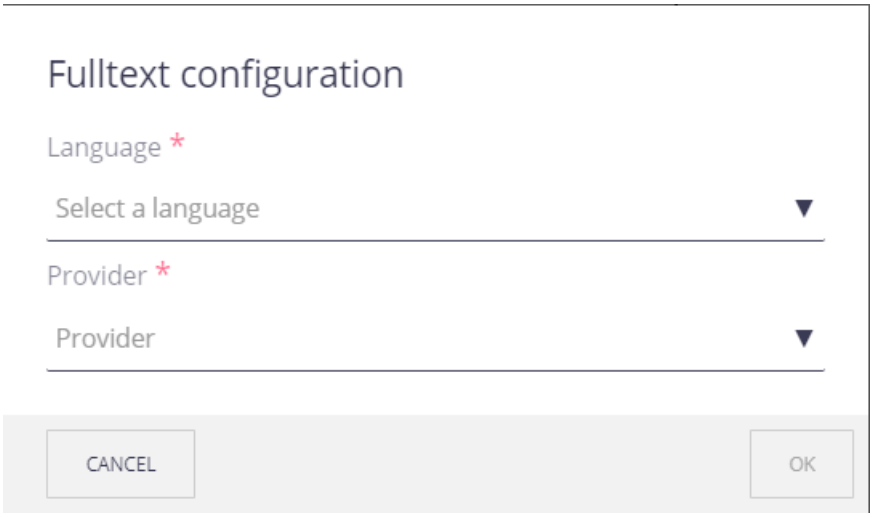

Illustration 3: Select language and provider

When OCR extraction is launched, a message appears at the bottom-left of your screen with the following message. Nuxeo is waiting for the result :

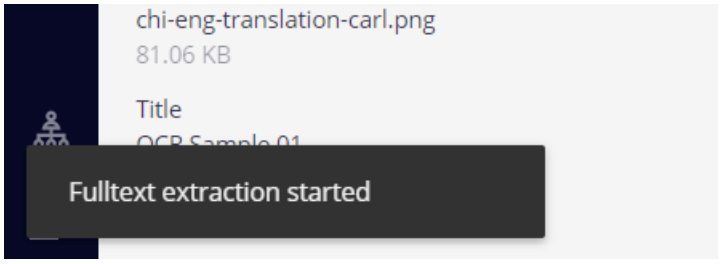

Illustration 4: Fulltext extraction starting message

When OCR result is available, a new message appears at the bottom-left of your screen :

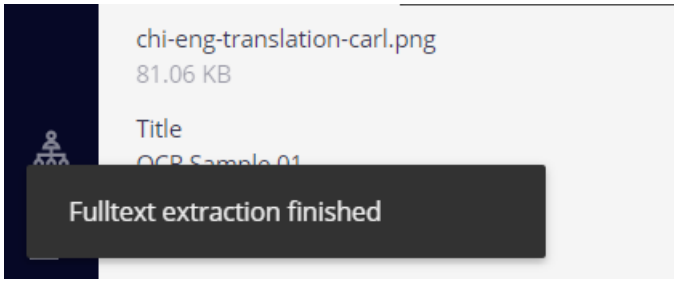

Illustration 5: Fulltext extraction finished message

3.3. Fulltext tab availability

After OCR extraction done, a new tab appears automatically on document page. This tab contains OCR data from **ocr** schema including ocr:fulltext and ocr:advanced metadata.

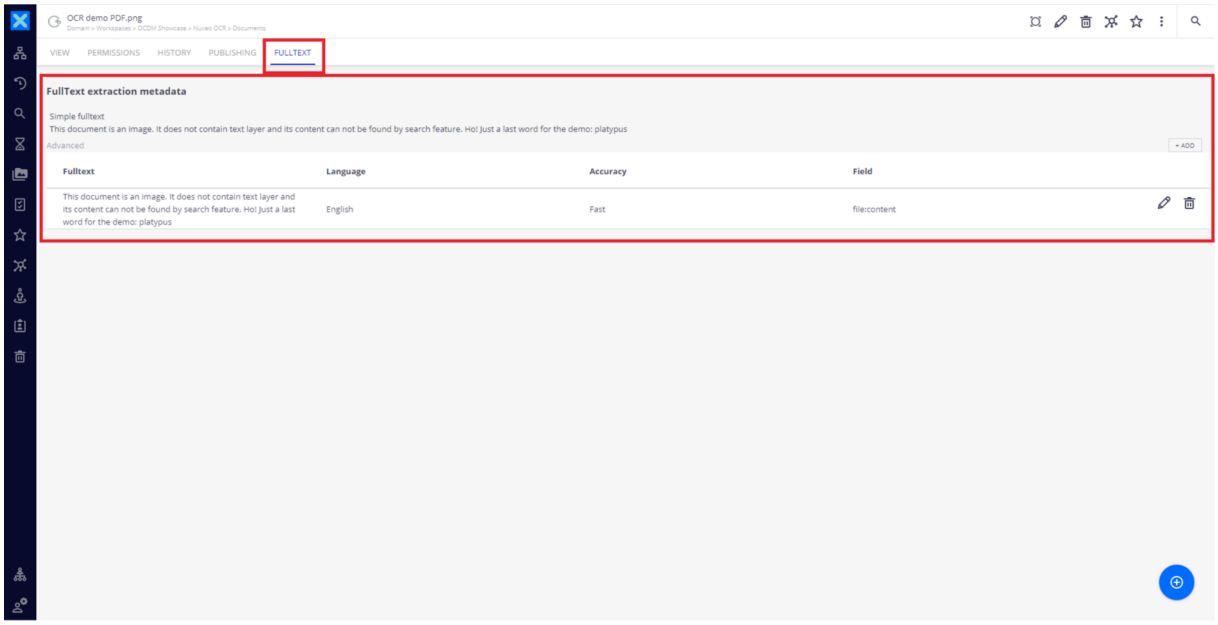

Illustration 6: Fulltext tab containing OCR extraction data

3.4. Audit and history tab

After OCR extraction done, new entry is available in audit trail. Available also in UI, each OCR extraction has a trace with Fulltext category.

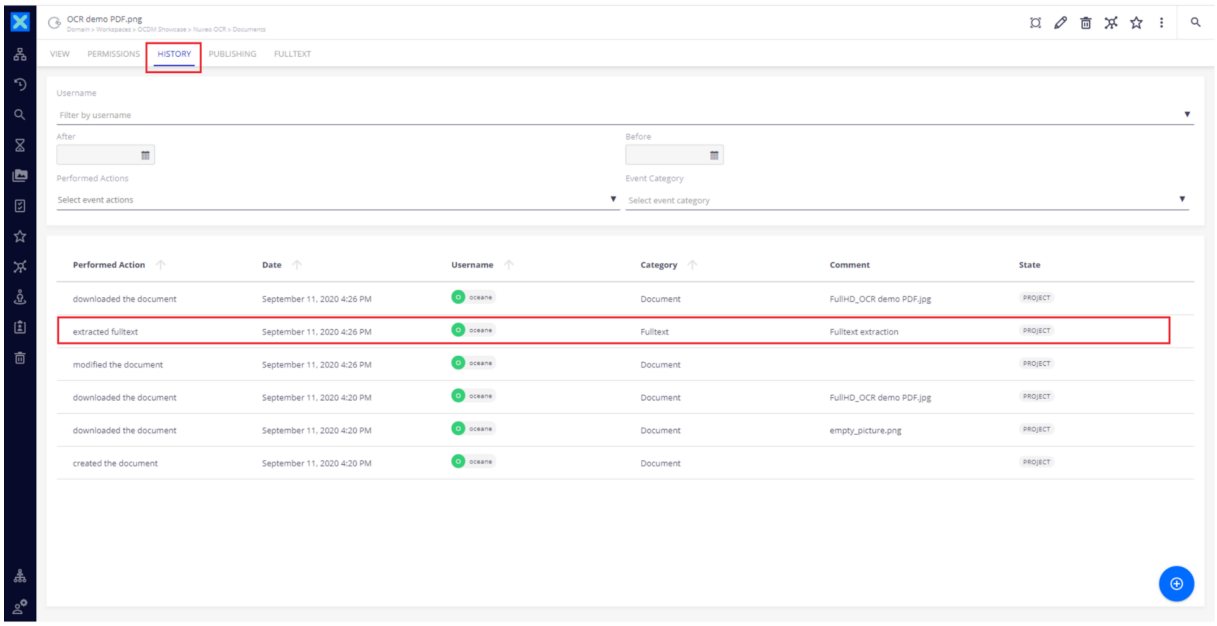

Illustration 7: Audit trail and history available for OCR extraction

3.5. Searches

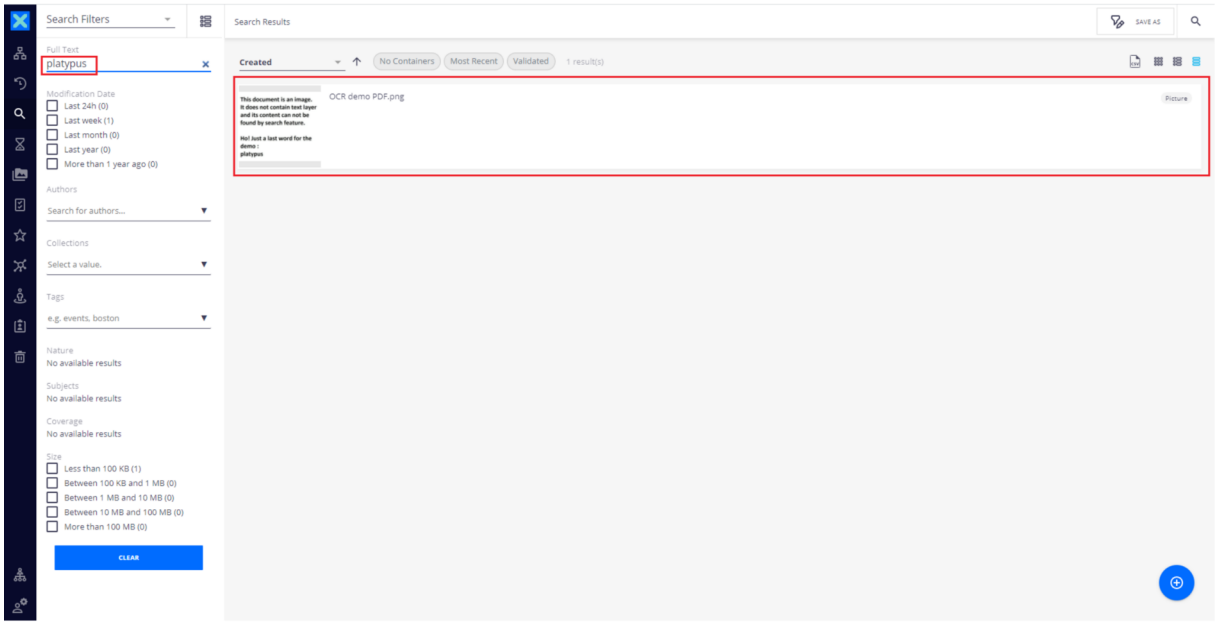

Illustration 8: Fulltext search including content extracted with OCR from blob