



BeeCon 2016

Integrating a simple OCR in Alfresco

Angel Borroy
developer @ keensoft



OCR for the Enterprise

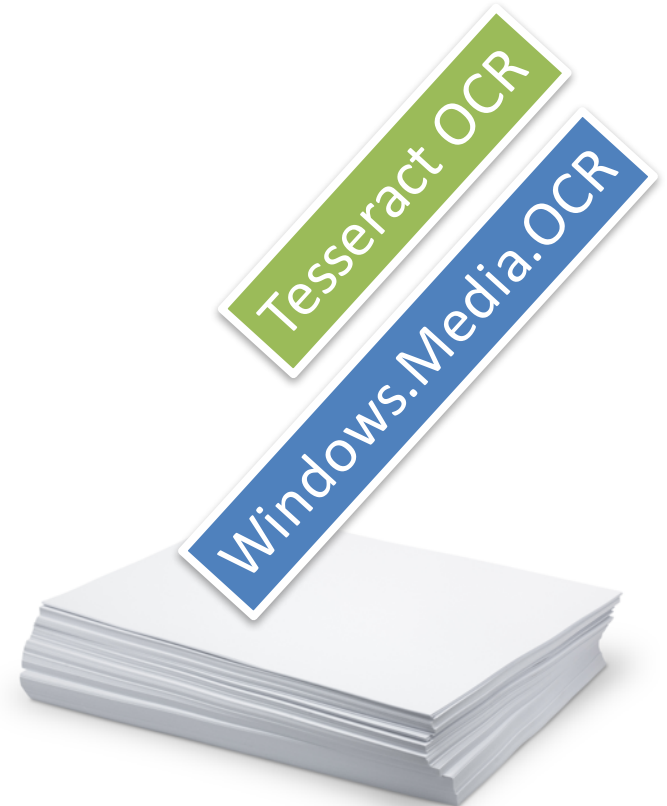
- Minimum license starting in **100,000 documents/year**
- Dedicated server required
- Hard learning curve
 - Regular expressions
 - Templates and workflows
 - Proprietary integration



BeeCon 2016

OCR for the Community

- Open Source
- No other server than Alfresco
- No learning curve, just drop off your documents on a folder and get **Searchable PDFs**
- Every hosting OS is supported

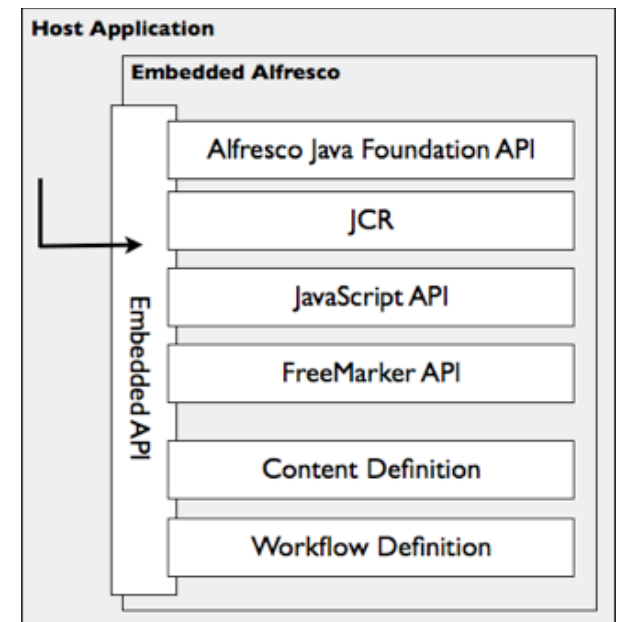


BeeCon 2016

Building a simple OCR Action

1 REPO AMP

- Content model (simple)
- Action
- Transformer





OCR Action: Key classes

```
<bean id="ocr-extract"
      class="es.keensoft.alfresco.ocr.OCRExtractAction"
      parent="action-executer" init-method="init">
  <property name="ocrTransformWorker" ref="transformer.worker.OCR" />
</bean>

<bean id="transformer.worker.OCR"
      class="es.keensoft.alfresco.ocr.OCRTransformWorker">
  <property name="serverOS" value="${ocr.server.os}" />
  <property name="executerWindows">
  <property name="executerLinux">
</bean>
```



OCR Action: Configuration Linux

alfresco-global.properties

```
# local ocr program
ocr.command=/usr/local/bin/pdfsandwich
ocr.output.verbose=true
ocr.output.file.prefix.command=-o
# rotating, cleaning, languages...
ocr.extra.commands=-lang spa
ocr.server.os=linux
```



BeeCon 2016



OCR Action: Configuration Windows

alfresco-global.properties

local ocr service

ocr.url=http://localhost:60064/api/OCR

ocr.output.verbose=true

rotating, cleaning, languages...

ocr.extra.commands=Spanish

ocr.server.os=windows



BeeCon 2016

OCR Action: Rule configuration

ocr

Edit Delete

Description:

- ✓ Active
- ✓ Run in background
- ☐ Rule applied to subfolders

When:

Items are created or enter this folder

▼

If all criteria are met:

All Items

▼

Perform Action:

Extract OCR Continue on error: Yes

Only apply for foreground



BeeCon 2016

OCR Action: Rule configuration




The screenshot shows the configuration for an OCR action rule. On the left, there are three icons: a Windows logo, a document with a list, and a clock with a document. A red arrow points from the clock icon to the 'Run in background' checkbox, which is highlighted with a red box. A green arrow points from the 'SYNCHRONOUS' label to the 'Active' checkbox. A red box labeled 'ASYNCHRONOUS' is positioned over the 'When:' section. The configuration details are as follows:

- ocr** (Title)
- Describe** (Section Header)
- ☒ **Active**
- ☒ **Run in background**
- ☐ **Rule applied to subfolders**
- When:**
 - or enter this folder
- If all criteria are met:**
 - All Items
- Perform Action:**
 - Extract OCR Continue on error: Yes

Buttons for **Edit** and **Delete** are located in the top right corner.



OCR Action: Results

Shared Files >  **Alfresco-Installation-Manual.pdf**  

Modified by Administrator on Thu 3 Mar 2016 16:08:47 | ★ Favorite | 👍 Like 0 | 💬 Comment | ➦ Share

⏮ Previous ⏭ Next 4 / 48 | - + 108% | Maximize | ⬇ Download | 🔗 | 🔍

1 INTRODUCTION

This document describes **Alfresco Community 5.0.a** installation process at the dependencies, based on Ubuntu servers and Amazon infrastructures.

2 COMPONENTS CATALOG

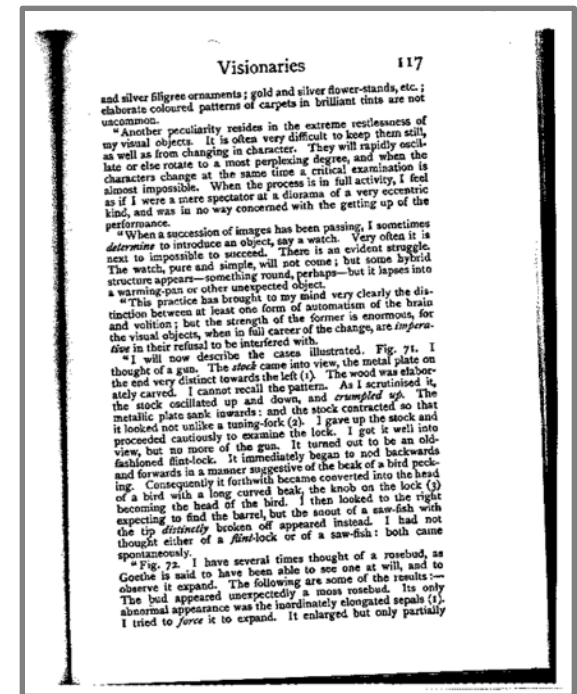
2.1 SERVERS AND PORTS

Search: 🔍 ⏮ ⏭ ⏹ Aa



What else?

- Study different original documents
 - Existing (incorrect) layer text
 - Image resolution below **200 dpi**
 - Landscape / portrait orientation
 - Paper size may change
- Plain OCR soft is not enough



* Image coming from <http://www.tobias-elze.de/pdfsandwich/>



BeeCon 2016



OCR Software: Mac OS X

<https://github.com/jbarlow83/OCRmyPDF>

- Generates a searchable PDF/A file from a regular PDF
- Keeps the exact resolution of the original images
- Keeps file size about the same
- Deskews and/or cleans the image before performing OCR
- Uses **Tesseract OCR** engine
- Open Source and developed with Python 3



BeeCon 2016

OCR Software: Linux

<http://www.tobias-elze.de/pdfsandwich/>

- Generates "sandwich" OCR pdf files
- Recognizes page layout (even for multicolumn)
- Uses `unpaper`, `convert`, `gs` and **tesseract**
- Open Source and developed using OCAML



BeeCon 2016



OCR Software: Windows

<https://github.com/Xandroid4Net/CommandLineOcr> (*non final*)

- **Windows.Media.Ocr**
 - Microsoft API runnable in Windows 8 and Windows 2012
 - Native in Windows 10 and Windows 2016



BeeCon 2016



OCR Software: Hosted services

<https://ocr.space/OCRAPI>

<http://www.ocrwebservice.com/api/restguide>

<http://www.bitocr.com/documentation.html>

...



Google Cloud Platform

<https://cloud.google.com/vision/>



BeeCon 2016

Real world use case (1)

OS Ubuntu 14.04 LTS

Version Alfresco 5.0.d

OCR soft pdfsandwich

Languages eng+spa+cat+fra



BeeCon 2016

Real world use case (2)

OS Ubuntu 15.10

Version Alfresco 5.0.d

OCR soft OCRmyPDF

Language eng





Real world use case (3)

OS Windows Server 2012 R2

Version Alfresco 5.1.e

OCR soft Windows.Media.Ocr

Language Spanish





Open Source OCR addon

<https://github.com/keensoft/alfresco-simple-ocr>

License LGPL v3.0

State Production

Languages (interface) English, Portuguese Brazilian, German and Spanish

Languages (OCR) 39 / 25

“No original Alfresco resources have been overwritten”

<https://github.com/OrderOfTheBee/addons/wiki/Inclusion-criteria-overview>



BeeCon 2016



OCR: Recap

- Generates automatically **PDF searchable** from PDF Image
- Open Source *addon* for Alfresco available
- Minimal configuration required
- Different Open Source Linux programs available
- Also Microsoft is providing the library *Windows.Media.Ocr*





Resources

GitHub

<http://github.com/keensoft/alfresco-simple-ocr>

Twitter

[@AngelBorroy](#)

Blog

<http://www.keensoft.es/en/category/blog-en/>

<http://angelborroy.wordpress.com>



BeeCon 2016



BeeCon 2016

Integrating a simple OCR in Alfresco

Angel Borroy
developer @ keensoft

