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
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
Building a simple OCR Action

1 REPO AMP

• Content model (simple)
• Action
• Transformer
OCR Action: Key classes

```xml
<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
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
OCR Action: Rule configuration

Only apply for foreground
OCR Action: Rule configuration

- **SYNCHRONOUS**
  - Run in background

- **ASYNCHRONOUS**
  - When:
    - or enter this folder
  - If all criteria are met:
    - All Items
  - Perform Action:
    - Extract OCR Continue on error: Yes
OCR Action: Results

1 INTRODUCTION
This document describes Alfresco Community 5.0.1 installation process at the dependencies, based on Ubuntu servers and Amazon infrastructures.

2 COMPONENTS CATALOG

2.1 SERVERS AND PORTS
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/
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
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
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
OCR Software: Hosted services

https://ocr.space/OCRAPI
http://www.ocrwebservice.com/api/restguide
http://www.bitocr.com/documentation.html

...
Real world use case (1)

**OS** Ubuntu 14.04 LTS

**Version** Alfresco 5.0.d

**OCR soft** pdfsandwich

**Languages** eng+spa+cat+fra
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

BeeCon 2016
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
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://angelborroy.wordpress.com
Integrating a simple OCR in Alfresco

Angel Borroy
developer @ keensoft