Wikiprint Book

Title: Instalacja OCRa dla faktur w systemie eDokumenty

Subject: eDokumenty - elektroniczny system obiegu dokumentów, workflow i CRM - DeployerGuide/Customization/OCRInvoice

Version: 67

Date: 05/06/26 00:25:13

Table of Contents

Instalacja OCRa dla faktur w systemie eDokumenty

Aktualna dokumentacja od wersji 6.53.0 znajduje się pod poniższym linkiem

OCR dla Faktur

Poniższa instrukcja przedstawia uruchomienie mechanizmu OCRowania faktur w systemie eDokumenty działających na systemie Linux. Mechanizm jest obsługiwany od wersji 5.2.77.

Poniższa instrukcja została przygotowana na bazie systemu Linux Debian9

```
apt-get update
```

```
apt-get -y install autoconf-archive automake g++ libtool libleptonica-dev pkg-config
apt-get -y install git
apt-get -y install poppler-utils
apt-get -y install libjpeg-dev libtiff-dev libpng-dev
apt-get -y install zbar-tools
```

Jeśli pakiety leptonica 1.74+ nie są dostępne w dystrybucji w takim przypadku, konieczna będzie komplikacja ze źródeł

```
mkdir /usr/lib/leptonica
cd /usr/lib/leptonica
wget https://github.com/DanBloomberg/leptonica/releases/download/1.85.0/leptonica-1.85.0.tar.gz
gunzip leptonica-1.85.0.tar.gz
tar -xf leptonica-1.85.0.tar
cd leptonica-1.85.0
./configure
make
make install
```

```
mkdir /usr/lib/tesseract
cd /usr/lib/tesseract
git clone https://github.com/tesseract-ocr/tesseract.git tesseract-ocr
cd tesseract-ocr/
./autogen.sh
./configure --disable-openmp
make
make install
ldconfig
```

```
cd /usr/local/share/tessdata/
wget https://github.com/tesseract-ocr/tessdata_fast/raw/master/script/Latin.traineddata
wget https://github.com/tesseract-ocr/tessdata_fast/raw/master/pol.traineddata
wget https://github.com/tesseract-ocr/tessdata_fast/raw/master/eng.traineddata
wget https://github.com/tesseract-ocr/tessdata_fast/raw/master/osd.traineddata
```

alternatywne źródło do pobrania:

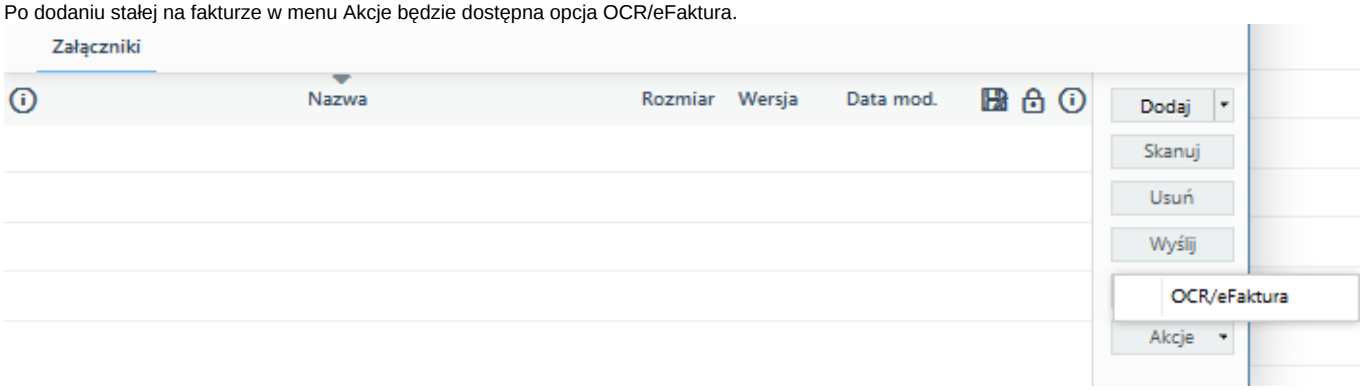
```
wget https://raw.githubusercontent.com/tesseract-ocr/tessdata/main/script/Latin.traineddata
wget https://raw.githubusercontent.com/tesseract-ocr/tessdata/main/pol.traineddata
wget https://raw.githubusercontent.com/tesseract-ocr/tessdata/main/eng.traineddata
wget https://raw.githubusercontent.com/tesseract-ocr/tessdata/main/osd.traineddata
```

Po pobraniu, zainstalowaniu oraz skompilowaniu pakietów ostatnim elementem jest dodanie stałej w config.inc domyślnie

```
vim /home/edokumenty/public_html/apps/edokumenty/config.inc
```

```
define('USE_NEW_OCR_FOR_EINVOICE', TRUE);
```

Po dodaniu stałej na fakturze w menu Akcje będzie dostępna opcja OCR/eFaktura.



Pakiety niezbędne do działania Bufora OCR - Python 3 (dla systemu Debian 10)

```
apt-get -y install rabbitmq-server
apt-get -y install python3-pip
apt-get -y install python3-pika
apt-get -y install python3-toml
apt-get -y install python3-pil
apt-get -y install python3-packaging
pip3 install pdfplumber --break-system-packages
pip3 install opencv-python-headless --break-system-packages
pip3 install pandas --break-system-packages

apt-get -y install supervisor

apt-get install pdftk
```

Przetwarzanie w tle (Bufor OCR)

Dotyczy Ready_ w wersji 6.52.1+

Wykorzystujemy supervisor do uruchomienia dwóch workerów (skrypty w języku Python), które znajdują się w katalogu domowym systemu (najczęściej: /home/edokumenty/bin).
Skrypty to: **worker_ocr.py** oraz **ocr_result.py**

Domyślne konfiguracje umieszczone są w katalogu home/edokumenty/etc/. Przed uruchomieniem należy usunąć z nazwy _default.

BUFFOR OCR osobna maszyna

W celu rozłożenie obciążenia, które w dużym stopniu generuje OCR możemy wydzielić go na osobną maszynę.

W tym celu na środowisku gdzie działa RabbitMQ tworzymy nowego użytkownika i nadajemy mu odpowiednie uprawniania:

```
rabbitmqctl add_user UZYTKOWNIK HASLO
rabbitmqctl set_user_tags UZYTKOWNIK administrator
rabbitmqctl set_permissions -p / UZYTKOWNIK ".*" ".*" ".*"
```

Następnie dane do nowo utworzonego konta uzupełniamy w konfiguracji na maszynie eDokumentyOCR

```
vim /home/edokumenty/etc/rabbimq.toml
```

Po uzupełnieniu danych konieczny jest restart workerów

```
supervisorctl reload
```

Znane problemy

1. Brak pakietu libpng12.so.0. W logach OCR pojawia się komunikat:

tesseract: error while loading shared libraries: libpng12.so.0: cannot open shared object file: No such file or directory

Sprawdzamy czy pakiet istnieje:

```
ls -ld $(locate -r libpng.*\.so.*)
```

Komenda powinna zwrócić nam:

```
lrwxrwxrwx 1 root root     19 kwi 18 22:12 /usr/lib/x86_64-linux-gnu/libpng16.so -> libpng16.so.16.28.0
lrwxrwxrwx 1 root root     19 kwi 18 22:12 /usr/lib/x86_64-linux-gnu/libpng16.so.16 -> libpng16.so.16.28.0
-rw-r--r-- 1 root root 206768 kwi 18 22:12 /usr/lib/x86_64-linux-gnu/libpng16.so.16.28.0
lrwxrwxrwx 1 root root     11 kwi 18 22:12 /usr/lib/x86_64-linux-gnu/libpng.so -> libpng16.so
```

Jeśli otrzymamy taką informację konieczne będzie ponowne kompilowanie leptonici oraz tesseract

Kompilowanie tesseract dla 1 wątku

```
./configure --disable-openmp
```

1. Problem z convertowanie jpg do PDF

W logach php mamy komunikat

```
[23-Sep-2020 12:28:46 Europe/Warsaw] ReadyCls\OCR\OcrEngine - pdftoppm fails with message: [1]
```

lub

```
convert-im6.q16: attempt to perform an operation not allowed by the security policy `PDF' @ error/constitute.c/IsCoderAuth
```

W pliku /etc/ImageMagick-6/policy.xml należy zakomentować linię

```
<!-- <policy domain="coder" rights="none" pattern="PDF" /> -->
```

UWAGA!!"' Ten plik najpewniej zostanie przywrócony do pierwotnej wersji przy każdym upgrade pakietu ImageMagick. Trzeba pamiętać aby po upgrade serwera ponownie to zakomentować.

POPPLER testowanie poprawności instalacji

Komenda do przeprowadzenie testu popplera.

```
pdftotext -bbox-layout NAZWAPLIKUWEJSCIOWE.pdf NAZWAPLIKUWYJSCIOWEGO.html
```