Adding OCR capability

OCR components must be installed on the same server as TS file indexing service.

Only GhostScript and Terrasect are required to proces PDF files.

TEMPORARY FIX: <tomcat>\catalina\catalina.properties add java.io.tmpdir=c:/Temp

Install: ImageMagick binaries

Download and unpack "portable" version (recommended c:\ImageMagick)

```
https://www.imagemagick.org/script/binary-releases.php
```

Register the location of the **convert** executeable in web.xml

```
<context-param>
   <param-name>ExecutableImageMagick</param-name>
   <param-value>c:\ImageMagick\convert</param-value>
</context-param>
```

Leaving the entry empty will prevent OCR handling of image files: png, jpg, jpeg

Install: Ghostscript binaries

Download and run installer

```
http://www.ghostscript.com/download/gsdnld.html
```

Note: You are not required to buy a license

Register the location of the **gswin64c** executeable in web.xml

```
<context-param>
   <param-name>ExecutableGhostscript</param-name>
   <param-value>c:\Program Files\gs\gs9.20\bin\gswin64c.exe</param-value>
</context-param>
```

Leaving the entry empty will prevent OCR handling of PDF files

Install: Tesseract binaries

For linux just use install from repository using

```
sudo yum install tesseract-ocr
```

If you are using Amazon linux please use this instead (thanks for help).

```
sudo yum --enablerepo=epel --disablerepo=amzn-main install libwebp
sudo yum --enablerepo=epel --disablerepo=amzn-main install tesseract
```

For Windows download installer or zip archieve

```
https://sourceforge.net/projects/tesseract-ocr-alt/files/
```

Register the location of the **tesseract** executeable in web.xml

```
<context-param>
   <param-name>ExecutableTerrasect</param-name>
   <param-value>c:\tesseract\tesseract</param-value>
</context-param>
```

Revision #1
Created 10 April 2025 13:37:02 by Theis Villumsen
Updated 10 April 2025 13:37:25 by Theis Villumsen