

Lav OCR-behandling af pdf-filer

Der skal spørges om hvilken sti der skal behandles filer i. Bat-rutine, som:

- Går igennem alle .pdf-filer i en mappe (og evt. undermapper).
- Checker om filen allerede indeholder søgbar tekst (dvs. OCR-læsbart).
- Hvis den ikke er OCR-læsbart, laver den en OCR-version vha. tesseract.
- Kopierer alle OCR-læsbart originale filer til en undermappe Pdf-ocr/.
- Lægger alle OCR-konverterede filer i samme Pdf-ocr\<relativ sti>\ mappe.

VEJLEDNING TIL OPSÆTNING (til den tekniske person)

Denne vejledning forklarer, hvordan systemet sættes op første gang.

☒ 1. Installer nødvendige programmer

• Tesseract OCR

- Hent fra: <https://github.com/tesseract-ocr/tesseract>
- Installer med standardindstillinger
- Vigtigt: Sæt flueben ved "*Additional language data*" og vælg f.eks. **danish**

• Ghostscript

- Hent fra: <https://www.ghostscript.com/download/gsdnld.html>
- Installer 64-bit version
- Standardinstallationen er fin

Sørg for at følgende to filer findes:

- `tesseract.exe` (f.eks. i `C:\Program Files\Tesseract-OCR\`)
 - `gswin64c.exe` (f.eks. i `C:\Program Files\gs\gs10.1.2\bin\`)
-

☒ 2. Tilføj programmer til systemets PATH

Så de kan kaldes direkte fra .bat-filen.

Sådan gør du:

1. Højreklik på **Denne computer > Egenskaber**
2. Klik på **Avancerede systemindstillinger**
3. Klik på **Miljøvariabler**
4. Find **Path** under "Systemvariabler" og klik "Rediger"
5. Tilføj følgende (hvis de findes):

```
C:\Program Files\Tesseract-OCR  
C:\Program Files\gs\gs10.1.2\bin
```

☒ 3. Læg .bat-filen et synligt sted

Gem pdf-ocr_behandling.bat et sted som fx:

```
C:\BrukerScripts\pdf-ocr_behandling.bat
```

Eller blot **på Skrivebordet**, så brugere nemt kan dobbeltklikke.

☒ 4. Test installationen

Åbn en kommandoprompt og skriv:

```
tesseract -v  
gswin64c -v
```

Du skal få versionsinfo. Hvis ikke: PATH er ikke sat korrekt.

☒ 5. Indhold af .bat-fil

Sørg for at .bat-filen indeholder denne version – se filen pdf-ocr_behandling

(Hvis du ikke har den, kan jeg sende hele indholdet igen)