Unter LINUX eine Textebene zu PDF-Dateien hinzufügen

(Idee aus LINUX-Welt 01/2023, Seite 104)

Einleitung

Für die meisten Drucker gibt es unter LINUX keine vom Hersteller gelieferte Tools, um beim Einscannen von Dokumenten durchsuchbare PDF-Dateien zu erhalten. Die Texte werden immer nur als Bilder in den PDFs eingebettet. Das macht es schwierig später etwas Wichtiges wiederzufinden. Natürlich gibt es auch unter LINUX entsprechende Tools, aber diese sind oft nur an der Kommandozeile zu bedienen. Das macht es für von grafischen Oberflächen verwöhnte Benutzer schwierig, damit zurecht zu kommen. Außerdem vergisst man ja oft die Befehle.
Deshalb hier eine einfache Grafische Oberfläche, um diese Arbeit mit ein paar Klicks zu erledigen.

Vorbereitungen

Zuerst die benötigten Programme installieren:

- ImageMagick (sollte bereits installiert sein)
- pdfsandwich
- tesseract-ocr

```
sudo apt install pdfsandwich tesseract-ocr-de
```

Dann noch ImageMagick das Recht geben, PDF-Dateien zu editieren:

```
sudo nano /etc/ImageMagick-6/policy.xml
```

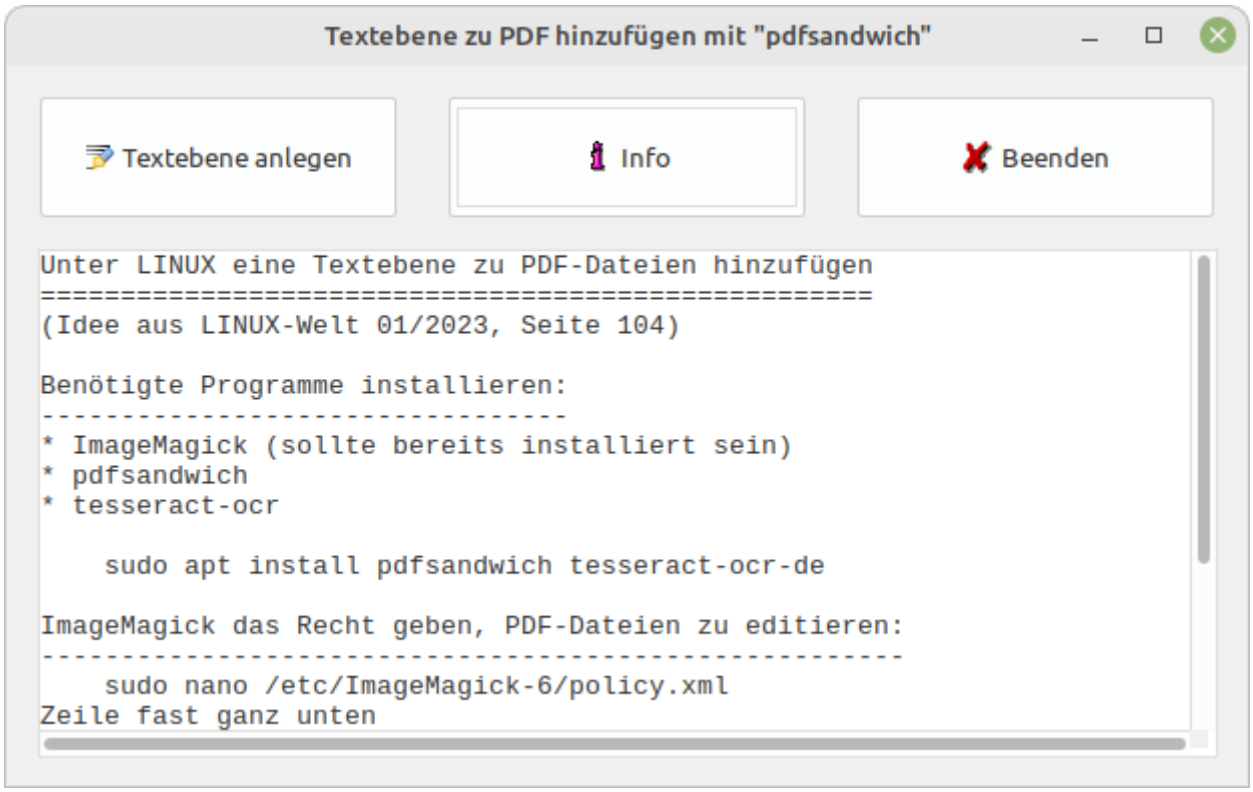
In der Zeile fast ganz unten
<policy domain="coder" rights="*none*" pattern="PDF" />
mit
<policy domain="coder" rights="*read*|*write*" pattern="PDF" />
überschreiben. Speichern mit **Strg+O** und Beenden mit **Strg+X**.

Das Programm **"pdftext"** irgendwo ablegen, z.B. in **`/usr/bin`**.

Eventuell zum Starten aus den Startmenü eine .desktop Datei dazu anlegen.

PDFTXT nutzen

Die GUI ist sehr einfach und versteht sich eigentlich von selbst:



Mit **"Textebene anlegen"** wird man aufgefordert, PDF-Dateien zum Bearbeiten auszuwählen. Man kann mehrere PDF-Dateien auswählen und alle mit einem Rutsch bearbeiten lassen.
Es geht allerdings auch, PDF-Dateien zum Bearbeiten aus dem Dateimanager per **Drag & Drop** auf das Programmfenster zu ziehen.

Das Ergebnis der Bearbeitung wird im gleichen Verzeichnis wie die Ursprungsdatei in eine Datei mit dem Zusatz "_ocr" gespeichert.