Begleitinformation zu

Bücher & Zeitschriften digitalisieren und aufbereiten u. a. mit ScanTailor unter Linux

Vorgetragen von Benno Vock am 18.09.2024
beim VDI Bezirksverein Schwarzwald e.V. / Freiburger Linux User Gruppe
Mailkontakt: btux@posteo.org

1. Warum Dokumente, Bücher und Zeitschriften digitalisieren?

Es gibt verschiedene Motivationen warum analoge Schriftstücke digitalisiert werden:

- Sicherung vor Verfall / Diebstahl / Feuer
- Platzsparende Aufbewahrung / leichter Transport
- Volltextsuche im Dokument
- Vereinfachte Weitergabe - Achtung: Urheberrecht beachten!!

2. Das Ergebnis: Wie und wofür?

Ziel der Digitalisierung sollte es sein, dass die Dokumente plattformübergreifend zugänglich sind.

- Hier bietet sich das PDF-Format als derzeit maximal zukunftsicher an.
- Die Dokumente sollen leicht verfügbar sein, z. B. Auf einem Tablet mobil nutzbar.
- Kein Zwang zur Nutzung von Online-Speicher / Cloud

3. Unterscheidung der Vorlagen

Wertvolle Vorlagen:

- Zerstörungsfrei mit Flachbettscanner oder Kamera
- Sehr langsames Verfahren!!

Vorlagen, die „körperlich“ nicht mehr erforderlich sind:

- Mechanische Aufbereitung für Massenbearbeitung im Einzugscanner.
- Anschließend Entsorgung der Papiervorlage (Buchliebhaber müssen hier seeehr stark sein)

4. Mein (Massen-) Workflow

1. Vorlage, d. h. Buch oder Zeitschrift aufbereiten
   - Scannen mit einem Einzugscanner, der 2-seitig scannt als PDF
2. Bestenfalls ist die Arbeit hier bereits erledigt. Weiter mit 4
3. Schwierige Vorlagen (z. B. durchscheinende Seiten) werden mit ScanTailor nachbearbeitet (leider nur schwarz/weiß möglich).
4. Texterkennung als Ebene über das PDF-Bild erstellen
   - z. B. Mit Tesseract / OCRmyPDF o. ä.

4.1 Vorlage, d. h. Buch oder Zeitschrift aufbereiten

- Bei Büchern und gebundenen Zeitschriften:
  - Buchdeckel entfernen und Bindung mit Hebelschere abtrennen.
- Bei gehefteten Zeitschriften:
  - Klammern entfernen und Heft in der Mitte aufschneiden.

4.2 Scannen mit einem Einzugscanner

- Scannen mit Hilfe von VueScan weil:
  - Umfangreiche Einstellmöglichkeiten sowohl bezogen auf Größe als auch Belichtung / Farbe.
    - Hier ist die Funktion „Extremwerte autom.“ in der Rubrik „Farbe“ hervorzuheben.
    - Mit den Einstellungen experimentieren!!!
  - Schiefe Scans oder zusammenklebende Seiten lassen sich einfach wiederholen ohne neu beginnen zu müssen
- Scannen mindestens als Graustufen, besser Farbe

4.3 Schwierige Vorlagen

- Schwierige Vorlagen (z. B. durchscheinende Seiten) werden mit ScanTailor nachbearbeitet:
  - Da ScanTailor nur Bilddateien bearbeiten kann, muss das gescannte PDF-Dokument in Bilddateien (vorzugsweise TIF) umgewandelt werden (z. B. Mit PDF-XChange).
    - Wichtig: Keine mehrseitigen TIFs
- ScanTailor kann nur schwarz/weiß ausgeben. Farbige Texte oder Zeichnungen müssen ausgespart werden.
- Scantailor gibt als Ergebnis TIF-Dateien aus.
- TIF-Dateien müssen wieder in PDF umgewandelt werden.
  - Bei mir: PDF-XChange

4.4 Texterkennung

- Texterkennung immer als extra Ebene in PDF
  - Sonst geht die Formatierung verloren.
- OCRmyPDF ist ein sehr mächtiges Werkzeug und nutzt im Hintergrund Tesseract
  - Verschiedene Parameter beachten!
    - Vorsicht vor den Optimierungsfunktionen, denn PDFs werden ggf. verschlimmbessert.
  - Alternative: PDF-XChange

5. Links

Vergleich ScanTailor vs. Abby Finereader: https://www.youtube.com/watch?v=IM1EqJ3MCII

Der größte Teil des Videos beschäftigt sich mit ScanTailor und ist ein sehr gutes Schulungsvideo dazu.
VueScan: https://www.hamrick.com/de/

Universelles Scanprogramm mit vielen Funktionen, die Massenscans vereinfachen.

Scanner: https://www.brother.de/scanner/ads-1800w

Nachfolger des von mir verwendeten Dokumentenscanners (ADS-1700W). Trotz seiner Größe sehr leistungsfähig und schnell.

ScanTailor Advanced: https://flathub.org/apps/com.github._4lex4.ScanTailor-Advanced

Deutlich leistungsfähiger als ScanTailor, daher von mir verwendet

PDF-Xchange: https://pdf-xchange.de/

Das Programm läuft mittels wine auf sehr gut unter Linux.

OCRmyPDF: https://github.com/ocrmypdf/OCRmyPDF

Das leistungsfähige OCR-Programm dürfte in jedem Linux-Repository vorhanden sein.

Darktable (Linux-Version im Repository): www.darktable.org aktuelle Version über Flatpak

Bietet sich an, wenn Vorlagen als Bilder aufgearbeitet werden sollen – sofern VueScan das nicht bereits erledigen konnte.

Weitere Links, die zwar nicht zum Vortrag gehören, aber meiner Meinung nicht fehlen sollten:

pi-hole, ein System für einen Raspberry Pi u. a. um unerwünschte Webseiten zu unterbinden: https://pi-hole.net/

Kuketz-Blog: Bei Fragen zum Datenschutz immer einen Besuch wert, besonders die „Empfehlungsecke“ www.kuketz-blog.de/empfehlungsecke/

Computertruhe e. V. www.computertruhe.de gibt gespendete, gebrauchte Hardware an bedürftige Menschen und gemeinnützige Organisationen weiter.