OCR mit freier Software

Scannen und dann?

Worum geht es?

- OCR – Optical Character Recognition
- Automatisches Extrahieren von Text aus Bildern
- Für verschiedene Zeichensätze, Sprachen und Schriften
- mit/ohne Erhalt des Layouts

Voraussetzung/Vorbearbeitung

- Möglichst guter Scan bzw. gutes Foto
  - Ausreichende Auflösung (200..400dpi)
  - Ausrichtung (Landscape, Kopfstand)
  - Geraderichten
  - Kontrasterhöhung, z.B. bei Schwarz auf Grau
  - Geeignetes Dateiformat

Aufgabenstellung

- Umsatzsteuerrecht → Abliefernachweis EU
- ca. 200 Lieferungen je Monat
- Versandbescheinigungen von Speditionen und Packetdiensten
- (halb-)automatische Zuordnung zu eigenen Lieferscheinen
- Stapelscanner → FTP-Server
- PDF-Format, 300dpi, schwarz/weiß

OCR-Software für Linux

- Kommerziell: Kofax, Abby, ...
- Frei, Debian (apt-cache search ocr)
  - gocr
  - ocrad
  - tesseract
  - cuneiform

gocr

- Verarbeitet nur pnm-Files
- convert aus imagemagick benutzen
- Verschiedene Ausgabeformate (ISO8859_1, TeX, HTML, UTF8, XML, ASCII)
- Mäßige Erkennungsleistung
- Trainierbar, nicht ausprobiert

ocrad

- Verarbeitet pnm-Files, wie gocr
- Schnell!
- Keine Abhängigkeiten
- Direkt als Filter benutzbar
- Mäßige Erkennungsleistung

cuneiform

- Von russ. Firma Cognitive Technologies
- Seit 2008 unter BSD-Lizenz
- Geforkte Kommandozeilen-Version
- verarbeitet gängige 1-Seiten-Bildformate
- Gute Erkennungsleistung
- Kann rtf-Dateien erzeugen → Textverarbeitung
- Erkennung für >20 (europäische) Sprachen

tesseract

- 1985..1995 von HP entwickelt
- 2005 von Google überarbeitet und freigegeben
- Apache-Lizenz
- apt-get install tesseract-ocr tesseract-ocr-deu
- tesseract -l deu input.png outputbase hocr
- Gute Erkennungsleistung
- Ziemlich langsam

hocr

- cuneiform und tesseract können hocr-Format erzeugen
- HTML mit Positionsinformationen
- Basis für layouterhaltende Weiterverarbeitung
- cuneiform → leider buggy
- Mehrere Konverter
- z.B. hocr2pdf aus exactimage

OCRmyPDF

- https://github.com/fritz-hh/OCRmyPDF
- Version 2.0 brandneu
- Version 1.1 benutzt, geht mit wheezy und nach etwas Fummelei sogar mit squeeze
- tesseract aus wheezy: tw. kaputtes UTF-8
- Patch mit Workaround
- OCRmyPDF input.pdf output.pdf
- Lieferte bei mit bessere Resultate als hocr2pdf