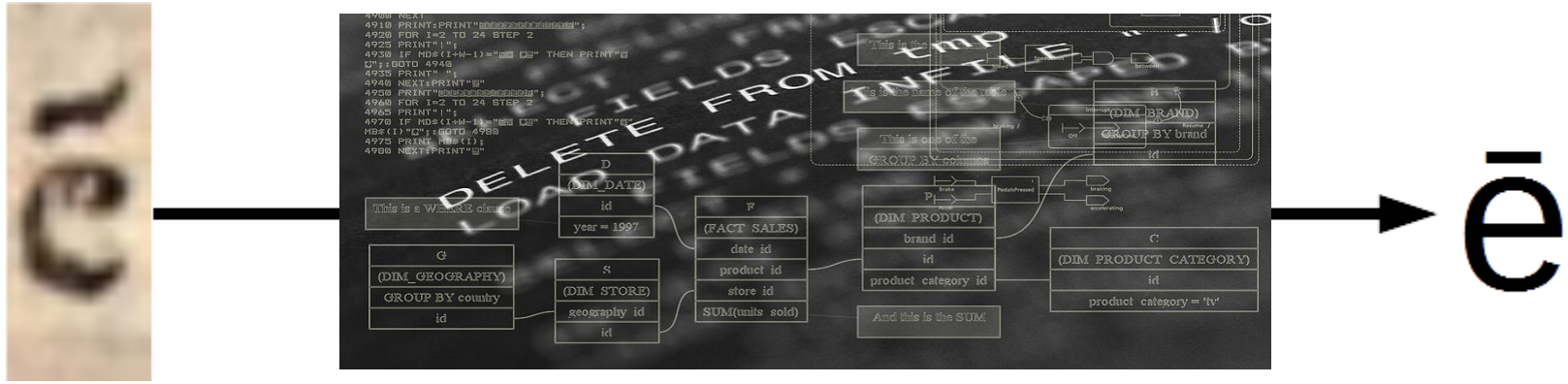
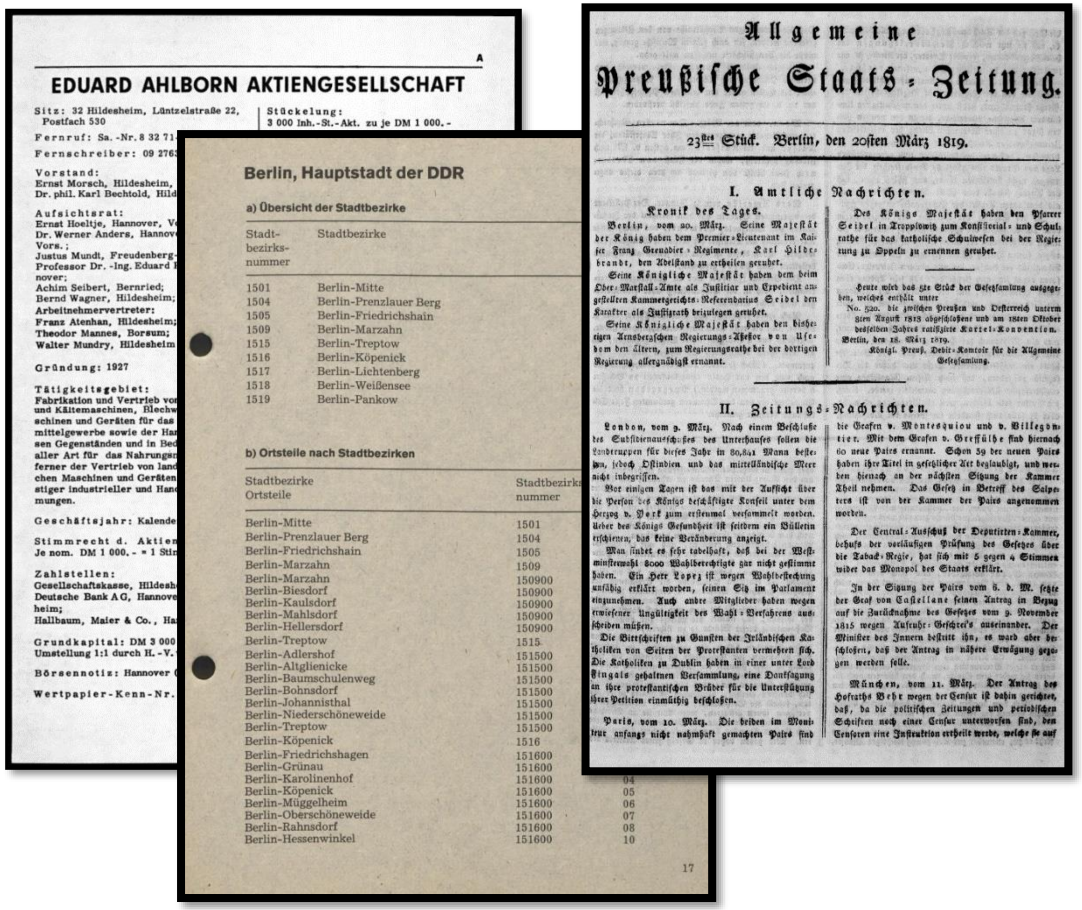
Vom Bild zum Text.
Automatisierte Texterkennung in historischen Drucken mit der freien Software Tesseract.



Stefan Weil, Philipp Zumstein, Universitätsbibliothek Mannheim

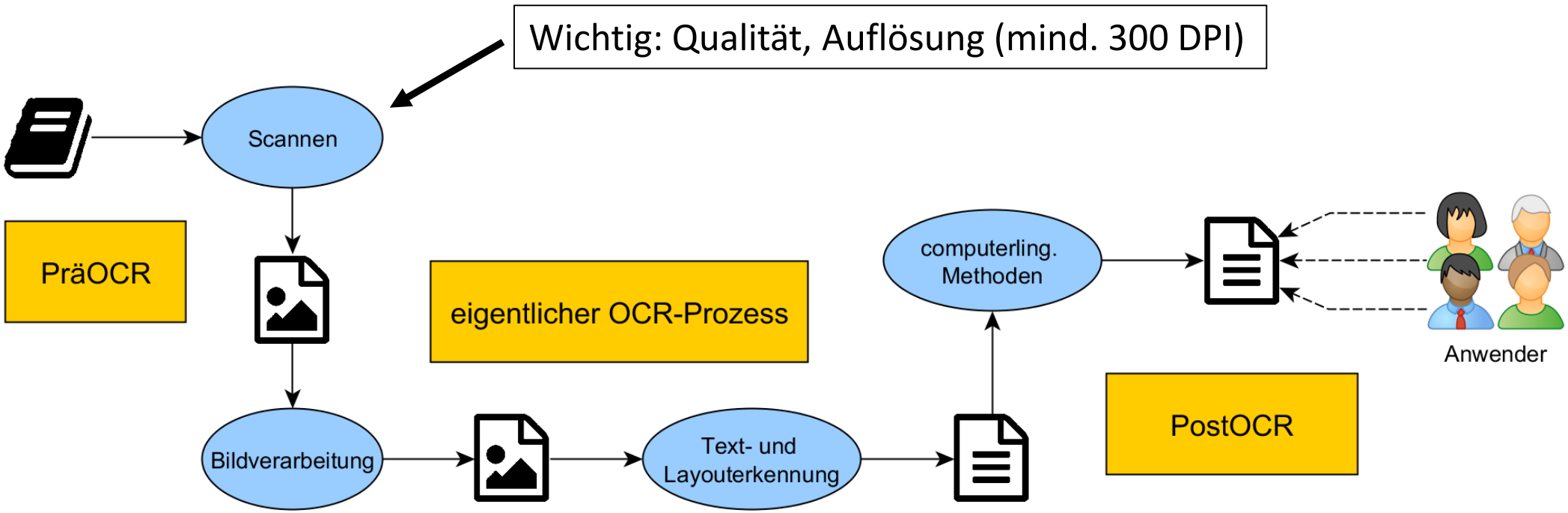
Motivation – Warum Texterkennung?
(OCR = „Optical character recognition“)



- Recherchemöglichkeit (Volltextsuche)
- Datenextraktion für Forschungsprojekte
- Linguistische Analysen

z. B. bei der UB Mannheim für Hoppenstedt Aktienführer, Gemeindeverzeichnisse und Preußische Staatszeitung / Deutscher Reichsanzeiger

Prozesskette



aus: Baierer, Zumstein. Verbesserung der OCR in digitalen Sammlungen von Bibliotheken)

OCR Software (Übersicht)

kommerzielle Software

fett = eingesetzt in Bibliotheken

ABBYY Finereader
BIT-Alpha
Readiris
OmniPage

Adobe Acrobat
CorelDraw
Microsoft OneNote
…

Tesseract
OCRopus / Kraken / Calamari
Ocrad
CuneiForm …

freie Software

ABBYY Cloud OCR
Google Cloud Vision
Microsoft Azure Computer Vision
OCR.space Online OCR …

Cloud OCR

Tesseract OCR

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Open Source
- Komplettlösung „All-in-one“
- Mehr als 100 Sprachen / mehr als 30 Schriften
- Liest Bilder in allen gängigen Formaten (nicht PDF!)
- Erzeugt Text, PDF, hOCR, ALTO, TSV
- Große, weltweite Anwender-Community
- Technologisch aktuell (Texterkennung mit neuronalem Netz)
- Aktive Weiterentwicklung u. a. im DFG-Projekt OCR-D

Tesseract an der UB Mannheim

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Verwendung im DFG-Projekt „Aktienführer“ https://digi.bib.uni-mannheim.de/aktienfuehrer/
- Volltexte für Deutscher Reichsanzeiger und Vorgänger https://digi.bib.uni-mannheim.de/periodika/reichsanzeiger

- DFG-Projekt „OCR-D“ http://www.ocr-d.de/, Modulprojekt „Optimierter Einsatz von OCR-Verfahren – Tesseract als Komponente im OCR-D-Workflow“: Schnittstellen, Stabilität, Performance und praktische Einsetzbarkeit

OCR-D
Koordinierte Förderinitiative zur Weiterentwicklung von Verfahren der Optical Character Recognition (OCR)

Tesseract – Installation

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Linux 
  Tesseract 4 ist Bestandteil der gängigen Linux-Distributionen

- Windows 
  Tesseract-Installer der Universitätsbibliothek Mannheim
  https://github.com/UB-Mannheim/tesseract/wiki

- macOS 
  Installation über MacPorts oder Homebrew 

- Beschreibung: https://github.com/tesseract-ocr/tesseract/wiki

Tesseract – Sprachen und Schriften

- Tesseract 4 unterscheidet zwischen „Sprachen“ und „Schriften“:
  - Eine Sprache wie beispielsweise **deu** kennt die typischerweise im Deutschen verwendeten Schriftsymbole und enthält u. a. ein deutsches Wörterbuch.
  - Eine Schrift wie beispielsweise **script/Latin** kennt idealerweise alle Schriftsymbole dieser Schrift und ein Wörterbuch verschiedener Sprachen, die diese Schrift nutzen. Das ist neu in Tesseract 4 und bietet Vorteile in vielen typischen Texten.
- Übersicht: Kapitel „Languages and Scripts“ im Tesseract-Handbuch https://digi.bib.uni-mannheim.de/tesseract/manuals/tesseract.1.html
- Drei Ausprägungen für jede Sprache bzw. Schrift:
  - Optimiert für Geschwindigkeit: https://github.com/tesseract-ocr/tessdata_fast
  - Genauigkeit + Training: https://github.com/tesseract-ocr/tessdata_best
  - Zwei OCR-Erkennungsalgorithmen: https://github.com/tesseract-ocr/tessdata

Tesseract – Support

- Dokumentation
  - Tesseract Wiki: http://github.com/tesseract-ocr/tesseract/wiki
  - UB Mannheim Wiki: https://github.com/UB-Mannheim/tesseract/wiki
  - Handbuchseiten: https://digi.bib.uni-mannheim.de/tesseract/manuals/

- Anwenderforum: http://groups.google.com/group/tesseract-ocr
- Entwicklerforum: http://groups.google.com/group/tesseract-dev

- Ticketing-System (in der Regel für Fehlermeldungen)
  - Tesseract: https://github.com/tesseract-ocr/tesseract/issues
  - Tesseract für Windows: https://github.com/UB-Mannheim/tesseract/issues

PDF-Dateien

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- All-in-one Lösung **OCRmyPDF**
  https://github.com/jbarlow83/OCRmyPDF
- Xpdf **XpdfReader** und **Xpdf command line tools**
  http://www.xpdfreader.com/
  - `pdfimages myfile.pdf image`
    erzeugt image-0000.ppm, image-0001.ppm, image-0002.ppm, …
  - `pdftopng myfile.pdf image`
    erzeugt image-000001.png, image-000002.png, image-000003.png, …
  - Diese Bilder können von Tesseract gelesen werden.

hOCRjs

Zweck: Visualisierung des erkannten Layouts,
überlagerte Darstellung von Digitalisat und erkanntem Text

- Website: http://kba.cloud/hocrjs/
- GitHub: https://github.com/kba/hocrjs
- Beispiel: http://kba.cloud/hocrjs/example/426117689_0459.html
- Beispiel: https://digi.bib.uni-mannheim.de/periodika/reichsanzeiger/ocr/film/tesseract-4.0.0-20181201/139-9547/0316.hocr?overlay=1

ScanTailor

Zweck: Digitalisate verbessern für die Texterkennung

- Website: https://scantailor.org/
- GitHub: https://github.com/scantailor/scantailor

Ground Truth

Manuell erfasster Text, idealerweise fehlerfrei.

- Verwendet für Qualitätsmetriken
  - Wortfehlerrate (WER) 0…100 % bzw. Wortgenauigkeit 100…0 %
  - Zeichenfehlerrate (CER) 0…100 % bzw. Zeichengenauigkeit 100…0 %
- Verwendet zum Trainieren der OCR-Software
- Texterfassung ist kein einfaches Abschreiben, siehe die Ground Truth Guidelines für OCR-D: http://www.ocr-d.de/gt_guidelines
- Beispiele:
  - Ground Truth Korpus von OCR-D: http://www.ocr-d.de/daten
  - UB Mannheim: https://digi.bib.uni-mannheim.de/fileadmin/digi/453089178/gt/ https://github.com/UB-Mannheim/Reichsanzeiger/wiki/Text-recognition#accuracy

Nützliche Programme

- hOCRjs
  http://kba.cloud/hocrjs/ 
- ScanTailor
  https://scantailor.org/ 
- Gimp
  https://www.gimp.org/ 
- OCRmyPDF
  https://github.com/jbarlow83/OCRmyPDF 
- IrfanView (privat und in Bildungseinrichtungen frei nutzbar) 
  https://www.irfanview.com/

Linksammlung

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Wikipedia-Artikel
  - https://de.wikipedia.org/wiki/Tesseract_(Software)
  - https://de.wikipedia.org/wiki/Texterkennung
  - https://en.wikipedia.org/wiki/Comparison_of_optical_character_recognition_software
- Wikis auf GitHub
  - https://github.com/tesseract-ocr/tesseract/wiki
  - https://github.com/UB-Mannheim/tesseract/wiki
- Umfangreiche Sammlung von Links zum Thema OCR
  - https://github.com/kba/awesome-ocr

Literatur

- Weil, S., & Zumstein, P. (2016). Mit freier Software Text in Digitalisaten erkennen. https://speakerdeck.com/zuphilip/mit-freier-software-text-in-digitalisaten-erkennen-ocr-praxis-an-der-ub-mannheim
- Baierer, K., & Zumstein, P. (2016). Verbesserung der OCR in digitalen Sammlungen von Bibliotheken. *027.7 Zeitschrift für Bibliothekskultur / Journal for Library Culture, 4*(2), 72-83. https://doi.org/10.12685/027.7-4-2-155

Bildnachweis

- Apple Logo, Microsoft Logo, Linux Logo, Logos für MacPorts und Homebrew: Wikimedia Commons https://commons.wikimedia.org/
- OCR-D Logo: http://www.ocr-d.de/
- Bilder zu ScanTailor: https://github.com/scantailor/scantailor, https://github.com/scantailor/scantailor/wiki/User-Guide
- Titelbild: Weil, S., & Zumstein, P. (2016)
- Prozesskette: Baierer, K., & Zumstein, P. (2016)