EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

2. OCR-BW-Workshop –
Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen

Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

09.06.2021

Projektziele

„Aufbau eines Kompetenzzentrums für Volltexterschließung von handschriftlichen und gedruckten Werken.“

Das Projekt OCR-BW unterstützt Archive, wissenschaftliche Bibliotheken und andere Institutionen in Baden-Württemberg bei der Anwendung von automatischer Texterkennungs- und Transkriptionssoftware.

UB Tübingen:	Transkription und Volltexterschließung von Autographen, Handschriften und Inkunabeln
UB Mannheim:	Volltexterkennung (OCR) von Druckwerken

➢ **https://ocr-bw.bib.uni-mannheim.de/**

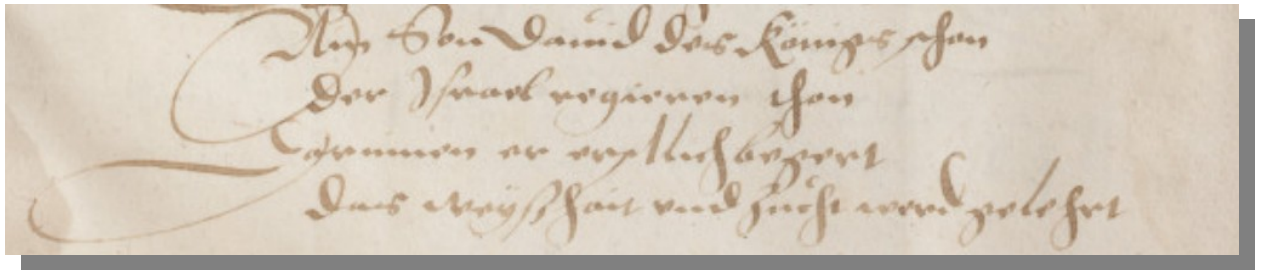
OCR-BW

EBERHARD KARLS UNIVERSITÄT TÜBINGEN

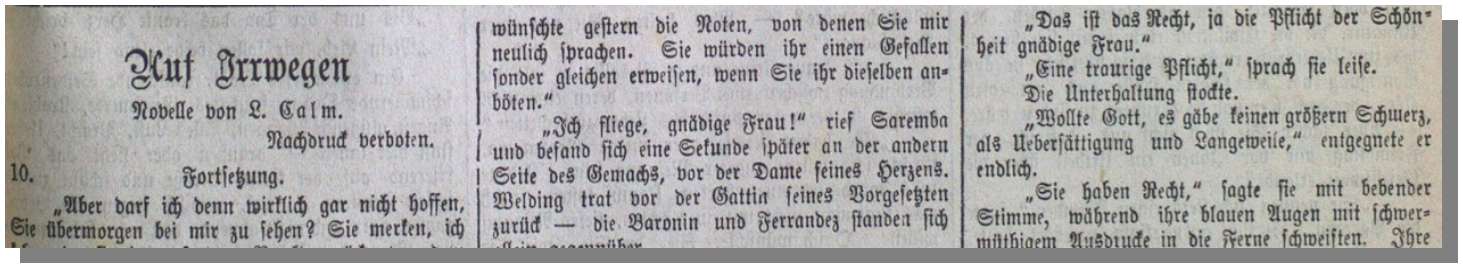
UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Arbeitsbeispiele für Transkription und Texterkennung

- Handschrift aus dem Bestand der WLB Stuttgart (16. Jhd.)



- Historische Zeitungen (19.–20. Jhd.) für das Stadtarchiv Ladenburg, das MARCHIVUM und die BLB Karlsruhe



Auf Irrwegen

Novelle von L. Calm.

Nachdruck verboten.

10. Fortsetzung.

„Aber darf ich denn wirklich gar nicht hoffen, Sie übermorgen bei mir zu sehen? Sie merken, ich wünschte gestern die Noten, von denen Sie mir neulich sprachen. Sie würden ihr einen Gefallen sonder gleichen erweisen, wenn Sie ihr dieselben anböten."

„Ich fliege, gnädige Frau!" rief Saremba und befand sich eine Sekunde später an der andern Seite des Gemachs, vor der Dame seines Herzens. Welding trat vor der Gattin seines Vorgesetzten zurück — die Baronin und Ferrandez standen sich

„Das ist das Recht, ja die Pflicht der Schönheit gnädige Frau."

„Eine traurige Pflicht," sprach sie leise.

Die Unterhaltung stockte.

„Wollte Gott, es gäbe keinen größern Schmerz, als Uebersättigung und Langeweile," entgegnete er endlich.

„Sie haben Recht," sagte sie mit bebender Stimme, während ihre blauen Augen mit schwermüthigem Ausdrucke in die Ferne schweiften. Ihre

OCR-BW

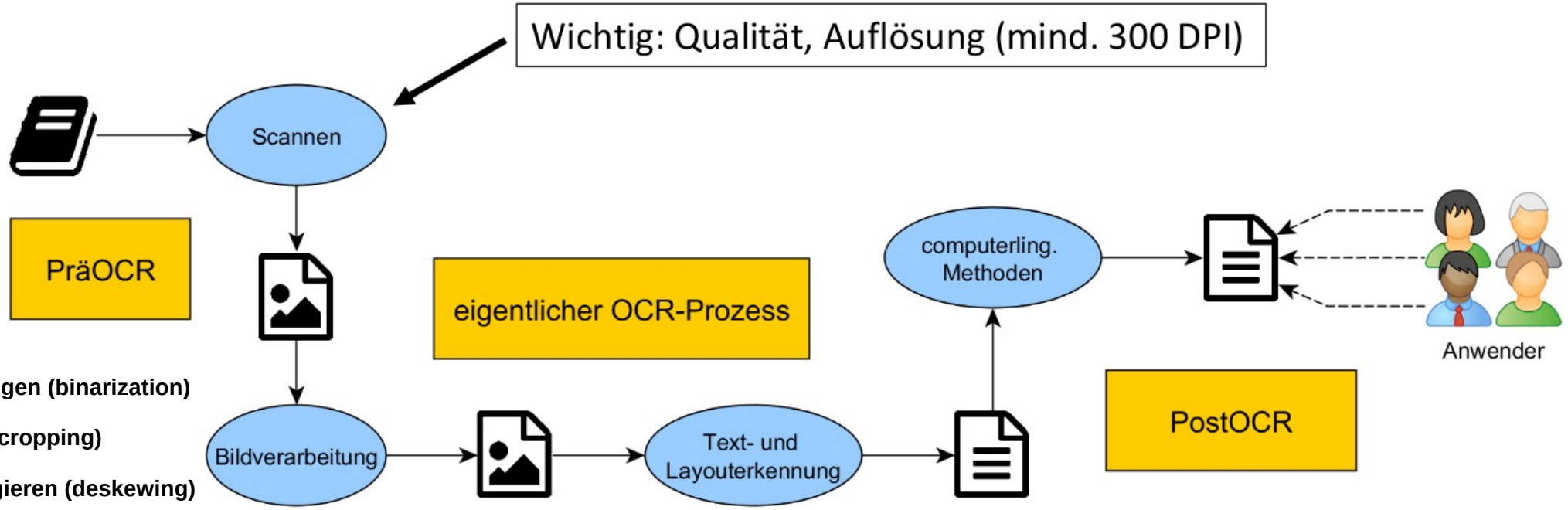
Werkstattbericht und Ausblick der UB Mannheim

Stefan Weil, Jan Kamlah
Universitätsbibliothek Mannheim

09.06.2021

Gefördert durch
Baden-Württemberg
MINISTERIUM FÜR WISSENSCHAFT, FORSCHUNG UND KUNST

Vom Werk zum Digitalisat mit Volltext



S/W-Bild erzeugen (binarization)

Beschneiden (cropping)

Neigung korrigieren (deskewing)

Entzerren (dewarping)

Flecken entfernen (despeckling)

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

Unterstützte Vorhaben und Projekte

- Unterstützung und Beratung von mehr als einem Dutzend Einrichtungen des Landes (Stadt-, Kreis-, Landesarchiv, Universitätsbibliotheken und Museen) sowie Einzelanfragen von Wissenschaftlern und Privatpersonen
- Größere Unterstützungsvorhaben
  - Badische Landesbibliothek (Unterstützung und Beratung bei der Integration und Anwendung von Tesseract)
  - MARCHIVUM (Unterstützung und Beratung bei der Erschließung von historischen Zeitungen)
- Vielen Anfragen konnte durch eine kurze Anleitung zur Anwendung von automatischer Texterkennung sowie der Auswahl geeigneter Modelle geholfen werden
- Die meisten Anfragen betrafen Drucke aus dem Zeitraum 1930–1950, sowie historische Drucke aus dem 17.–19 Jahrhundert

Unterstützte Vorhaben und Projekte

- Unterstützte Projekte
  - Tesseract
    - Verbesserung der Performance und Stabilität
    - neue Modelle für Fraktur-Schriften
  - OCR-D
  - FDMLab (Landesarchiv Baden-Württemberg)
  - Akademie-Projekt (historische Finanz- und Wirtschaftsdaten, Vorprojekt)

Softwaretechnische Werkzeuge

- Ground-Truth
  - GTMake (Automatisches Erstellen von GT-Lines)
  - GTCheck (Nachkorrektur der GT per Weboberfläche)
  - GTReval (Automatische Korrektur der GT basierend auf speziell trainierten Modellen und Regeln)
- Texterkennung
  - ocrd_all & ocrd_pagetopdf (Beiträge zum OCR-D Projekt)
  - BackgroundSubtraction4OCR (Reduktion des Bildhintergrunds)
  - TesseractXplore (GUI für Tesseract)

Verbesserte Ground Truth und neue Modelle für Tesseract

Ground Truth

- Neue Datensätze: Fibeln, Weisthuemer
- Aufgewertete Datensätze: GT4HistOCR (23561 Zeilen), AustrianNewspapers (44891 Zeilen), NZZ (68 Zeilen)

Neue Modelle

- GT4HistOCR (2019), verwendeter Datensatz: GT4HistOCR
- Frak2021 (2021), verwendete Datensätze: GT4HistOCR, AustrianNewspapers und Fibeln
- Diverse werksspezifische Modelle

Ausblick

- OCR-BW Nachfolgeprojekt (07/2021–06/2022)
  - Unterstützung der Einrichtungen des Landes
  - Erstellung eines Online-Nachschlagewerkes
  - Validierung von Alternativen zu Transkribus gemeinsam mit der UB Tübingen

- OCR-D Implementierungsprojekte (2021–2023)
  - Werksspezifisches Training von neuen Modellen
  - Implementierung von OCR-D in Kitodo, OCR-Service, OCR-on-demand für DFG-Viewer und Kitodo Presentation

Literatur

- Weil, S. (2019). Training Fraktur. GitHub. https://github.com/tesseract-ocr/tesstrain/wiki
- Weil, S. (2019). Vom Bild zum Text. Automatisierte Texterkennung in historischen Drucken mit der freien Software Tesseract. https://nbn-resolving.org/urn:nbn:de:0290-opus4-163511
- 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