3 Jahre OCR-BW

- Ein Rückblick aus Mannheimer Perspektive


Foto: Valentin Marquardt/Universität Tübingen

Referentin: Larissa Will

E-Mail: larissa.will@uni-mannheim.de

Datum: 22.06.2022

Eine Kooperation von:

Gefördert durch:

Mannheimer Projektteam

- Stefan Weil (Projektleiter): gesamte Projektlaufzeit
- Jan Kamlah (Projektmitarbeiter): 2019-2021
- Larissa Will (Projektmitarbeiterin): 2021-2022

Meilensteine im Projekt

UNIVERSITÄTSBIBLIOTHEK MANNHEIM

1. Softwareentwicklung und -verbesserungen
2. Ground-Truth-Erstellung und -Verbesserung
3. Beratungstätigkeit für Forschende und kulturbewahrende Institutionen
4. Erstellung von Schulungsmaterialien

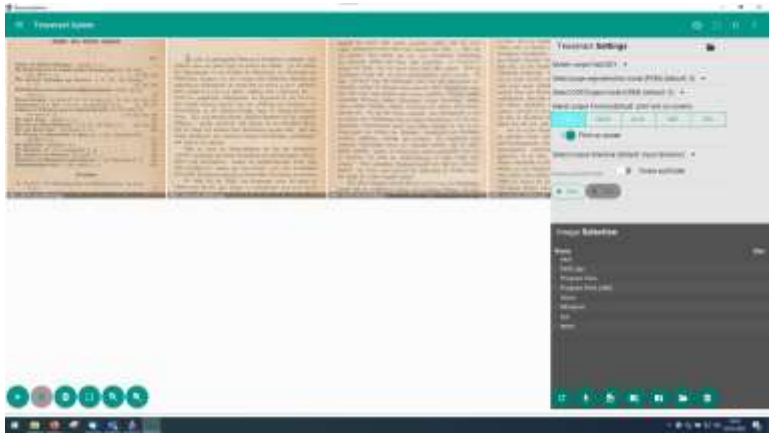
1. Softwareentwicklungen und -verbesserungen

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

- Hilfstools für die Texterkennung
  - BackgroundSubtractor4OCR
  - GTReval
  - GTCheck
  - GTMake
- OCR-D Weiterentwicklungen

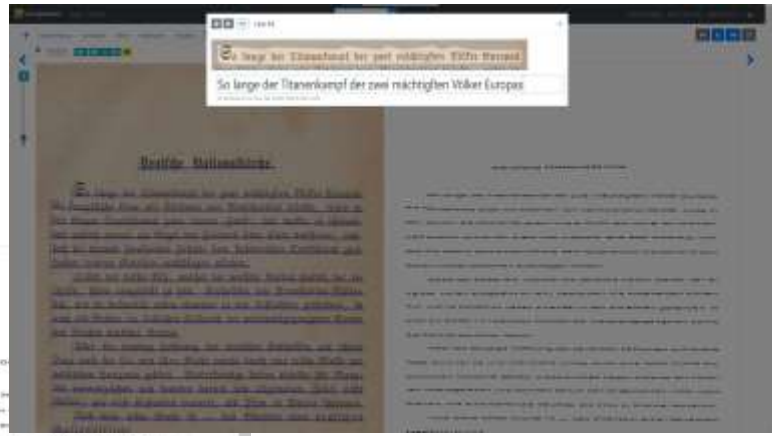
1. Softwareentwicklungen und -verbesserungen

- OCR-Software Tesseract
  - Trainingsoptimierungen
  - TesseractXplore (graphische Oberfläche)



1. Softwareentwicklungen und -verbesserungen

- Transkriptionsplattform eScriptorium
  - Bereitstellung einer Instanz
  - Deutsche Übersetzung für die Benutzeroberfläche
  - Kleinere Korrekturen und Verbesserungen



1. Softwareentwicklungen und – verbesserungen – Ausblick

- Transkriptionsplattform PERO-OCR
  - Austausch mit Entwicklern
  - Erste kleinere Beiträge
  - Bereitstellung einer Instanz

13.8.35
vielen Dank für den
freundl. Brief,
den ich gestern Abend
erhielt. Ich fahre

2. Ground-Truth-Erstellung und -Verbesserung

UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Training von Tesseract-Modellen für historische Drucke, z. B.:
  - Frak2021
  - GT4HistOCR
  - Austrian Newspaper
  - Fraktur 5000000
- Training von Kraken-Modellen für historische Drucke, z. B.:
  - Fraktur 2022-02-20
  - Austrian Newspaper
  - Handschriftenmodell von Edwin Hennig (uat2_56)
- Für Kraken und Tesseract sind noch Modelle für Behördenschriftgut (Schreibmaschine) in Arbeit
- Allgemeines Handschriftenmodell für eScriptorium/Kraken steht noch auf unserer Wunschliste

3. Beratungstätigkeit für Forschende und kulturbewahrende Institutionen

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- BLB Karlsruhe: Unterstützung bei der Implementation einer neuen Texterkennungskomponente
- MARCHIVUM: Texterkennung für Mannheimer Zeitungen
- Friedrich-Ebert-Stiftung: Erzeugung durchsuchbarer Volltexte für Danziger Volksstimme und ihrer Vorgängerzeitung Volkswacht Danzig
- Internet Archive: Unterstützung beim Einsatz von Tesseract
- Unterstützung verschiedener Forschungsvorhaben

4. Erstellung von Schulungsmaterialien

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- eScriptorium-Dokumentation:
  - Anleitung zur lokalen Installation (Windows, MacOS, Linux)
  - Anleitung und Hinweise zur Nutzung
  - Videoanleitung
- Tesseract-Dokumentation (Veröffentlichung folgt):
  - Anleitung zur Installation und Nutzung unter Windows (inkl. TesseractXplore)
  - Anleitung zur Installation und Nutzung unter Linux
- OCRmyPDF-Dokumentation (Veröffentlichung folgt):
  - Anleitung zur Installation und Nutzung unter Windows und Linux (WSL)