



Kompetenzzentrum OCR

Automatische Texterkennung als Serviceangebot

Im Projekt OCR-BW haben die Universitätsbibliotheken Mannheim und Tübingen seit 2019 das „Kompetenzzentrum Volltexterkennung von handschriftlichen und gedruckten Werken“ aufgebaut und beraten seitdem Informationseinrichtungen und wissenschaftliche Projekte in Baden-Württemberg und darüber hinaus zu diesem Thema.

FAIRe Ground-Truth-Daten

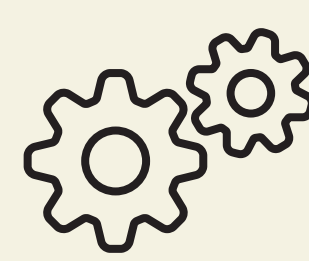
Die Erzeugung von Ground-Truth-Daten für Segmentierungs- und Texterkennungsmodelle ist aufwändig und kostenintensiv. Es ist daher für eine nachhaltige und nachvollziehbare Wissenschaft wünschenswert, wenn die Daten gemäß den FAIR-Prinzipien bereitgestellt werden.



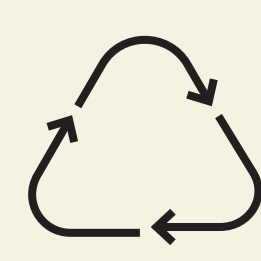
Findable



Accessible



Interoperable



Reusable

Dies fördert Transparenz, Chancengleichheit und ethisch verantwortliche KI-Entwicklung. Durch die Anwendung dieser Prinzipien auf die Daten wird der Grundstein für vertrauenswürdige und hochwertige Modelle gelegt.

Das Serviceangebot des Kompetenzzentrums erstreckt sich von der Datenauswahl über die Erzeugung von Trainingsmaterial, dem anschließenden Texterkennungsprozess bis hin zur Nachnutzung der Projektergebnisse.

Datenauswahl

Um spätere Probleme zu vermeiden, müssen bei der Auswahl der Werke technische sowie rechtliche Aspekte berücksichtigt werden:

- Rechtliche Aspekte: Urheberrecht, Lizenzen, Datenschutz
- Beurteilung der Datenqualität: Drucktechnische Herausforderungen, Quellenmaterial
- Berücksichtigung von Digitalisierungsempfehlungen: Farbraum, Auflösung, Ausrichtung, Dateiformat

Erzeugung von Ground-Truth-Daten

Die Erstellung von GT-Daten hat großen Einfluss auf die Qualität der Texterkennung. Es ist nicht nur wichtig, was transkribiert wird, sondern auch wie es transkribiert wird:

- Standardisierung der GT-Daten: Auswahl einer passenden Transkriptionsrichtlinie nach OCR-D
- Anpassung der Richtlinie an spezifischen Anwendungsfall
- Erstellung einer Dokumentation zur Nachvollziehbarkeit

Veröffentlichung und Nachnutzung

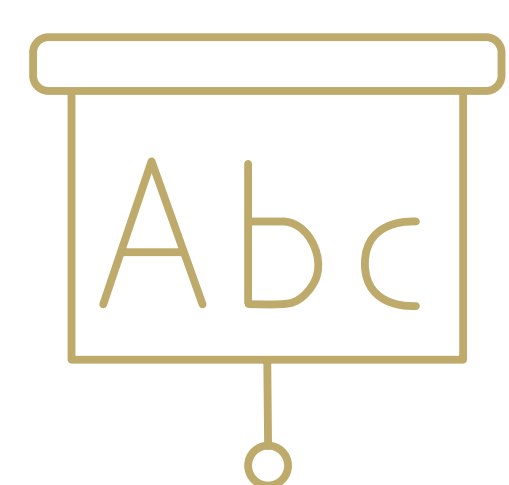
Bei der Veröffentlichung sollte die Auffindbarkeit, Zitierbarkeit und langfristige Verfügbarkeit der Daten sichergestellt werden:

- Bereitstellung von GT, Texterkennungsergebnissen und Modellen z. B. über Zenodo
- Anreicherung mit Metadaten
- Lizenz vergeben (möglichst offen)
- Wahl eines Dateiformats (Empfehlung: nicht-proprietäres Format)
- Achtung von Urheber- und Persönlichkeitsrechten



Beratung

- Beratung und Unterstützung bei Fragen zur automatisierten Texterkennung
- offene Sprechstunde via Zoom, jeden zweiten Donnerstag im Monat



Schulung

- Durchführung von Einführungskursen in eScriptorium und Transkribus
- Schulungsmaterialien über ZOERR und GitHub



Unterstützung

- Hinweise zur Digitalisierung
- Bearbeitung von Testseiten
- Empfehlung von Tools und Modellen
- Texterkennung im kleineren Umfang
- Bereitstellung einer eScriptorium-Instanz durch die UB Mannheim
- für kleinere Projekte Bereitstellung von Ressourcen

Unser Service



Wenn Sie Fragen rund um das Thema Texterkennung haben, sprechen Sie uns gerne an!

Weitere Informationen finden Sie auf unserer Website: <https://ocr-bw.bib.uni-mannheim.de/>.