OCR-D Implementierungsprojekt

Integration von Kitodo und OCR-D zur produktiven Massendigitalisierung (2021–2023/24)

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

Stand OCR-D-Kitodo-Implementierung

Stefan Weil, Universitätsbibliothek Mannheim
7. November 2023

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Status Gesamtprojekt

OCR-D + Kitodo.Production

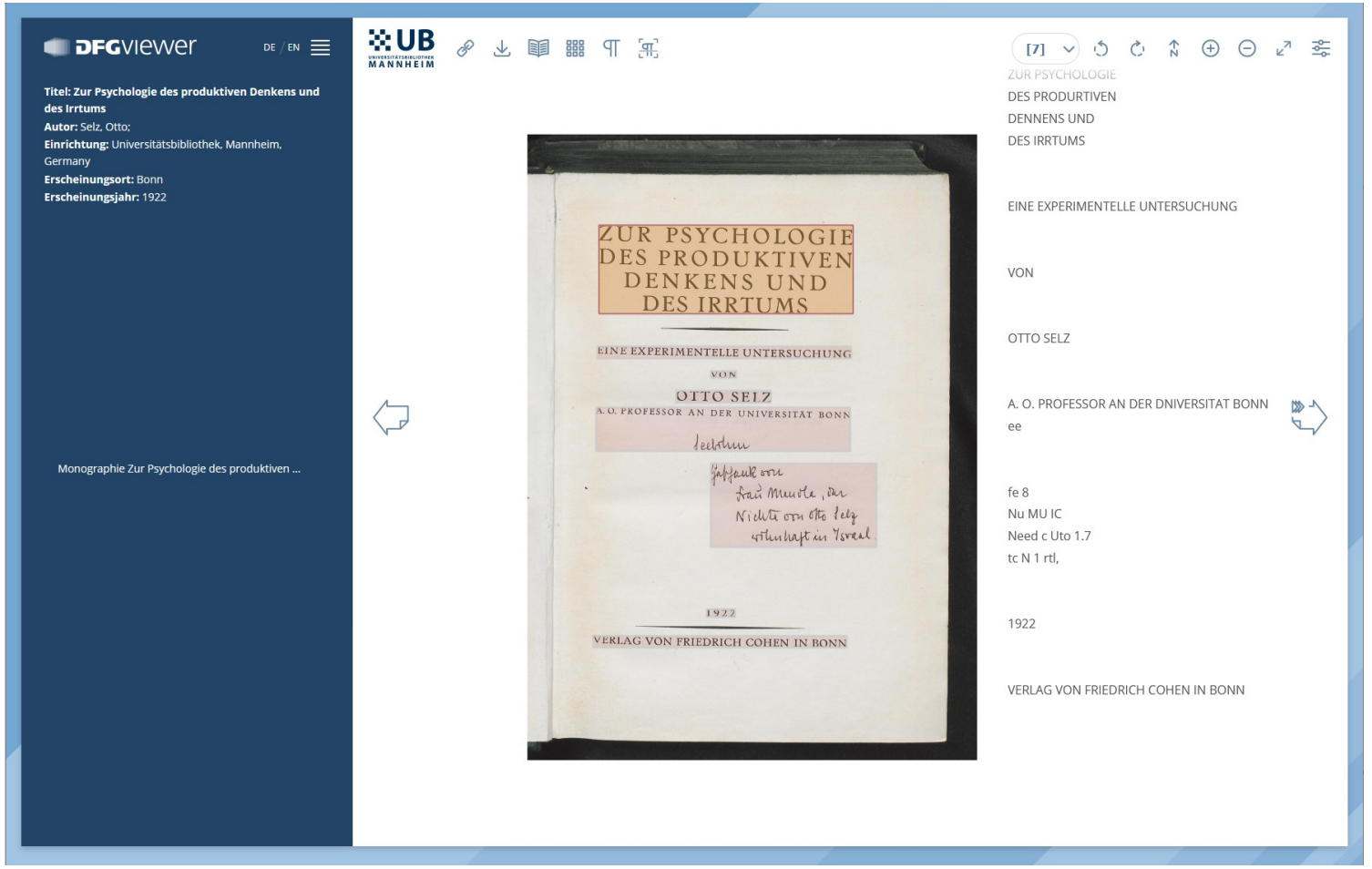
- UB Braunschweig: Projektabschluss war 09/2023 mit dem Workshop „Integration von OCR-D in Kitodo: vom Digitalisat bis zur Veröffentlichung der Texte“
- SLUB Dresden: Projektende 03/2024

OCR-D + Kitodo.Presentation

- UB Mannheim: Projektende 01/2024

Status OCR-D + Kitodo.Presentation

- Referenzimplementierung mit DFG-Viewer und OCR-on-Demand läuft seit über einem Jahr
- Dauer (abhängig von der ausgewählten OCR Technik und dem Seitenaufbau):
  - 1 s bis 14 s pro Seite
  - 39 min für 694 Seiten (parallelisierbar)



Aktuell angebotene Funktionalität

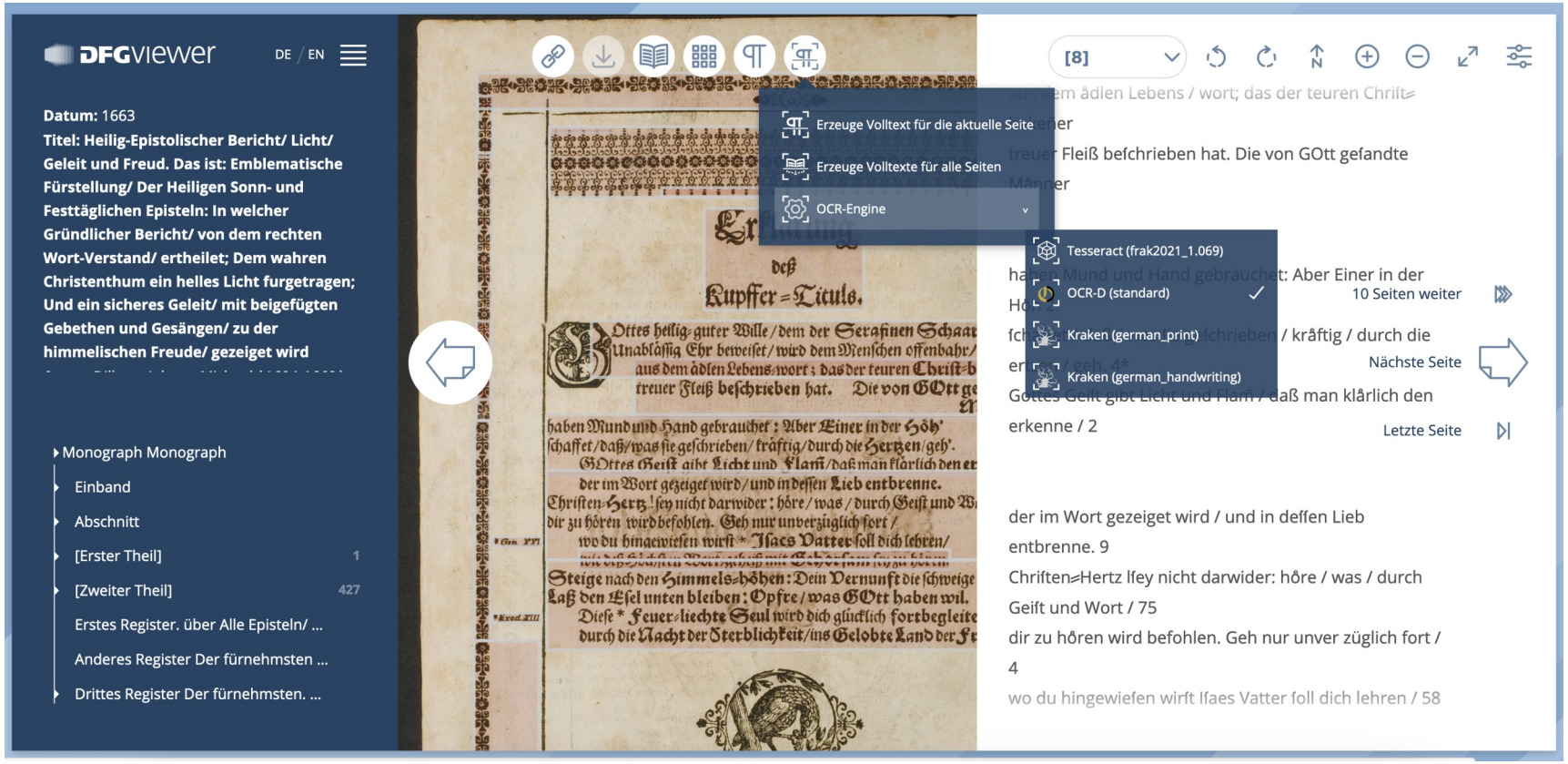
OCR für einzelne oder alle Seiten mit vier unterschiedlichen OCR-Prozessen:

- Tesseract mit dem Modell frak2021_1.069 der UB Mannheim
- OCR-D mit Tesseract-Prozessor und dem Modell frak2021_1.069
- Kraken mit dem Modell german_print der UB Mannheim
- Kraken mit dem Modell german_handwriting der UB Mannheim

Beispiel: Druck 1693, Tesseract / frak2021_1.069

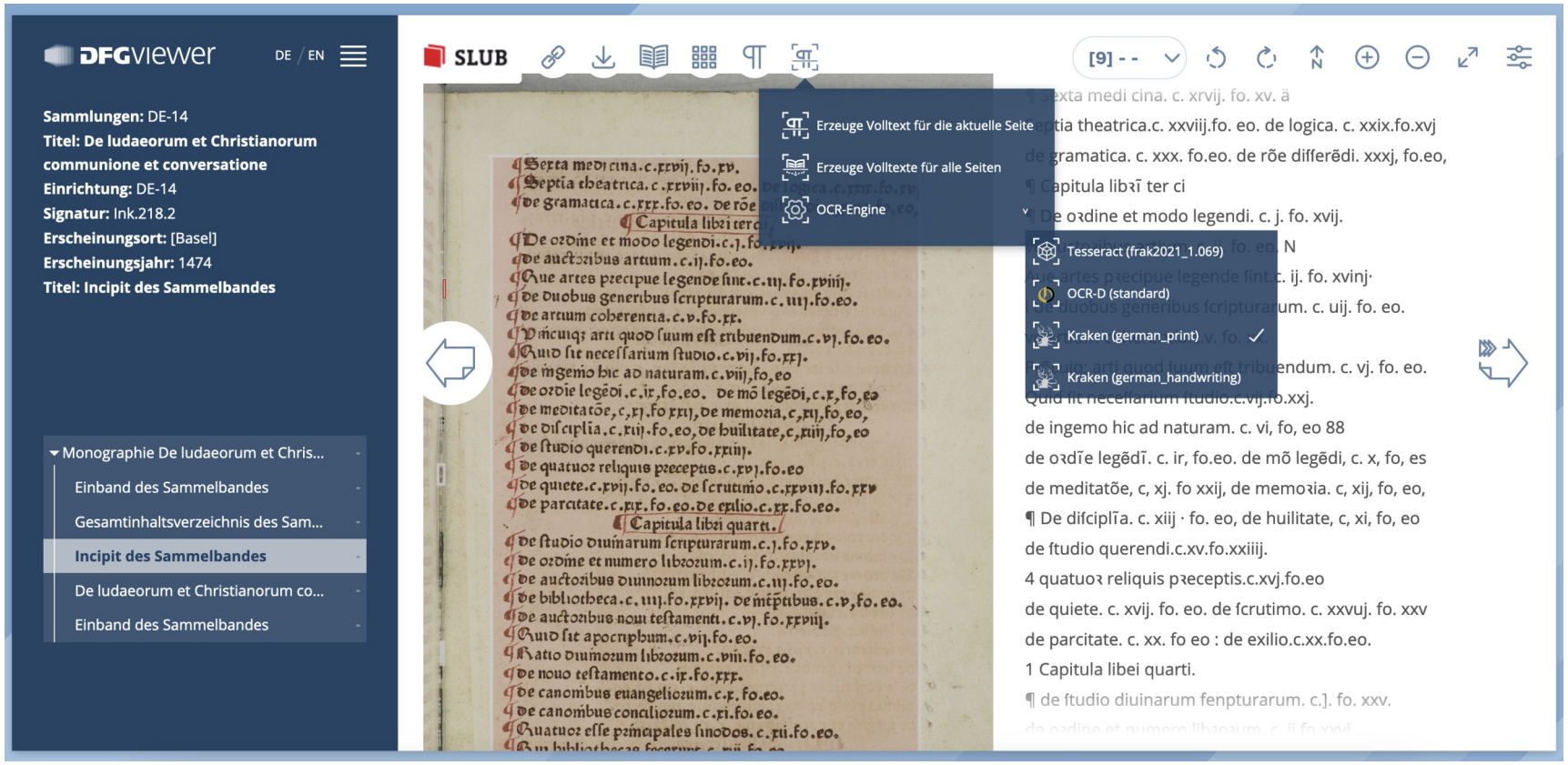
- URL: https://digitale.bibliothek.uni-halle.de/vd17/oai/?verb=GetRecord&metadataPrefix=mets&identifier=94835

Beispiel: Druck 1663, OCR-D / frak2021_1.069



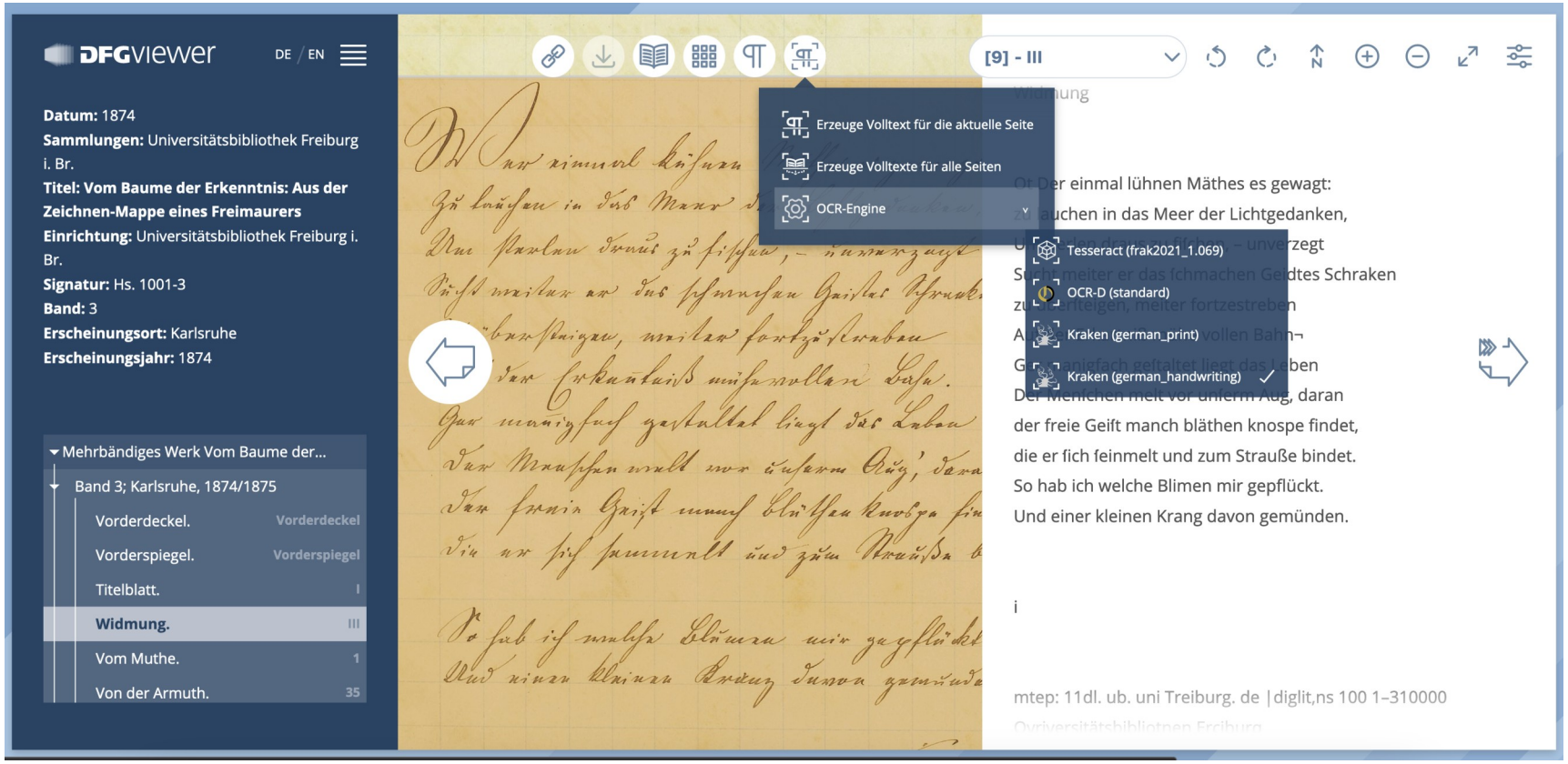
- URL: http://oai.hab.de/?verb=GetRecord&metadataPrefix=mets&identifier=oai:diglib.hab.de:ppn_549837965

Beispiel: Druck 1474, Kraken / german_print



- URL: https://digital.slub-dresden.de/oai/?verb=GetRecord&metadataPrefix=mets&identifier=oai:de:slub-dresden:db:id-312439970

Beispiel: HS 1874, Kraken / german_handwriting



- URL: http://dl.ub.uni-freiburg.de/diglit/hs1001-3/mets

Technische Herausforderung: TYPO3 Versionen

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Wunsch: Implementierung auf Basis einer aktuellen Version von TYPO3
- Frei nutzbare Versionen werden nicht mehr unterstützt
  TYPO3 8: bis 2020-03-31
  TYPO3 9: bis 2021-09-30
  TYPO3 10: bis 2023-04-30
- Kitodo.Presentation mit TYPO3 11 (bis 2024-10-31) hat in jüngster Zeit große Fortschritte gemacht, bedarf aber noch weiterer Anstrengungen
- Aktuelle Version TYPO3 12 (seit 2023-10-10, bis 2026-04-30) muss noch mit Kitodo.Presentation evaluiert werden
- Konkrete Implementierungen (z. B. DFG-Viewer) aktualisieren

Nutzungsstatistik 2024/10/23 bis 2024/11/06

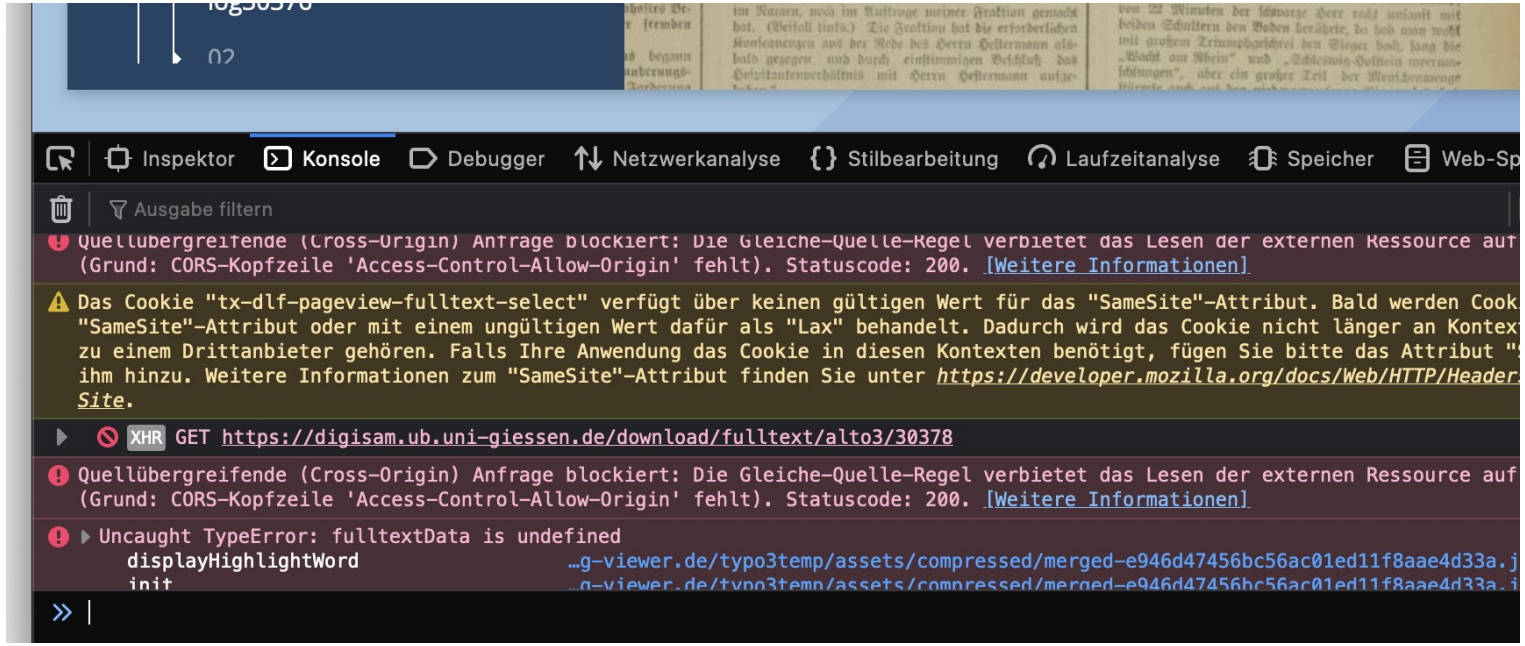
- 40 Aufrufe der URL https://dfg-viewer.bib.uni-mannheim.de/ (ohne Bots und UB Mannheim)
- 30 unterschiedliche IP-Adressen, darunter keine wissenschaftliche Einrichtungen
- => 2 bis 3 Nutzungen täglich

Ist das Angebot trotz mehrerer Vorträge zu wenig bekannt?

Oder erfüllt die momentane Implementierung noch nicht die Bedarfe und Ansprüche der Nutzenden?

Nutzung DFG-Viewer nicht immer einfach

- Vielfach muss man die benötigte METS-URL suchen, manchmal findet man sie gar nicht
- Relativ oft fehlt die Konfiguration für Cross-Origin Resource Sharing (CORS) => Volltexte der Einrichtung werden im DFG-Viewer nicht angezeigt



Wunschliste / Ideensammlung

- Weitere Optimierung der Webseiten für verbesserte Bedienbarkeit
- Maschinelle Übersetzung der Volltexte
  Open Source: MarianNMT (https://marian-nmt.github.io/), z. B. in Firefox verwendet
- Text to speech – Audioausgabe der Volltexte (Barrierefreiheit)
  diverse Open Source Projekte kommen dafür in Frage
- Feedback-Möglichkeit zu den OCR-Resultaten
- Korrigierbare Volltexte (auch für Nachtraining verwendbar)
- Versionierung der maschinell erzeugten oder korrigierten Volltexte
  Git-Repositories mit Revisionen, Branches und Tags, die für die Anzeige auswählbar sind

Ausblick

- Die bisher besten an der UB Mannheim trainierten generischen Modelle für Tesseract – GT4HistOCR und frak2021 – werden in Kürze durch neu trainierte Modelle german_print ergänzt.
- Das Training für german_print verwendet aktuell mehr als 550.000 Textzeilen Ground Truth aus diversen GT-Sammlungen mit Texten vom 15. bis 20. Jahrhundert.
- Beispiel für Zwischenergebnis alt/neu:
  -im Centralbureau ein Umſatz von 390,542, & erzielt iſt gegen ca.
  -300,000 & in demſelben Zeitraum des Vorjahres. Am ſchwächſten
  +im Centralbureau ein Umſatz von 390,542,88 ℳ erzielt iſt gegen ca.
  +300,000 ℳ in demſelben Zeitraum des Vorjahres. Am ſchwächſten
  (https://upload.wikimedia.org/wikipedia/commons/f/fc/Reichsanzeiger-1876-10-26_Hausfrauenverein.png)
- Details: https://github.com/UB-Mannheim/tesstrain/wiki/.

Weitere Informationen

- Referenzimplementierung:
  → *https://dfg-viewer.bib.uni-mannheim.de/*
- Installationsanleitung:
  → *https://github.com/UB-Mannheim/kitodo-presentation/wiki*
- Projektplan:
  → *https://github.com/orgs/UB-Mannheim/projects/2*