



OCR-D Implementierungsprojekt

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



Stand OCR-D-Kitodo-Implementierung



Status Gesamtprojekt

OCR-D + Kitodo.Production

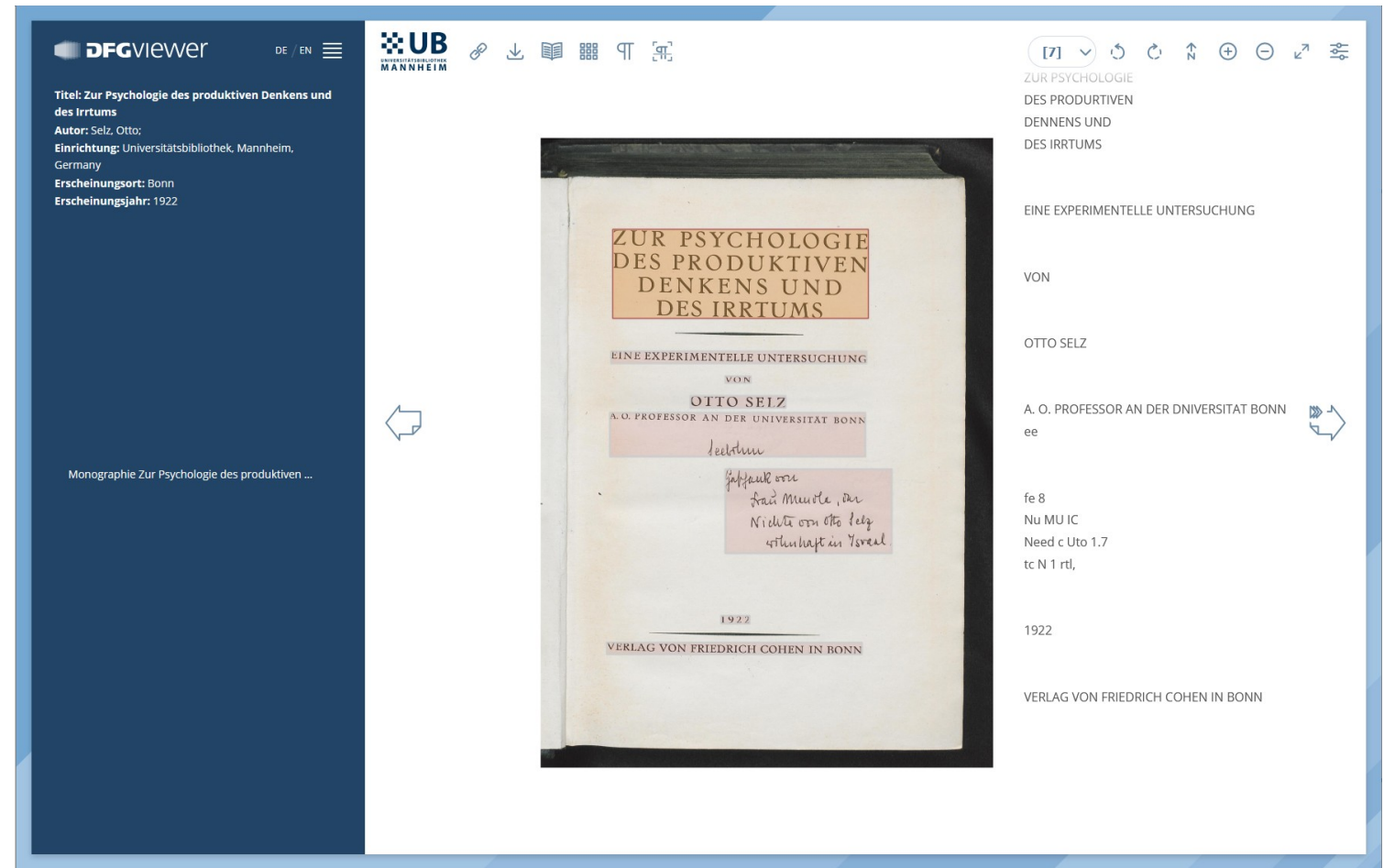
- UB Braunschweig: Projektabschluss war 09/2023 mit dem Workshop „Integration von OCR-D in Kitodo: vom Digitalisat bis zur Veröffentlichung der Texte“
- SLUB Dresden: Projektende 03/2024

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)



The screenshot displays the DFGviewer interface for a scanned book cover. The central image shows the title page of 'ZUR PSYCHOLOGIE DES PRODUKTIVEN DENKENS UND DES IRRTUMS' by Otto Selz, published in 1922. The interface includes a dark blue sidebar on the left with metadata and a right sidebar with a hierarchical table of contents. The table of contents lists the title, author, and publisher information.

Left Sidebar (Metadata):

- DFGviewer DE / EN
- UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM
- Titel: Zur Psychologie des produktiven Denkens und des Irrtums
- Autor: Selz, Otto;
- Einrichtung: Universitätsbibliothek, Mannheim, Germany
- Erscheinungsort: Bonn
- Erscheinungsjahr: 1922
- Monographie Zur Psychologie des produktiven ...

Right Sidebar (Table of Contents):

- [7]
- ZUR PSYCHOLOGIE
- DES PRODUKTIVEN
- DENNENS UND
- DES IRRTUMS
- EINE EXPERIMENTELLE UNTERSUCHUNG
- VON
- OTTO SELZ
- A. O. PROFESSOR AN DER UNIVERSITÄT BONN
- ee
- fe 8
- Nu MU IC
- Need c Uto 1.7
- tc N 1 rtl,
- 1922
- VERLAG VON FRIEDRICH COHEN IN BONN

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

DFGviewer DE / EN

ULB Sachsen-Anhalt

Datum: 1693
Sammlungen: 65
Titel: Das bißher weit und breit herumgegangene/ auch offft gedruckt und abgeschriebene Fontangen-Lied
Einrichtung: 65
Erscheinungsort: [S.I.]
Erscheinungsjahr: 1693

▼ Monographie Das bißher weit und br...
Titelblatt
Vorwort
Abschnitt 8

Erzeuge Volltext für die aktuelle Seite
Erzeuge Volltexte für alle Seiten
OCR-Engine

Tesseract (frak2021_1.069) ✓
OCR-D (standard)
Kraken (german_print)
Kraken (german_handwriting)

Hochgeehrter Christlicher Leser.
Nachdem vorher stehendes Fontangen-Lied
nah in die Hände und zu Gesicht gekommen
ist/wie dasselbe von vielen beltebet / und offft
ben worden/dahen aber auch vermercket/di
elicher Noetlichen Mod:Arten und Traumen etlichen
lichen Orten / unverständlich gewesen/als habe ich ged
seyn/wenn nicht allein das ganze Lied fein richtig ged
dern auch dasselbe mit etlichen kurzen Anmerkunge
würde/derguten Hoffnung zu GOtt/daß vielleicht
wol die Fontangen-Träger desto mehr mögen besch
Abscheu davor beweget/als auch die bißher dran keine
vielmehr Eckel und Greuel gehabt/in ihrem lobwür
ligem Effer möchten befärcket werden. Wer der
Liedes sey/ist mir unbewußt; es ist aber leicht abzun
müsse ein GOtt- und Tugendliebender Mensch seyn

Okten / unverständlich gewefen / als habe ch
ht / gut zu
feyn / wenn nicht allein das ganze Lled fein richtig
gedruckt / fon
auch dasselbe mit etlichen kurtzen
Tesseract (frak2021_1.069) ✓
Wol derguten Hoffnung zu GOtt / daß vielleicht
hi
Wol die Fontangen-Träger desto mehr mögen
be
Abnehen davor beweget/als auch die biß her dran
keinen Gefallen /
vielmehr Eckel und Greuel gebabt / in ihrem
lobwürdigen / ga hei
ligem Effer mochten befärcket werden. Wer der Autor
diefes
Lledes fey / ist mir unbewuft; es ist aber leicht
abzunehmen / es
müsse ein GOtt- und Tugendliebender Mensch feyn /
einer aus
der Anzahl derer jenigen / die da / wenn sie ja bey
Affentlicher Laffer

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

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

The screenshot shows the DFGviewer interface. On the left, there is a dark sidebar with the following text:

DFGviewer DE / EN

Datum: 1663
 Titel: Heilig-Epistolischer Bericht/ Licht/ Geleit und Freud. Das ist: Emblematische Fürstellung/ Der Heiligen Sonn- und Festtäglichen Episteln: In welcher Gründlicher Bericht/ von dem rechten Wort-Verstand/ ertheilet; Dem wahren Christenthum ein helles Licht furgetragen; Und ein sicheres Geleit/ mit beigefügten Gebethen und Gesängen/ zu der himmelischen Freude/ gezeigt wird

Below the sidebar is a table of contents:

Monograph Monograph	
Einband	
Abschnitt	
[Erster Theil]	1
[Zweiter Theil]	427
Erstes Register. über Alle Episteln/ ...	
Anderes Register Der fürnehmsten ...	
Drittes Register Der fürnehmsten. ...	

The main area shows a manuscript page with a large initial 'S' and text in German. A dropdown menu is open over the page, showing options: 'Erzeuge Volltext für die aktuelle Seite', 'Erzeuge Volltexte für alle Seiten', and 'OCR-Engine'. The 'OCR-Engine' dropdown is further open, showing a list of engines:

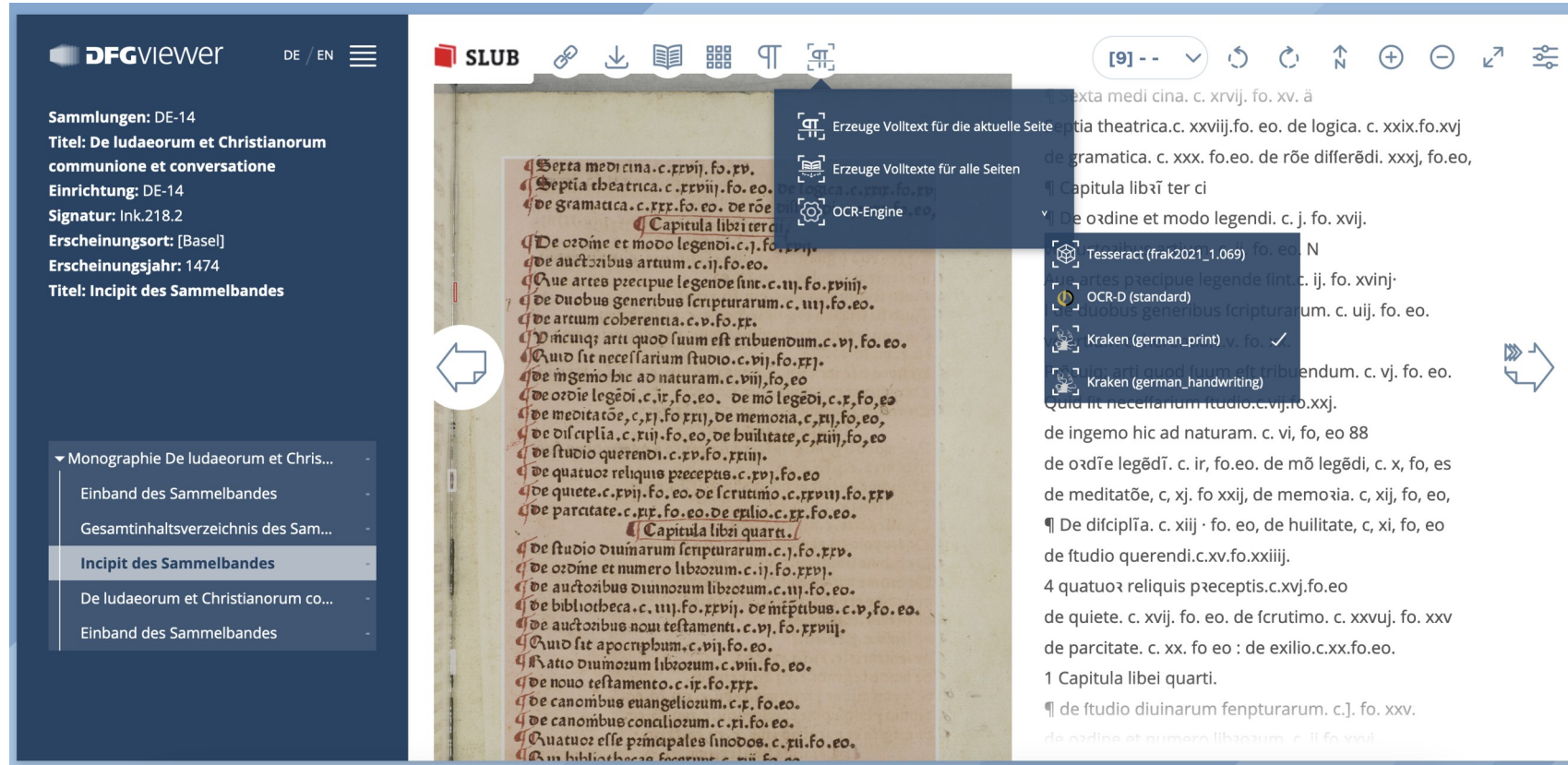
Tesseract (frak2021_1.069)	
OCR-D (standard)	10 Seiten weiter
Kraken (german_print)	Nächste Seite
Kraken (german_handwriting)	Letzte Seite

On the right side of the interface, there is a list of OCR results for the current page (8):

haben Mund und Hand gebraucher; Aber Einer in der Schafft/das/was sie geschrieben/ kräftig/durch die Hergen/ geh'.
 Gottes Geist aibe Lichte und Flam/ das man klärtlich den ee der im Wort gezeigt wird/ und in dessen Lieb entbrenne.
 Christen-Hertz! sey nicht darwider: höre / was / durch Geist und W dir zu hören wird befohlen. Geh nur unverzüglich fort / wo du hingewiesen wirft * Ilaes Vatter soll dich lehren/
 Steige nach den Himmels-höhen: Dein Vernunft die schweige Laß den Lsel unten bleiben: Opfre / was GOTT haben wil. Diese * Feuer-lichte Seul wird dich glücklich fortbegleite durch die Nacht der Sterblichkeit/ ins Gelobte Land der Fe

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

Beispiel: Druck 1474, Kraken / german_print



The screenshot shows the DFGviewer interface for a manuscript. On the left, there is a sidebar with metadata for 'Sammlungen: DE-14' and 'Titel: De Iudaeorum et Christianorum communione et conversatione'. The main area displays a manuscript page with Latin text in Gothic script. A dark blue overlay is positioned over the text, containing a list of OCR options:

- Erzeuge Volltext für die aktuelle Seite
- Erzeuge Volltexte für alle Seiten
- OCR-Engine
- Tesseract (frak2021_1.069)
- OCR-D (standard)
- Kraken (german_print) v. fo. ✓
- Kraken (german_handwriting)

The 'Kraken (german_print)' option is selected. The background text includes sections like 'De ordine et modo legendi' and 'Capitula libri tertii'.

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

Beispiel: HS 1874, Kraken / german_handwriting

The screenshot shows the DFGviewer interface for a handwritten manuscript. On the left, a dark sidebar contains metadata: **Datum:** 1874, **Sammlungen:** Universitätsbibliothek Freiburg i. Br., **Titel:** Vom Baume der Erkenntnis: Aus der Zeichen-Mappe eines Freimaurers, **Einrichtung:** Universitätsbibliothek Freiburg i. Br., **Signatur:** Hs. 1001-3, **Band:** 3, **Erscheinungsort:** Karlsruhe, **Erscheinungsjahr:** 1874. Below this is a table of contents for 'Band 3; Karlsruhe, 1874/1875' with 'Widmung.' selected as page III. The main area shows a handwritten manuscript page with a digital overlay. A dropdown menu is open over the manuscript, offering options: 'Erzeuge Volltext für die aktuelle Seite', 'Erzeuge Volltexte für alle Seiten', and 'OCR-Engine'. The OCR-Engine dropdown is further expanded, showing 'Tesseract (frak2021_1.069) - Unverzegt', 'OCR-D (standard)', 'Kraken (german_print) vollen Bah...', and 'Kraken (german_handwriting)' which is checked. The manuscript text is in German, starting with 'Wo immer köfzen zu köfzen in des Mann...' and 'Der einmal lühen Mäthes es gewagt: Suchen in das Meer der Lichtgedanken, ...'. At the bottom right, there is a URL: 'mtep: 11dl. ub. uni Treiburg. de | diglit,ns 100 1-310000' and 'Universitätsbibliothek Freiburg'.

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

Technische Herausforderung: TYPO3 Versionen

- 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

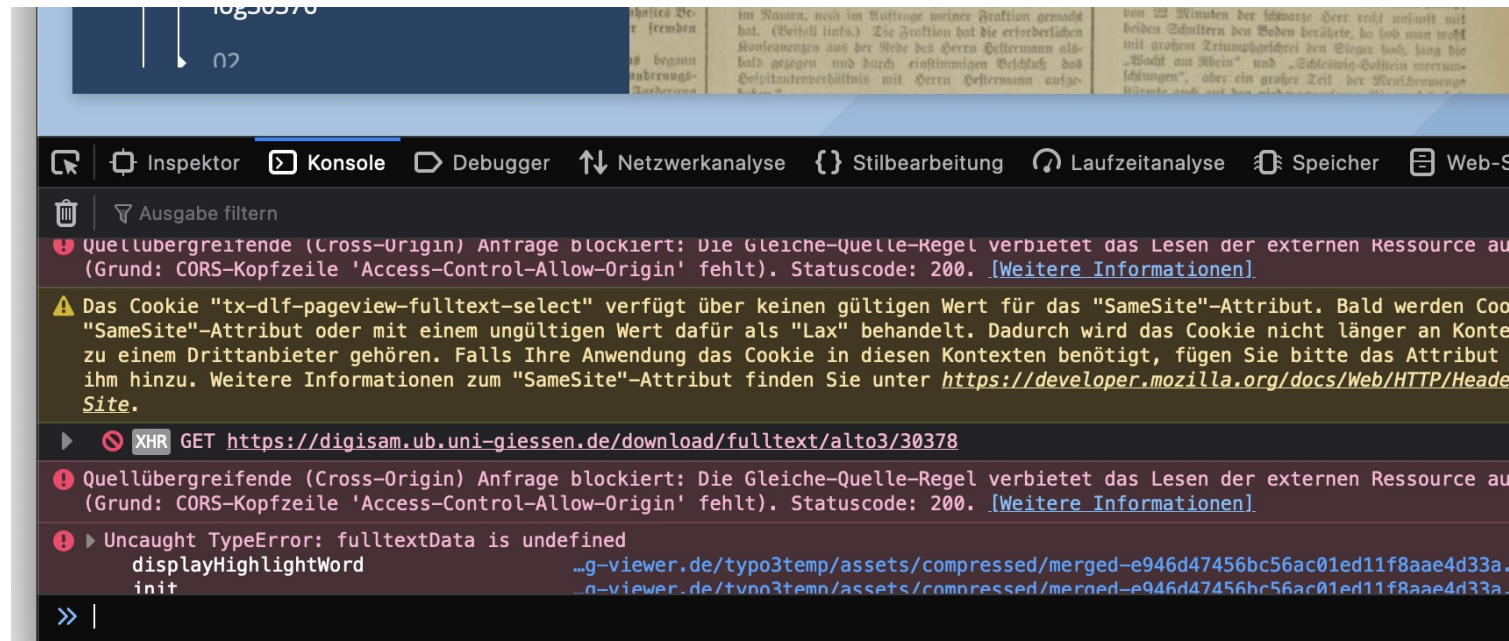
- 40 Aufrufe der URL <https://dfg-viewer.bib.uni-mannheim.de/> (ohne Bots und UB Mannheim)
- 30 unterschiedliche IP-Adressen, darunter keine wissenschaftliche Einrichtungen
- => 2 bis 3 Nutzungen täglich

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 ift gegen ca. -300,000 & in demselben Zeitraum des Vorjahres. Am fchwächften
 - +im Centralbureau ein Umsatz von 390,542,88 *M* erzielt ift gegen ca. +300,000 *M* in demselben Zeitraum des Vorjahres. Am fchwächften(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>