



# 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. The OCR data is overlaid on the image, identifying the title, author, and publisher. A sidebar on the left provides detailed metadata, and a sidebar on the right lists the document's classification and call numbers.

**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 ...

**ZUR PSYCHOLOGIE  
DES PRODUKTIVEN  
DENKENS UND  
DES IRRTUMS**

EINE EXPERIMENTELLE UNTERSUCHUNG  
VON  
**OTTO SELZ**  
A.O. PROFESSOR AN DER UNIVERSITÄT BONN

*Lehrbuch*

*gefördert von  
der Musee, der  
Nichte von Otto Selz  
erhalten in Israel*

1922  
VERLAG VON FRIEDRICH COHEN IN BONN

[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

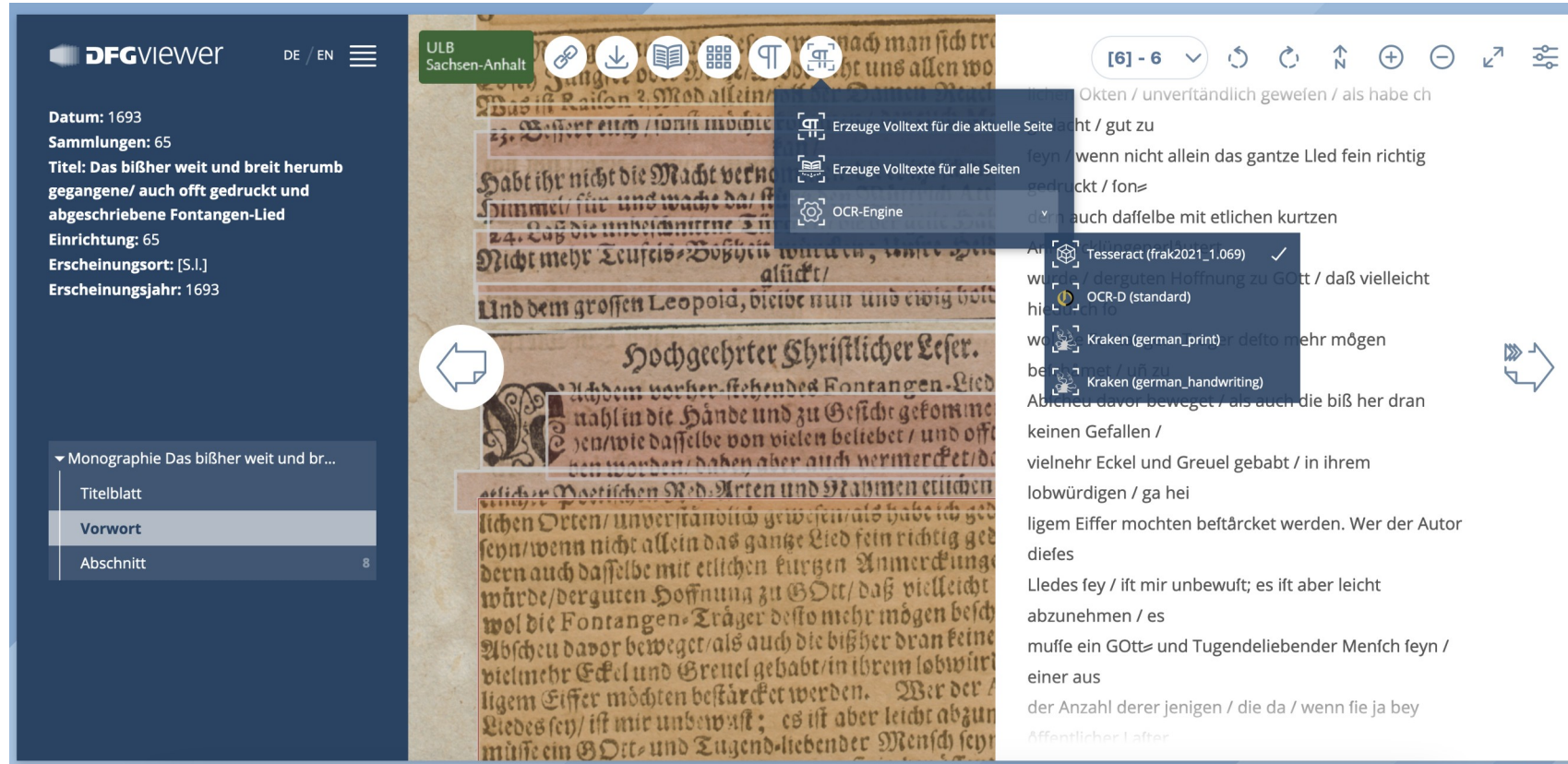
VERLAG VON FRIEDRICH COHEN IN BONN

# 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  
wie daselbe von vielen beltebet / und offft  
ben worden / dahin aber auch vermercket / da  
elicher Noetlichen Mod. Arten und Traumen etlichen  
lichen Orten / unverständlich gewesen / als habe ich ge  
seyn / wenn nicht allein das ganze Lied fein richtig ged  
dern auch daselbe mit etlichen kurzen Anmerkunge  
würde / der guten 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 Eiffer möchten befärcket werden. Wer der  
Liedes sey / ist mir unbewußt ; es ist aber leicht abzun  
müsse ein Gott. und Tugendliebender Mensch seyn

Okten / unverständlich gewesen / als habe ch  
ht / gut zu  
feyn / wenn nicht allein das ganze Lied fein richtig  
gedruckt / fon  
auch daselbe mit etlichen kurzen  
Tesseract (frak2021\_1.069) ✓  
w... der guten Hoffnung zu Gott / daß vielleicht  
hi...  
w... Kraken (german\_print) r desto mehr mögen  
be...  
Kraken (german\_handwriting)  
Abneid davor beweget / als auch die biß her dran  
keinen Gefallen /  
vielmehr Eckel und Greuel gebabt / in ihrem  
lobwürdigen / ga hei  
ligem Eiffer mochten befärcket werden. Wer der Autor  
diefes  
Lledes fey / ist mir unbewußt ; 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 Leser

- 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 sidebar with the title and a table of contents. The main area displays a manuscript page with a large initial 'S' and text in German. A context menu is open over the text, showing options like 'Erzeuge Volltext für die aktuelle Seite' and 'OCR-Engine'. A dropdown menu for 'OCR-Engine' is also visible, listing 'Tesseract (frak2021\_1.069)' and 'OCR-D (standard)' as selected options. On the right, a list of OCR results is shown, including page numbers and navigation arrows.

**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

Monograph Monograph

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

Erzeuge Volltext für die aktuelle Seite

Erzeuge Volltexte für alle Seiten

OCR-Engine

Tesseract (frak2021\_1.069)

OCR-D (standard) ✓ 10 Seiten weiter

Kraken (german\_print) geschrieben / kräftig / durch die Nächste Seite

Kraken (german\_handwriting)

erkenne / 2 Letzte Seite

haben Mund und Hand gebraucher; Aber Einer in der Hörschafft/das/was sie geschrieben / kräftig / durch die Herzen / geh'. Gottes Geist aibe Lichte und Flam / das man klärtlich den er 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 hinaewiesen wirft \* Ihaes Vatter soll dich lehren/ Steige nach den Himmels-höhen: Dein Vernunft die schweige Laß den Esel 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

der im Wort gezeigt wird / und in deffen Lieb entbrenne. 9

Christen-Hertz lsey nicht darwider: höre / was / durch Geist und Wort / 75

dir zu hören wird befohlen. Geh nur unver züglich fort / 4

wo du hingewiefen wirft Ihaes Vatter foll dich lehren / 58

- URL: [http://oai.hab.de/?verb=GetRecord&metadataPrefix=mets&identifizier=oai:diglib.hab.de:ppn\\_549837965](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, a dark sidebar contains 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 semi-transparent overlay with a list of OCR options is positioned over the text. The options include '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. ✓', and 'Kraken (german\_handwriting)'. The 'Kraken (german\_print)' option is selected with a checkmark. Navigation icons for zooming and panning are visible on the right side of the viewer.

- 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

**DFGviewer** DE / EN

**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

**Mehrbändiges Werk Vom Baume der...**

Band 3; Karlsruhe, 1874/1875	
Vorderdeckel.	Vorderdeckel
Vorderspiegel.	Vorderspiegel
Titelblatt.	I
<b>Widmung.</b>	<b>III</b>
Vom Muthe.	1
Von der Armuth.	35

**[9] - III**

- Erzeuge Volltext für die aktuelle Seite
- Erzeuge Volltexte für alle Seiten
- OCR-Engine
- Tesseract (frak2021\_1.069) - Unverzegt
- OCR-D (standard)
- Kraken (german\_print) vollen Bah...
- Kraken (german\_handwriting) ✓

metep: 11dl. ub. uni Treiburg. de | diglit,ns 100 1-310000  
 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

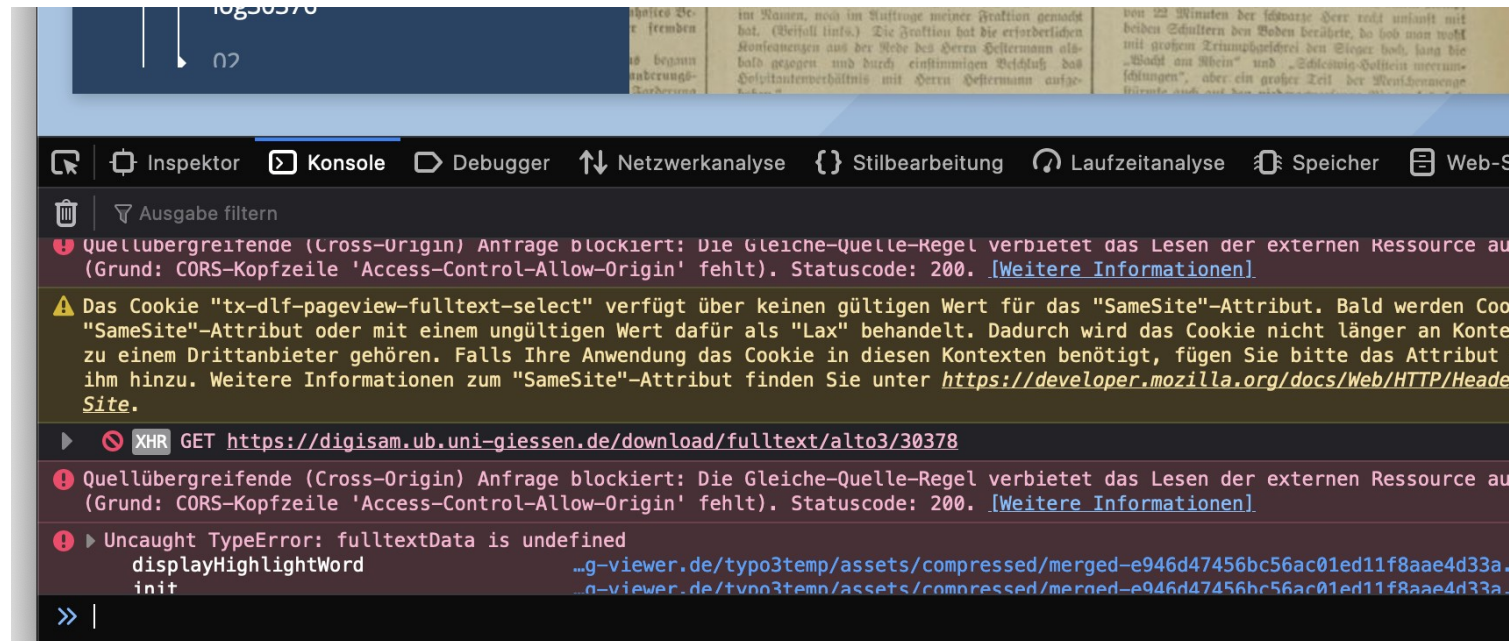
- 40 Aufrufe der URL <https://dfg-viewer.bib.uni-mannheim.de/> (ohne Bots und UB Mannheim)
- 30 unterschiedliche IP-Adressen, darunter keine wissenschaftliche Einrichtungen
- => 2 bis 3 Nutzungen täglich

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](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>