

OCR-D Implementierungsprojekt

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

OCR On-Demand für DFG-Viewer und Kitodo.Presentation

Einführung

dfg-viewer.de – Referenzimplementierung auf der Grundlage von Kitodo.Presentation

Homepage DFG Viewer

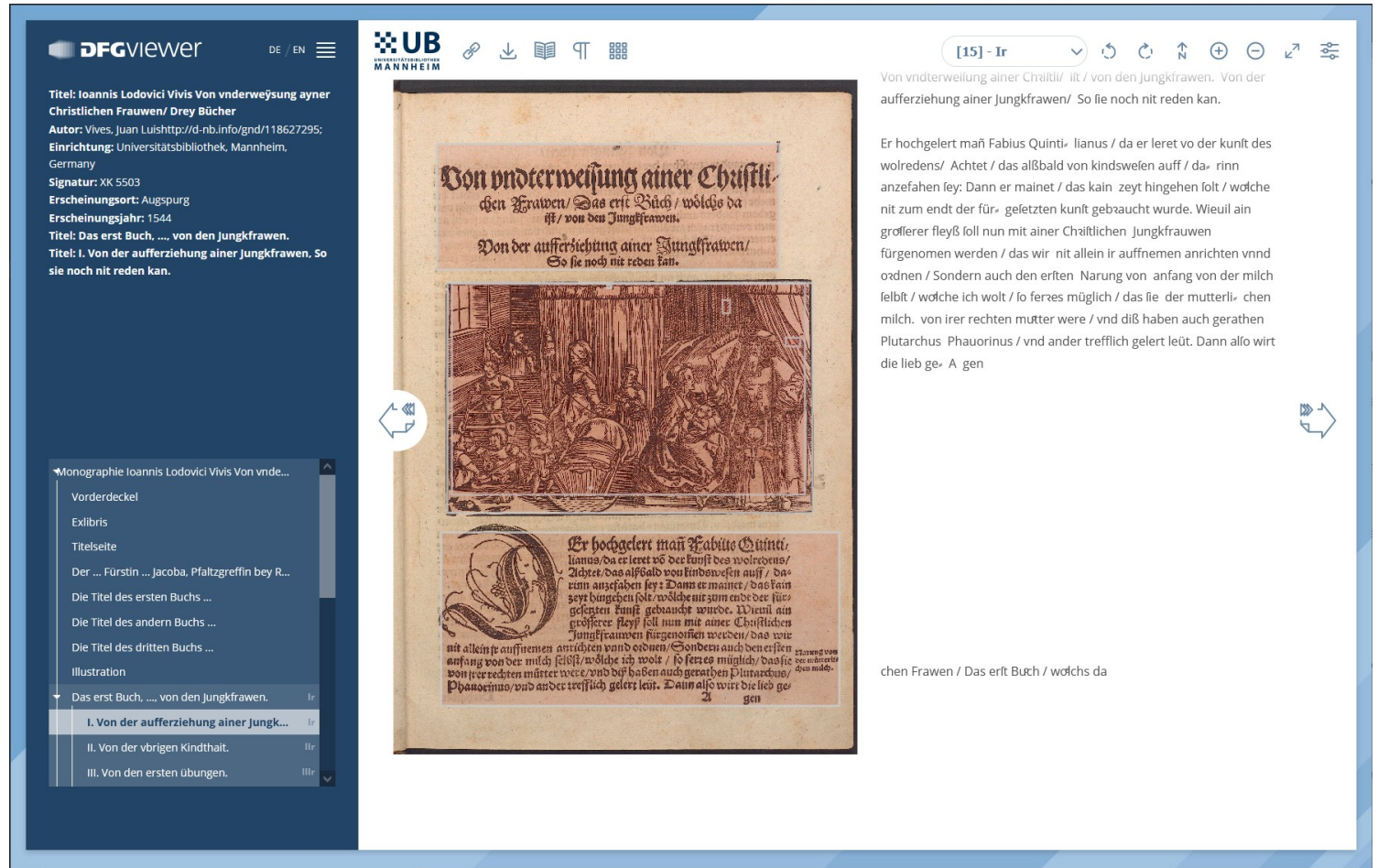
- Entwicklung und Betrieb durch die SLUB Dresden
- Förderung durch die Deutsche Forschungsgemeinschaft (DFG)



DFG-Viewer: Beispiel mit Volltext

Der DFG-Viewer

- ... ist in der Lage, die (meisten) Digitalisate aus Archiven und Bibliotheken darzustellen.



The screenshot displays the DFG-Viewer interface. On the left, a dark sidebar contains a table of contents for the manuscript. The main area shows a digital scan of a manuscript page with a title, a woodcut illustration, and a block of text. On the right, a full-text view of the selected page is shown, including a search bar and navigation icons.

DFGviewer DE / EN

Titel: Ioannis Lodovici Vivis Von vnderweysung ayner Christlichen Frauen/ Drey Bücher
Autor: Vives, Juan Luis <http://d-nb.info/gnd/118627295>;
Einrichtung: Universitätsbibliothek, Mannheim, Germany
Signatur: XK 5503
Erscheinungsort: Augspurg
Erscheinungsjahr: 1544
Titel: Das erst Buch, ..., von den Jungkfrauen.
Titel: I. Von der auffziehung ainer Jungkfrauen. So sie noch nit reden kan.

Monographie Ioannis Lodovici Vivis Von vnde...
 Vorderdeckel
 Exlibris
 Titelseite
 Der ... Fürstin ... Jacoba, Pfaltzgreffin bey R...
 Die Titel des ersten Buchs ...
 Die Titel des andern Buchs ...
 Die Titel des dritten Buchs ...
 Illustration
 Das erst Buch, ..., von den Jungkfrauen. Ir
 I. Von der auffziehung ainer Jungk... II
 II. Von der vbrigen Kindthait. IIIr
 III. Von den ersten übungen. IIIr

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Don vnderweysung ainer Christi- chen Frauen/ Das erst Buch / wölche da ist / von den Jungkfrauen.
Don der auffziehung ainer Jungkfrauen/ So sie noch nit reden kan.

Er hochgeleret mañ Fabius Quinti- lianus/ da er leret vo der kunft des wolredens/ Achtet / das als bald von kindswelen auff / da- rinn anzefahen sey: Dann er mainet / das kain zeyt hingehen solt / wölche nit zum endt der für- gelezten kunft gebraucht wurde. Wieuol ain grosserer fleiß soll nun mit ainer Christi- chen Jungkfrauen fürgenomen werden / das wir nit allein ir auffnemen anrichten vnnd ordnen / Sondern auch den ersten Narung von anfang von der milch selbst / wölche ich wolt / so ferres möglich / das sie der mutterli- chen milch. von irer rechten mutter were / vnd diß haben auch gerathen Plutarchus Phaurinus / vnd ander trefflich geleert leüt. Dann also wirt die lieb ge- A gen

[15] - Ir

Von vnderweysung ainer Christi- lit / von den Jungkfrauen. Von der auffziehung ainer Jungkfrauen/ So sie noch nit reden kan.

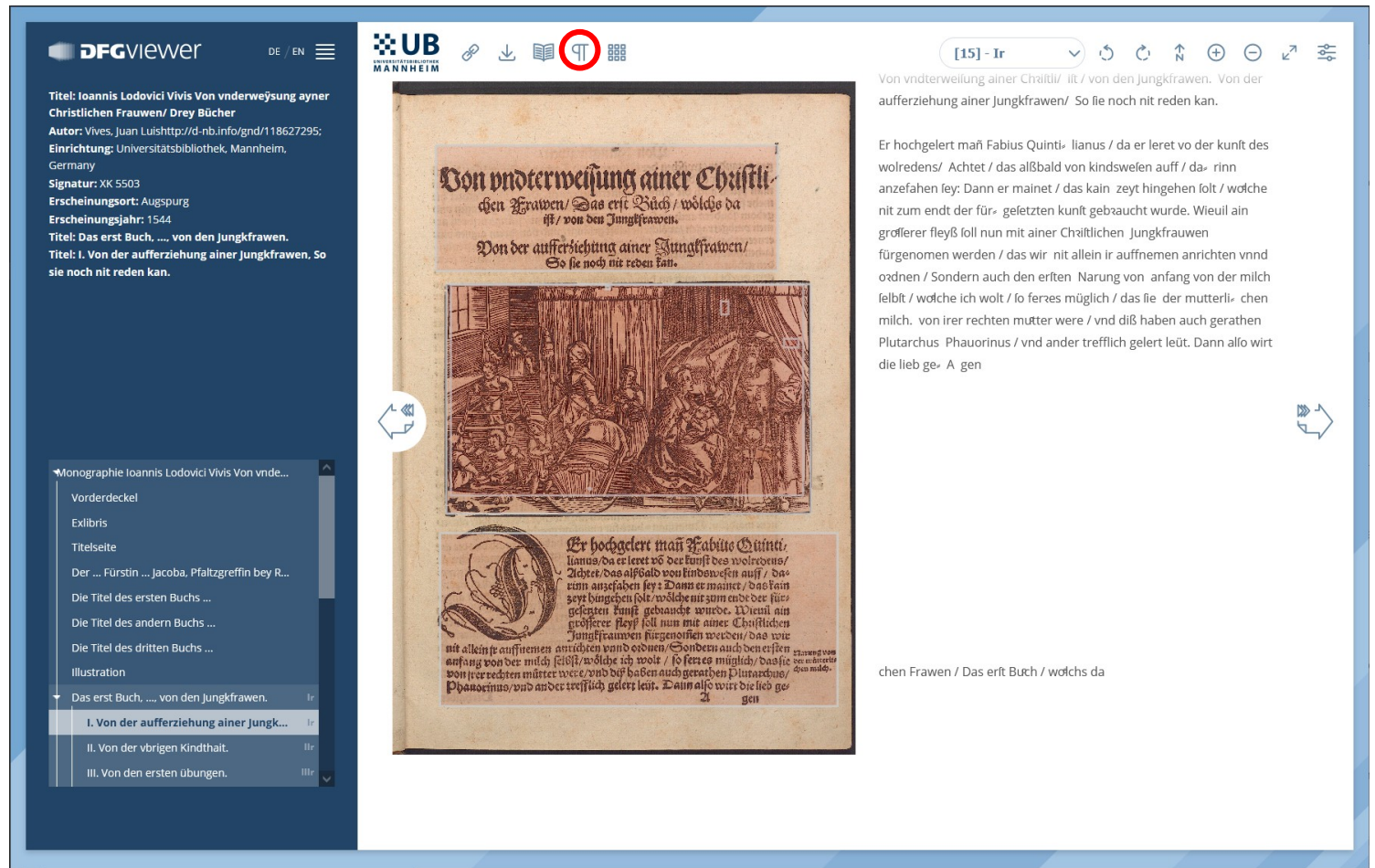
Er hochgeleret mañ Fabius Quinti- lianus/ da er leret vo der kunft des wolredens/ Achtet / das als bald von kindswelen auff / da- rinn anzefahen sey: Dann er mainet / das kain zeyt hingehen solt / wölche nit zum endt der für- gelezten kunft gebraucht wurde. Wieuol ain grosserer fleiß soll nun mit ainer Christi- chen Jungkfrauen fürgenomen werden / das wir nit allein ir auffnemen anrichten vnnd ordnen / Sondern auch den ersten Narung von anfang von der milch selbst / wölche ich wolt / so ferres möglich / das sie der mutterli- chen milch. von irer rechten mutter were / vnd diß haben auch gerathen Plutarchus Phaurinus / vnd ander trefflich geleert leüt. Dann also wirt die lieb ge- A gen

chen Frauen / Das erst Buch / wölchs da

DFG-Viewer: Beispiel mit Volltext

Der DFG-Viewer

- ... ist in der Lage, die (meisten) Digitalisate aus Archiven und Bibliotheken darzustellen.
- Dabei können optional auch die vorhandenen Volltexte angezeigt werden (im Bild rot markiert).



The screenshot shows the DFG-Viewer interface. On the left, a dark sidebar contains a table of contents for the monograph 'Ioannis Lodovici Vivis Von vnderweysung ayner Christlichen Frauen/ Drey Bücher'. The selected item is 'I. Von der auffziehung ainer Jungkfrauen. So sie noch nit reden kan.', which is highlighted in red. The main area displays a digital image of a manuscript page. The page features a title in German: 'Von vnderweysung ainer Christlichen Frauen/ Das erst Buch / wölche da ist / von den Jungkfrauen.' Below the title is a woodcut illustration of a classroom scene. The text on the page is in German and discusses the education of young women. A red circle highlights a specific part of the text in the original image. On the right side of the viewer, there is a search bar with '[15] - Ir' and various navigation icons. Below the search bar, the text from the manuscript is displayed in a clean, readable font, with the same red circle highlighting the corresponding text.

DFG-Viewer: Beispiel ohne Volltext

- Wenn keine Volltexte verfügbar sind, wird das entsprechende Symbol nicht angezeigt.

The screenshot shows the DFGviewer interface for a book record. On the left, a dark blue sidebar contains the following metadata:
Titel: Zur Psychologie des produktiven Denkens und des Irrtums
Autor: Selz, Otto;
Einrichtung: Universitätsbibliothek, Mannheim, Germany
Erscheinungsort: Bonn
Erscheinungsjahr: 1922
Below the sidebar, the text 'Monographie Zur Psychologie des produktiven D...' is visible.
The main content area displays a high-resolution image of the book cover. The cover text reads:
ZUR PSYCHOLOGIE
DES PRODUKTIVEN
DENKENS UND
DES IRRTUMS
EINE EXPERIMENTELLE UNTERSUCHUNG
VON
OTTO SELZ
A. O. PROFESSOR AN DER UNIVERSITÄT BONN
Below the printed text is a handwritten note: 'Lebstum', 'Gefault von Frau Mueller, die Nichts von Otto Selz isthaft in Israel.'
At the bottom of the cover, it says '1922' and 'VERLAG VON FRIEDRICH COHEN IN BONN'.
The interface includes a top navigation bar with the UB Mannheim logo, a search bar containing '171', and various icons for navigation and actions. A red question mark is positioned above the navigation bar. On the right side of the image, the text 'Volltext ist nicht vorhanden' is written vertically in red.

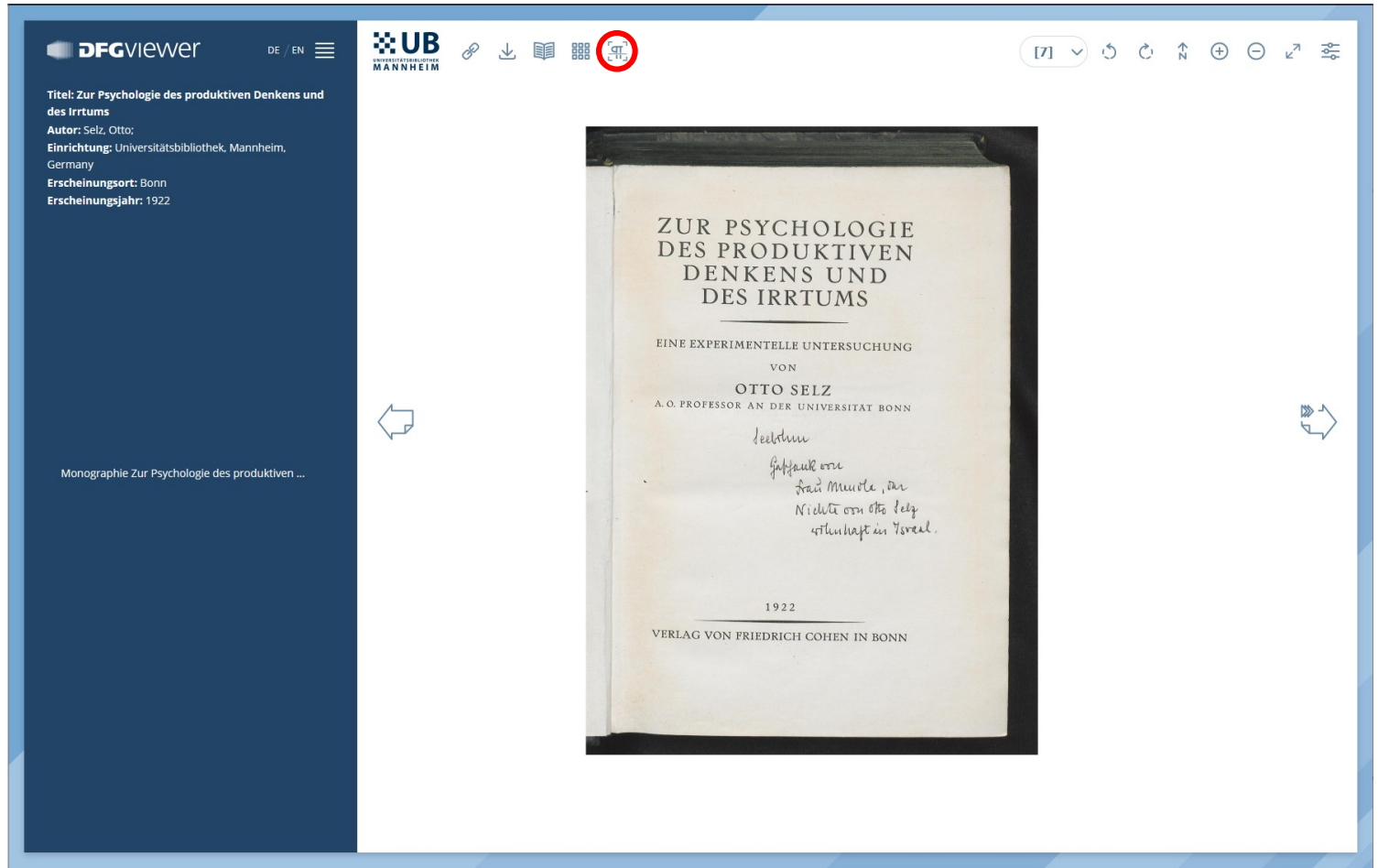
OCR On-Demand

DFG-Viewer: OCR On-Demand – Idee

- Idee: Nutzende können OCR für Werke ohne bzw. mit unbrauchbaren Volltexten anfordern.
- Der Volltext der aktuellen Seite steht nach wenigen Sekunden zur Verfügung, der Volltext des gesamten Werkes nach kurzer Zeit (abhängig vom Umfang).
- Dazu können verschiedene Optionen zur Volltextgenerierung angeboten werden.

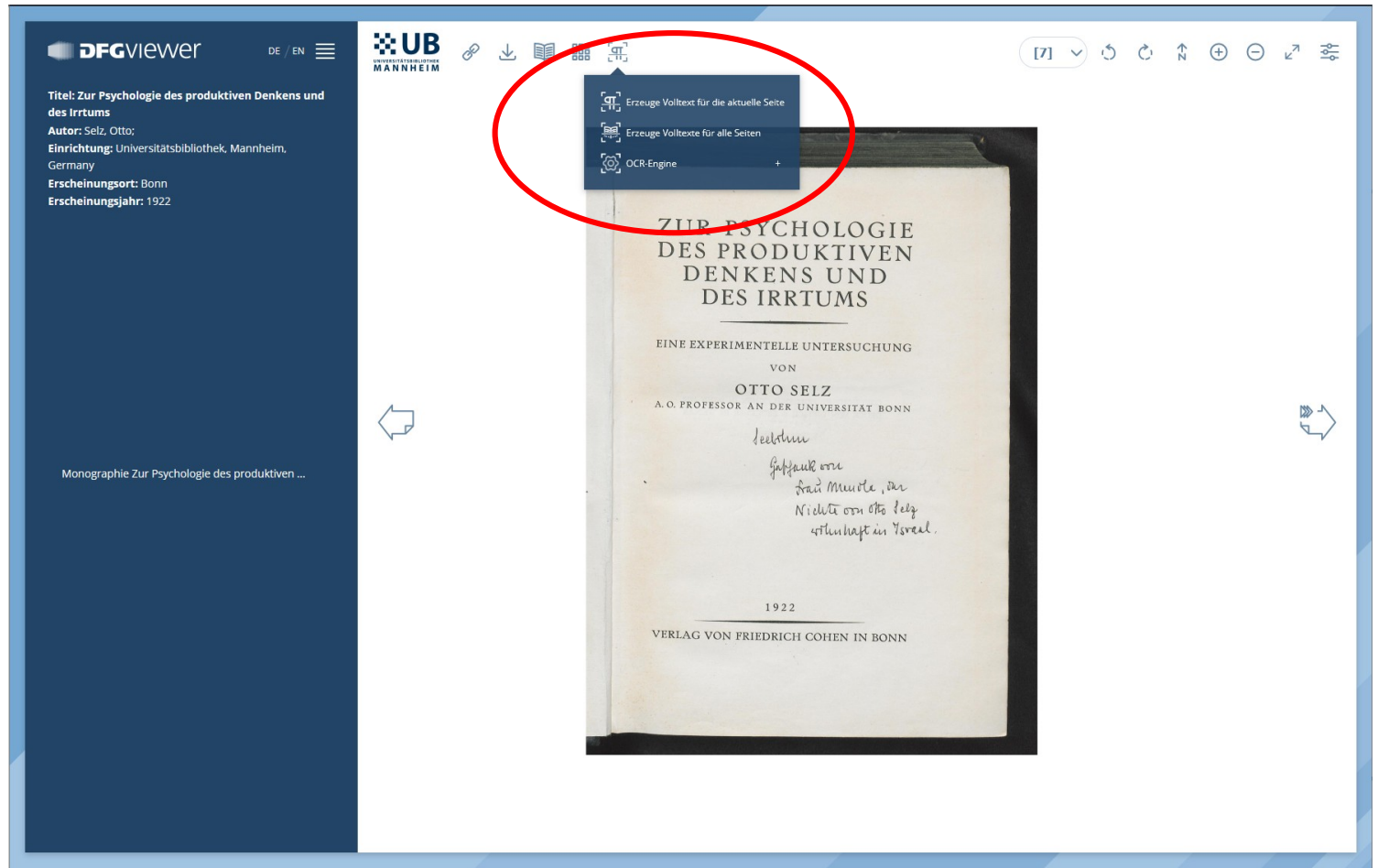
DFG-Viewer mit OCR On-Demand: Anwendung

- Wenn keine Volltexte verfügbar sind, können diese spontan erzeugt werden.
- Dazu wird ein neues Bedienelement eingeblendet (im Bild rot markiert).



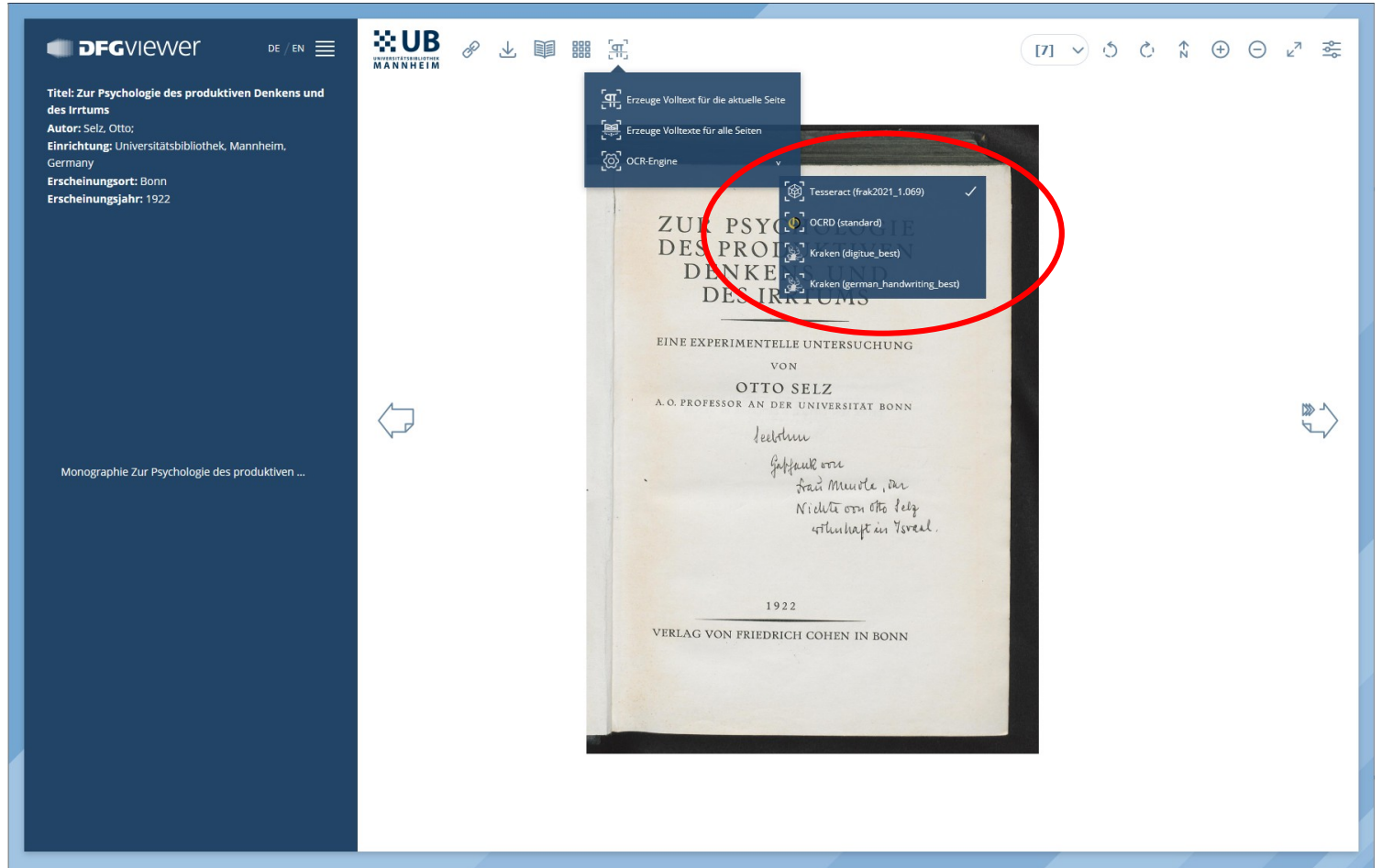
DFG-Viewer mit OCR On-Demand: Anwendung

- Das neue Bedienelement ermöglicht die Volltexterstellung für die jeweilige Seite oder das gesamte Werk.



DFG-Viewer mit OCR On-Demand: Anwendung

- Das neue Bedienelement ermöglicht die Volltexterstellung für die jeweilige Seite oder das gesamte Werk.
- Optional können verschiedene Techniken zur Volltexterstellung ausgewählt werden.



DFG-Viewer mit OCR On-Demand: Anwendung

- Nach erfolgreicher Volltexterkennung wird dieser angezeigt.
- 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. On the left, a dark blue sidebar contains the following text:

DFGviewer DE / EN

Titel: Zur Psychologie des produktiven Denkens und des Irrtums
Autor: Selz, Otto;
Einrichtung: Universitätsbibliothek, Mannheim, Germany
Erscheinungsort: Bonn
Erscheinungsjahr: 1922

Below the sidebar, the text "Monographie Zur Psychologie des produktiven ..." is visible.

The central area shows a scanned image of a book cover with the following text:

ZUR PSYCHOLOGIE
DES PRODUKTIVEN
DENKENS UND
DES IRRTUMS

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

Below the cover image, there is a handwritten note on a pink sticky note: "Jahres von 1922. Nicht von Otto Selz. Vertrieben in Israel".

At the bottom of the cover, it says: "1922 VERLAG VON FRIEDRICH COHEN IN BONN".

On the right side of the interface, there is a list of OCR results:

[7] ZUR PSYCHOLOGIE
DES PRODUKTIVEN
DENNENS UND
DES IRRTUMS

EINE EXPERIMENTELLE UNTERSUCHUNG

VON

OTTO SELZ

A. O. PROFESSOR AN DER DNIVERSITÄT BONN
ee

fe 8
Nu MU IC
Need c Uto 1.7
tc N 1 rtl,

1922

VERLAG VON FRIEDRICH COHEN IN BONN

Praktische Umsetzung

OCR-Engines

OCR-Engine = Technik zur Erzeugung von Volltexten

Ziel:

- Die Auswahl der angebotenen OCR-Engines obliegt der Einrichtung.
- Das Hinzufügen neuer OCR-Engines soll ohne großen Aufwand oder Eingriffe in den Programmcode möglich sein.



OCR-Engines: Umsetzung

- Alle benötigten Dateien befinden sich in der Verzeichnisstruktur von Kitodo.Presentation.
- Hier befinden sich die **OCR-Engines**, bei denen es sich um einfache **Shell-Skripte** handelt, und eine **JSON-Datei**, in der festgelegt werden kann, welche der OCR-Engines aktiv sein sollen sowie weitere Informationen zu den OCR-Engines.

OCR-Engines: Umsetzung

- Alle benötigten Dateien befinden sich in der Verzeichnisstruktur von Kitodo.Presentation.
- Hier befinden sich die **OCR-Engines**, bei denen es sich um einfache **Shell-Skripte** handelt, und eine **JSON-Datei**, in der festgelegt werden kann, welche der OCR-Engines aktiv sein sollen sowie weitere Informationen zu den OCR-Engines.
- Beispiel:

```
# ls -l
FullTextGenerationScripts
kraken-basic.sh
kraken-digitue_best.sh
kraken-
german_handwriting_best.sh
ocrd-basic.sh
ocrEngines.json
tesseract-basic.sh
```



Shell-Skripte = OCR-Engines

JSON-Datei mit Auflistung der aktiven OCR-Engines

OCR-Engines: Beispiel

```

{
  "ocrEngines": [
    {
      "name": "Tesseract",
      "de": "Tesseract (frak2021_1.069)",
      "en": "Tesseract (frak2021_1.069)",
      "class": "tesseract",
      "data": "tesseract-basic"
    },
    {
      "name": "Kraken (digitue_best)",
      "de": "Kraken (digitue_best)",
      "en": "Kraken (digitue_best)",
      "class": "kraken",
      "data": "kraken-digitue_best"
    },
    {
      "name": "Kraken
      (german_handwriting_best)",
      "de": "Kraken
      (german_handwriting_best)",
      "en": "Kraken
      (german_handwriting_best)",
      "class": "kraken",
      "data": "kraken-
      german_handwriting_best"
    }
  ]
}

```

Stefan Weil & Christos Sidiropoulos (Universitätsbibliothek Mannheim)

OCR-Engines: Beispiel

```

{
  "ocrEngines": [
    {
      "name": "Tesseract",
      "de": "Tesseract (frac2021_1.069)",
      "en": "Tesseract (frac2021_1.069)",
      "class": "tesseract",
      "data": "tesseract-basic"
    },
    {
      "name": "Kraken (digitue_best)",
      "de": "Kraken (digitue_best)",
      "en": "Kraken (digitue_best)",
      "class": "kraken",
      "data": "kraken-digitue_best"
    },
    {
      "name": "Kraken
(german_handwriting_best)",
      "de": "Kraken
(german_handwriting_best)",
      "en": "Kraken
(german_handwriting_best)",
      "class": "kraken",
      "data": "kraken-
german_handwriting_best"
    }
  ]
}

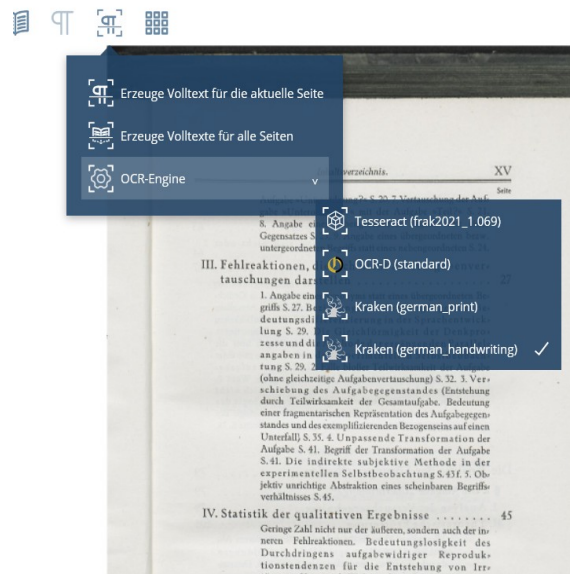
```

- } Eine der aktiven OCR-Engines mit zusätzlichen Informationen
- } Name und optionale Übersetzungen der Engine
- } Bezeichnung des Shell-Skriptes

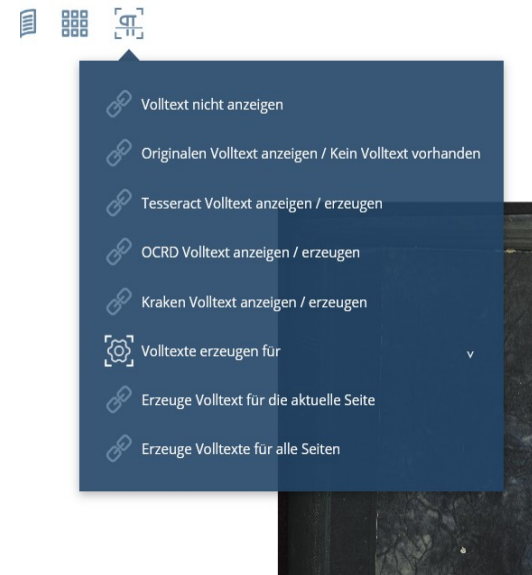
User Interface

User Interface: Variationen

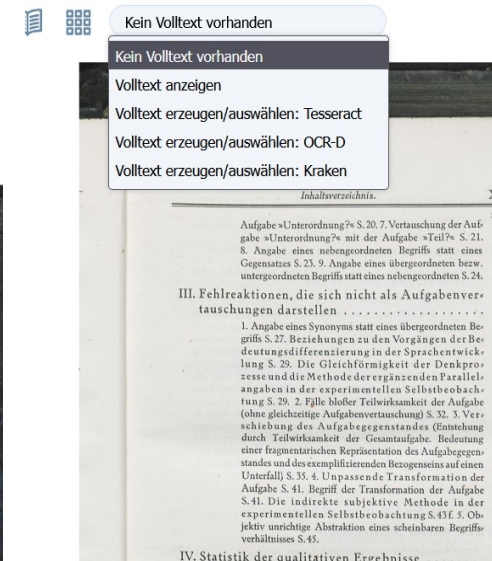
Aktueller Stand



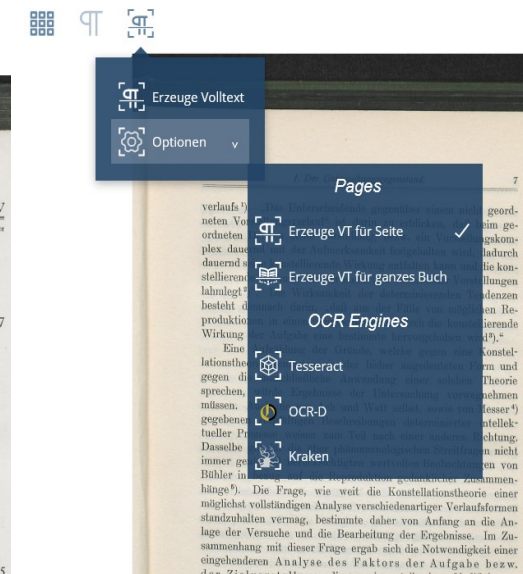
Option 1



Option 2



Option 3



Überblick

Laufende Arbeiten & Ausblick

- Restrukturierung der Oberfläche
- Indizierung verarbeiteter Werke:
 - Bereitstellung der Werke mit neu erzeugten Volltexten in „Sammlungen“ nach den besitzenden Einrichtungen
 - Suche in den Metadaten der Werke mit neu erzeugten Volltexten
 - Suche in den Volltexten der Werke mit neu erzeugten Volltexten
- OAI-Schnittstelle für Abruf neuer OCR-Ergebnisse durch besitzende Einrichtungen

Wunschliste

- Maschinelle Übersetzung der Volltexte
- Text to speech – Audioausgabe der Volltexte (Barrierefreiheit)
- Feedback-Möglichkeit zu den OCR-Resultaten
- Korrigierbare Volltexte (auch für Nachtraining verwendbar)

Weitere Informationen

- Prototyp:
→ <https://dfg-viewer.bib.uni-mannheim.de/>
- Installationsanleitung:
→ <https://github.com/UB-Mannheim/kitodo-presentation/wiki>
- Docker Konfiguration:
→ <https://github.com/UB-Mannheim/kitodo-presentation-docker>
- DFG-Viewer mit OCR-On-Demand:
→ <https://github.com/UB-Mannheim/dfg-viewer/tree/6.x-ocr>
- Kitodo.Presentation mit OCR-On-Demand:
→ <https://github.com/UB-Mannheim/kitodo-presentation/tree/4.x-ocr>
- Projektplan:
→ <https://github.com/orgs/UB-Mannheim/projects/2>

Vielen Dank!

Stefan Weil

stefan.weil@uni-mannheim.de

Christos Sidiropoulos

christos.sidiropoulos@uni-mannheim.de