Kitodo Praxistreffen 2022 in Braunschweig

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

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Stefan Weil

2022-10-20

DFG

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

OCR-Projekte an der UB Mannheim

UB
UNIVERSITÄTSBIBLIOTHEK
MANNHEIM

- OCR-D: Tesseract als Komponente im OCR-D-Workflow (2018–2019)
- OCR-BW – Kompetenzzentrum OCR der Universitätsbibliotheken Mannheim und Tübingen (2019–2022)
- OCR-D: Workflow für werkspezifisches Training auf Basis generischer Modelle mit OCR-D sowie Ground-Truth-Aufwertung (2021–2023)
- OCR-D: **Integration von Kitodo und OCR-D zur produktiven Massendigitalisierung** (2021–2023) gemeinsam mit SLUB Dresden und UB Braunschweig

dfg-viewer.de: Referenzimplementierung auf Basis von Kitodo.Presentation

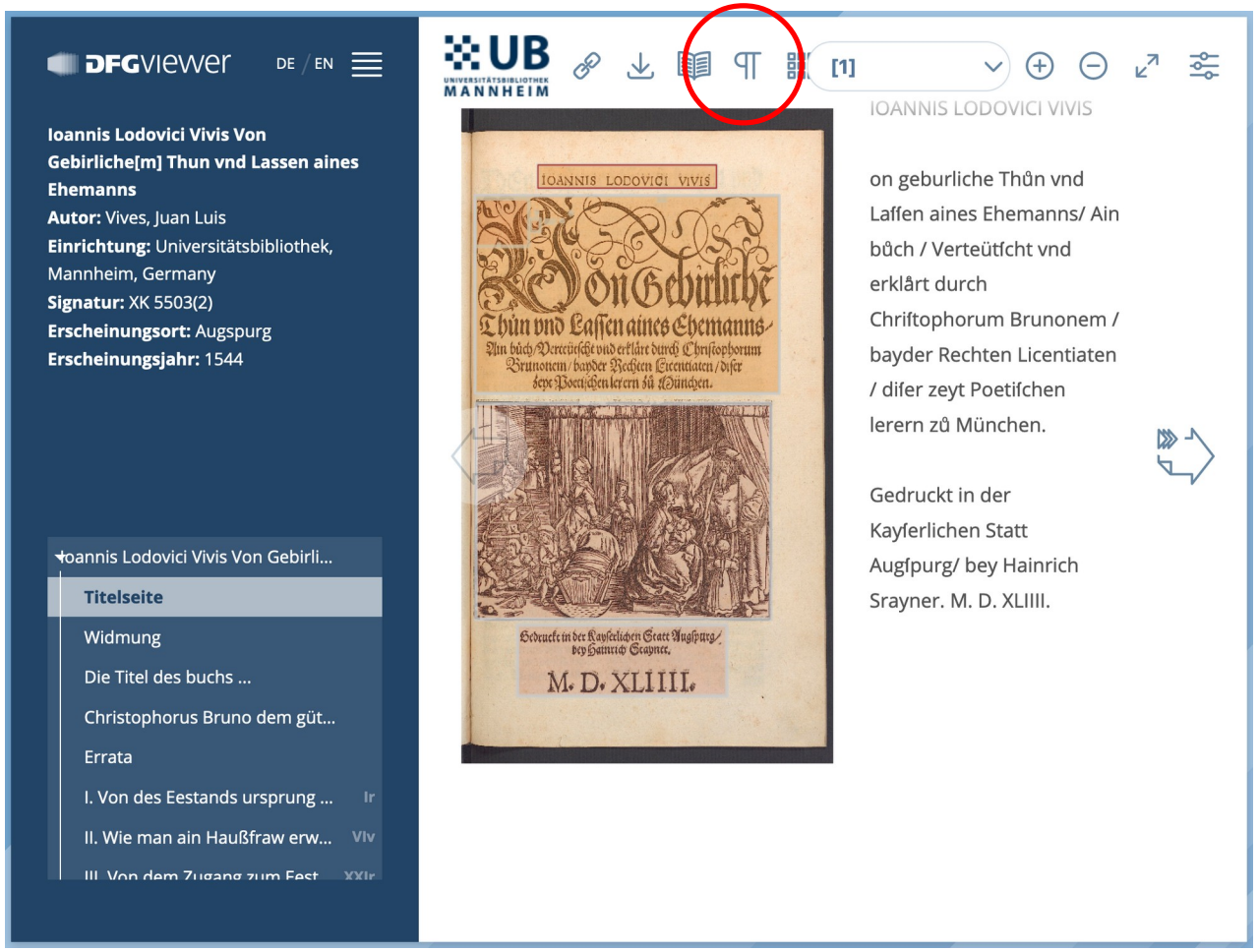
UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

Startseite DFG-Viewer

- Entwicklung u. Betrieb durch SLUB Dresden
- Förderung durch DFG

DFG-Viewer: Beispiel mit Volltext

- Der DFG-Viewer kann die (meisten) Digitalisate von Archiven und Bibliotheken darstellen.
- Dabei zeigt er optional auch vorhandene Volltexte an (im Bild rot markiert).

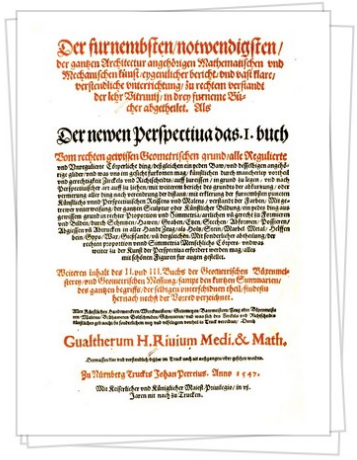


Aber: zahlreiche Digitalisate ohne Volltext

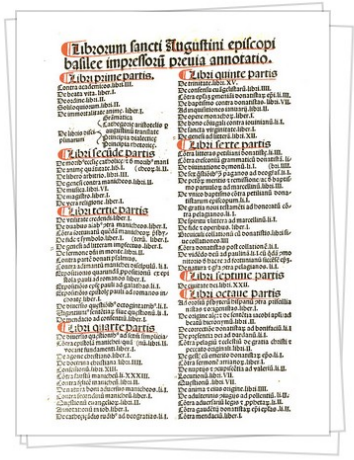
- Viele Digitalisate (> 90 % ?) haben noch keine Volltexte.
- „Für einige der digitalisierten Objekte sind digitale Volltexte hinterlegt.“
- Auch die Demos im DFG-Viewer sind ohne Volltexte.

DFG-Viewer Demonstrator

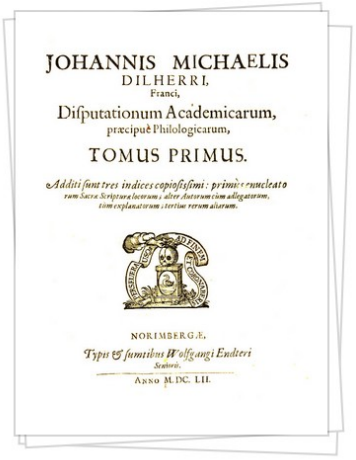
Sehen Sie hier zunächst drei Demonstrationen des DFG-Viewers mit ausgewählten Werken verschiedener Projektpartner.



Der furnembsten, notwendigsten, der gantzen Architectur angehoerigen Mathematischen und...
oai:de:slub-dresden:db:id-263566811
SLUB Dresden



Prima [- Undecima] pars librorum divi Aurelii Augustini...
bsb00020619
BSB München

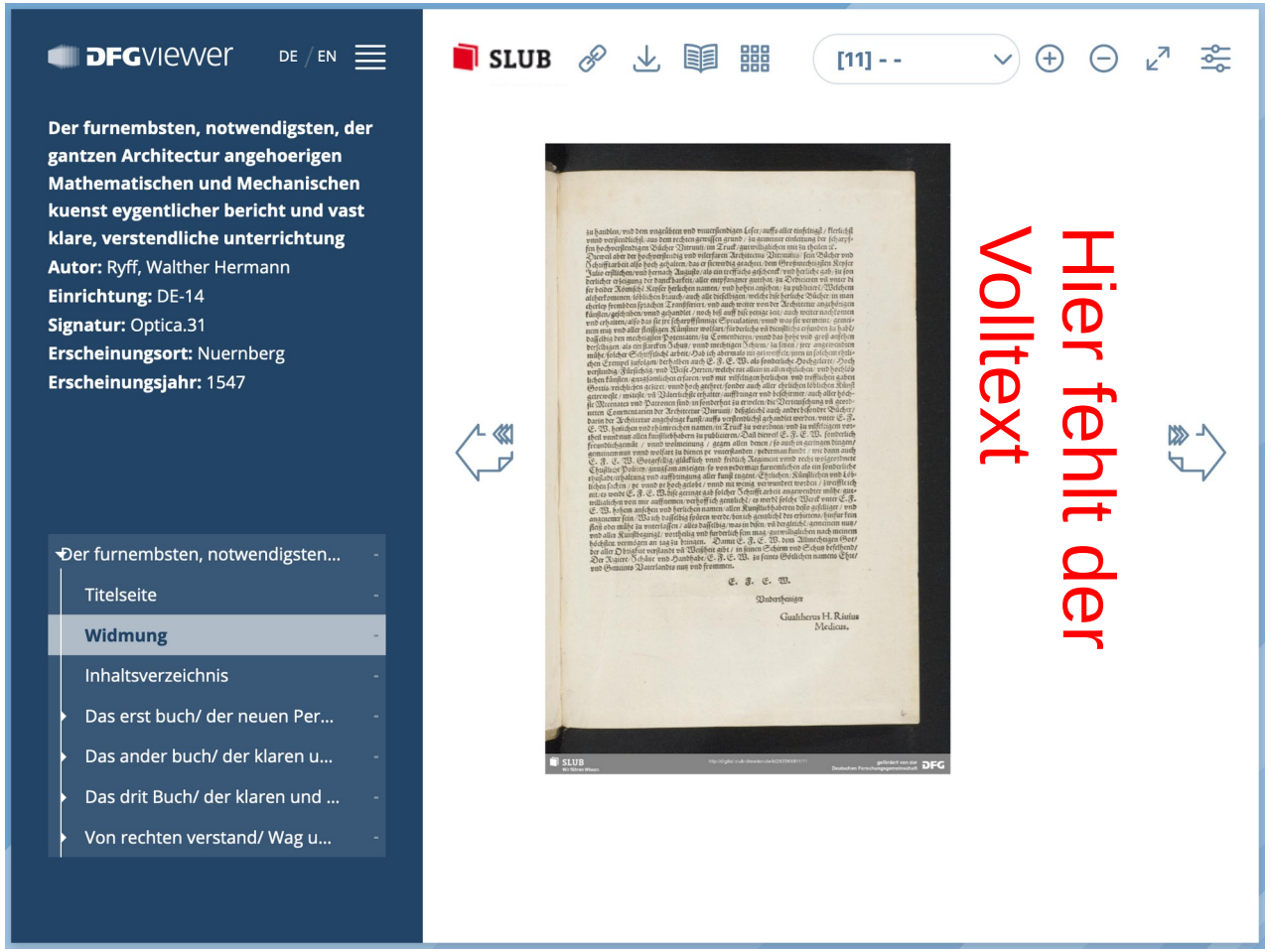


Johannis Michaelis Dilherri, Franci, Disputationum Academicarum, praecipue Philologicarum, Tomus...
oai:diglib.hab.de:ppn_549836969
HAB Wolfenbüttel

DFG-Viewer: Beispiel ohne Volltext

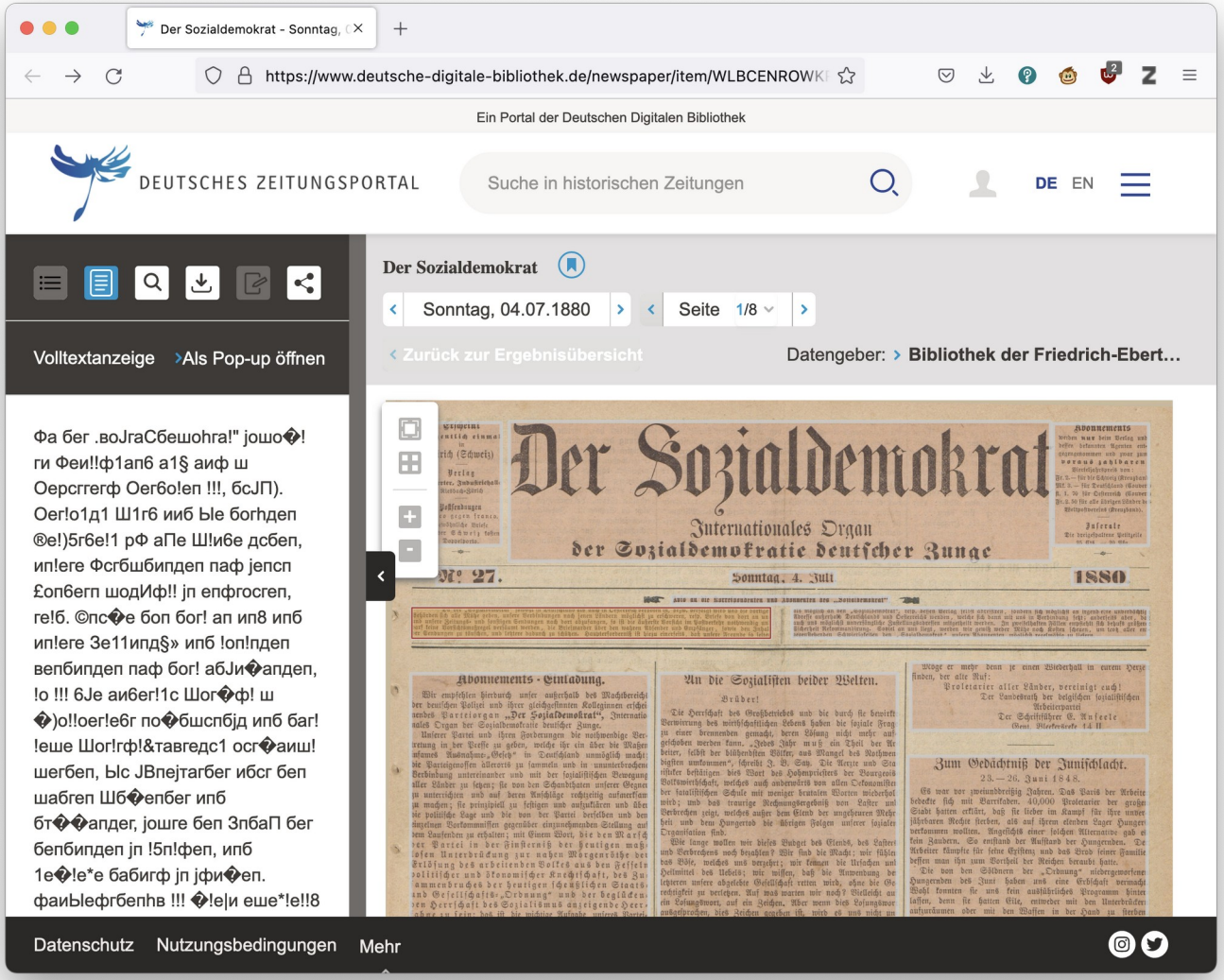
?

- Wenn keine Volltexte bereitgestellt werden, wird auch das entsprechende Symbol nicht angezeigt.



Zeitungsportal: Beispiel unbrauchbarer Volltext

- Bereitgestellte Volltexte sind gelegentlich unbrauchbar
- Mögliche Ursache: falsche Einstellung für Schrift (z. B. Antiqua / Fraktur) oder Sprache bei der OCR-Erstellung



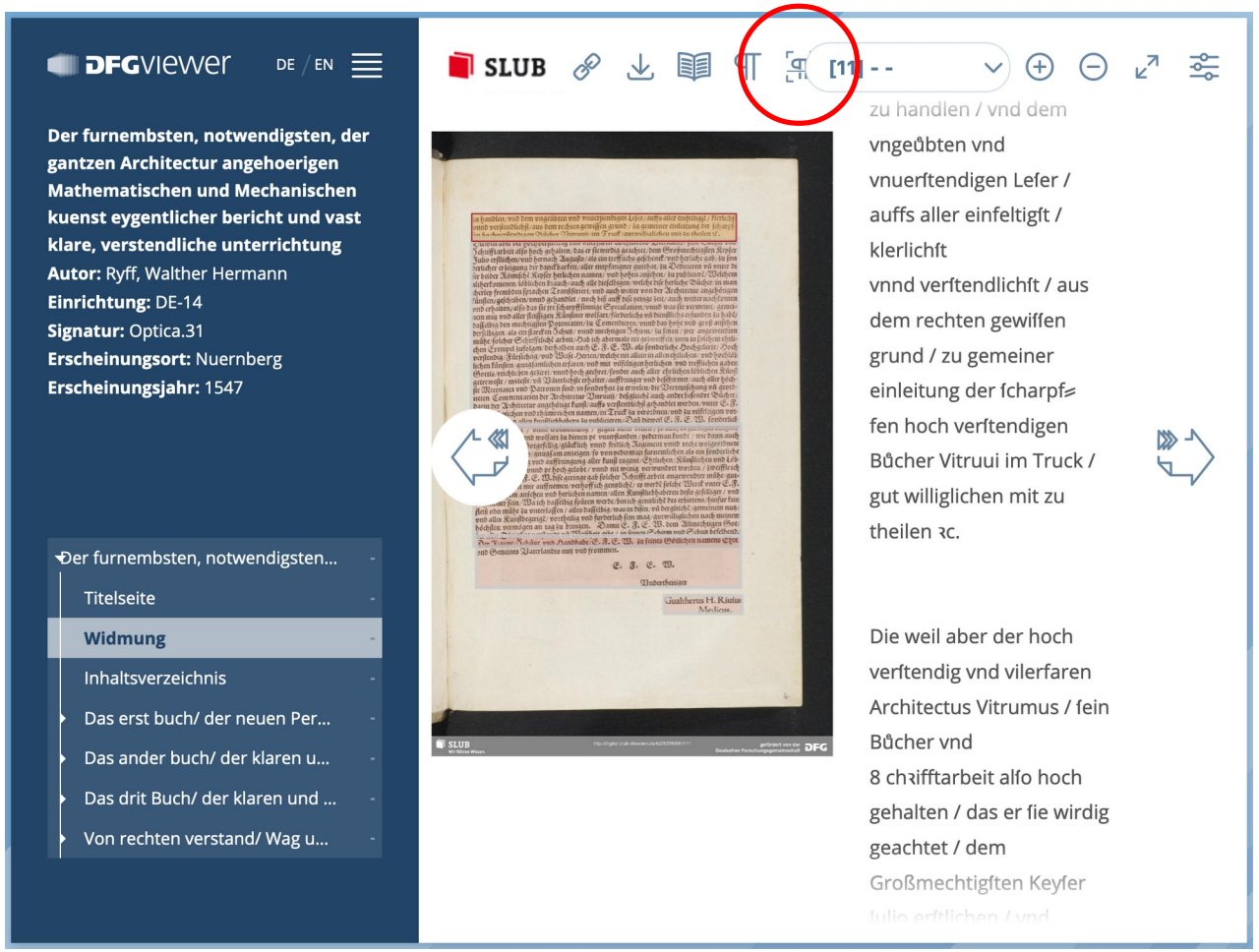
DFG-Viewer (und Kitodo.Presentation) mit OCR On-Demand – Rückblick

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Idee: Nutzende können OCR für Werke ohne bzw. mit unbrauchbaren Volltexten anfordern
- Volltext für aktuelle Seite wird nach wenigen Sekunden, für das gesamte Werk zeitnah (abhängig vom Umfang), bereitgestellt
- Erster Prototyp erstellt im KIT Summer of Code for Society (KIT SOC4S) durch Maxim Popov (2020)
  https://github.com/KIT-SOC4S/dfg-viewer
  https://github.com/KIT-SOC4S/kitodo-presentation

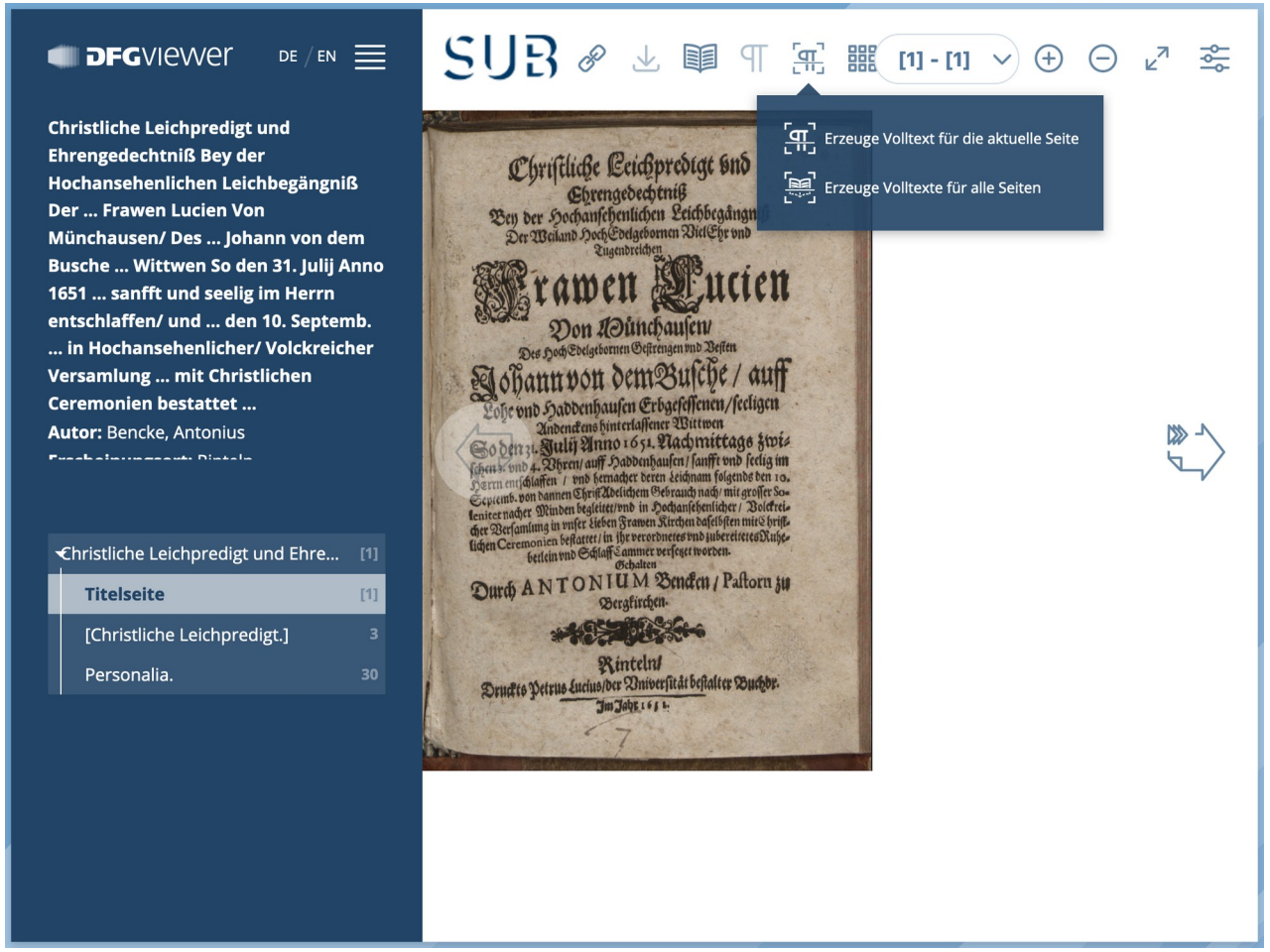
DFG-Viewer: Beispiel mit OCR On-Demand

- Wenn keine Volltexte bereitgestellt werden, können diese spontan erzeugt werden, wahlweise für die aktuelle Seite oder das ganze Werk.
- Dauer für Beispiel: 1 s bis 7 s pro Seite, 39 min für 694 Seiten (parallelisierbar)



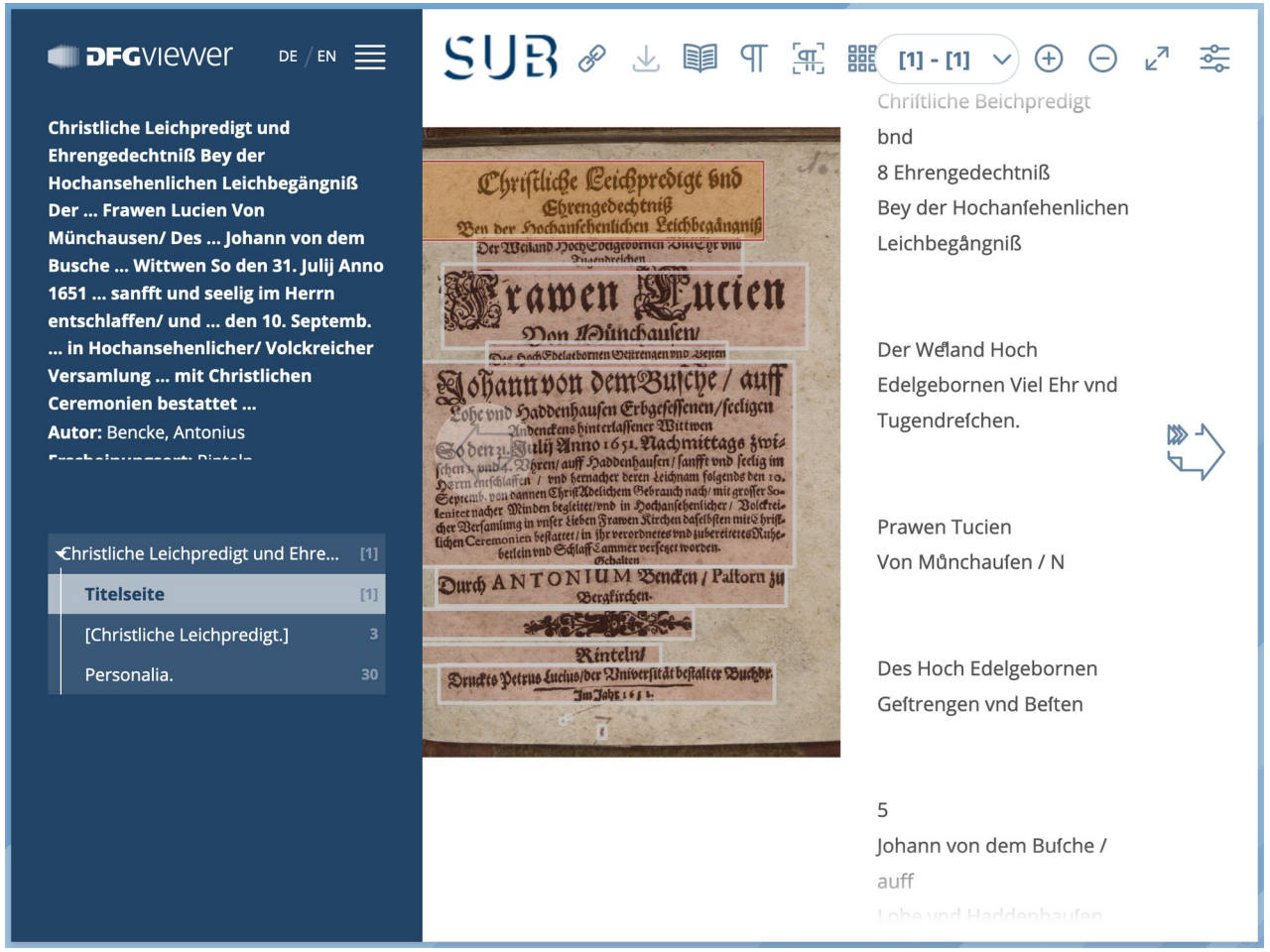
DFG-Viewer: OCR OnDemand im Einsatz

- Neues Bedienelement erzeugt Volltext für die aktuelle Seite oder das ganze Werk.



DFG-Viewer: OCR On-Demand Ergebnis

- Neues Bedienelement erzeugt Volltext für die aktuelle Seite oder das ganze Werk.



DFG-Viewer (und Kitodo.Presentation) mit OCR On-Demand – aktueller Stand

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Weiterentwicklung des Prototypen im Rahmen des DFG-Projektes *OCR-D: Integration von Kitodo und OCR-D zur produktiven Massendigitalisierung*“ (SLUB Dresden, UB Braunschweig, UB Mannheim) seit Januar 2022 durch Christos Sidiropoulos (UB Mannheim)
  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
- Projektplan
  https://github.com/orgs/UB-Mannheim/projects/2

DFG-Viewer (und Kitodo.Presentation) mit OCR On-Demand – laufende Arbeiten

UB UNIVERSITÄTSBIBLIOTHEK MANNHEIM

- Stabilerer Betrieb des Prototypen
- Wechsel auf TYPO3 10 mit neuesten Softwarereleases
- Auswahl zwischen alternativen OCR-Prozessen
- 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

Bei guter OCR-Qualität bieten sich neue Funktionalitäten an:

- Maschinelle Übersetzung der Volltexte
- Text to speech – Audioausgabe der Volltexte (Barrierefreiheit)

Erste Versuche dazu waren vielversprechend, zeigten aber auch, dass nicht alle gängige Anwendungen mit Spezialitäten wie beispielsweise dem langen s klar kommen.

Wunschliste (Fortsetzung)

Außerdem:

- Wie sieht die optimale Bedienerführung aus?
- Feedback-Möglichkeit zu den OCR-Resultaten
- Korrigierbare Volltexte (auch für Nachtraining verwendbar)

- Ihre Anregungen / Wünsche