

Kitodo Praxistreffen 2022 in Braunschweig

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



Stefan Weil
2022-10-20



OCR-Projekte an der UB 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

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).

The screenshot displays the DFG-Viewer interface. On the left, a dark blue sidebar contains the book's metadata: **Ioannis Lodovici Vivis Von Gebirliche[m] Thun vnd Lassen aines Ehemanns**, **Autor:** Vives, Juan Luis, **Einrichtung:** Universitätsbibliothek, Mannheim, Germany, **Signatur:** XK 5503(2), **Erscheinungsort:** Augspurg, **Erscheinungsjahr:** 1544. Below this is a table of contents with the title page highlighted in grey.

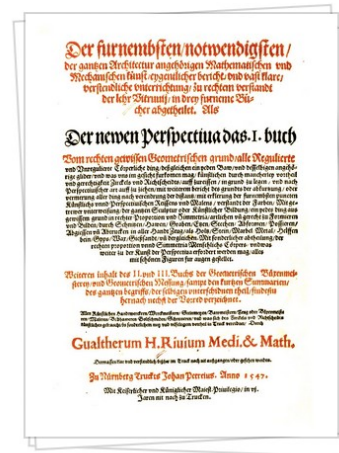
The main content area shows the title page of the book, which is a woodcut print. The title is written in a large, ornate Gothic script: **VON GEBIRLICHE** Thun vnd Lassen aines Ehemanns. The page is framed by a decorative border. A red circle highlights the 'Full Text' icon (represented by a document with a magnifying glass) in the top navigation bar. To the right of the image, the text of the title page is transcribed: **IOANNIS LODOVICI VIVIS** on geburliche Thün vnd Laffen aines Ehemanns/ Ain büch / Verteüfcht vnd erklärt durch Chriftophorum Brunonem / bayder Rechten Licentiaten / dife zeyt Poetifchen lerern zü München. Gedruckt in der Kayferlichen Statt Augfpurg/ bey Hainrich Srayner. M. D. XLIIII.

Aber: zahlreiche Digitalisate ohne Volltext

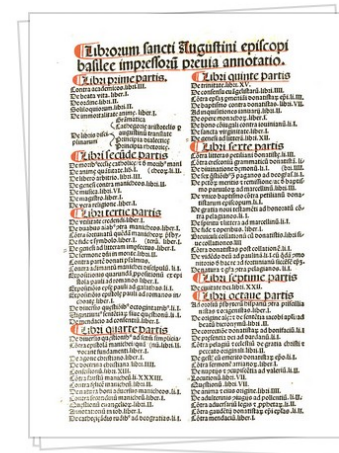
- Viele Digitalisate (> 90 % ?) haben noch keine Volltexte.
- „Für einige der digitalisierten Objekte sind digitale Volltexte hinterlegt.“
- Auch die Demos im DFG-Viewer sind ohne Volltexte.

DFG-Viewer Demonstrator

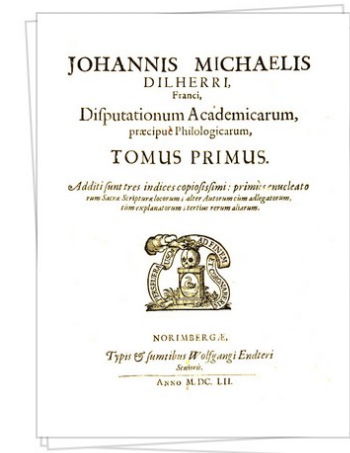
Sehen Sie hier zunächst drei Demonstrationen des DFG-Viewers mit ausgewählten Werken verschiedener Projektpartner.



Der furnembsten, notwendigsten, der ganzen Architectur angehorigen 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.

DFGviewer DE / EN

Der furnembsten, notwendigsten, der ganzen Architectur angehorigen Mathematischen und Mechanischen kuenst eygentlicher bericht und vast klare, verständliche unterrichtung
Autor: Ryff, Walther Hermann
Einrichtung: DE-14
Signatur: Optica.31
Erscheinungsort: Nuernberg
Erscheinungsjahr: 1547

Der furnembsten, notwendigsten...	-
Titelseite	-
Widmung	-
Inhaltsverzeichnis	-
Das erst buch/ der neuen Per...	-
Das ander buch/ der klaren u...	-
Das drit Buch/ der klaren und ...	-
Von rechten verstand/ Wag u...	-

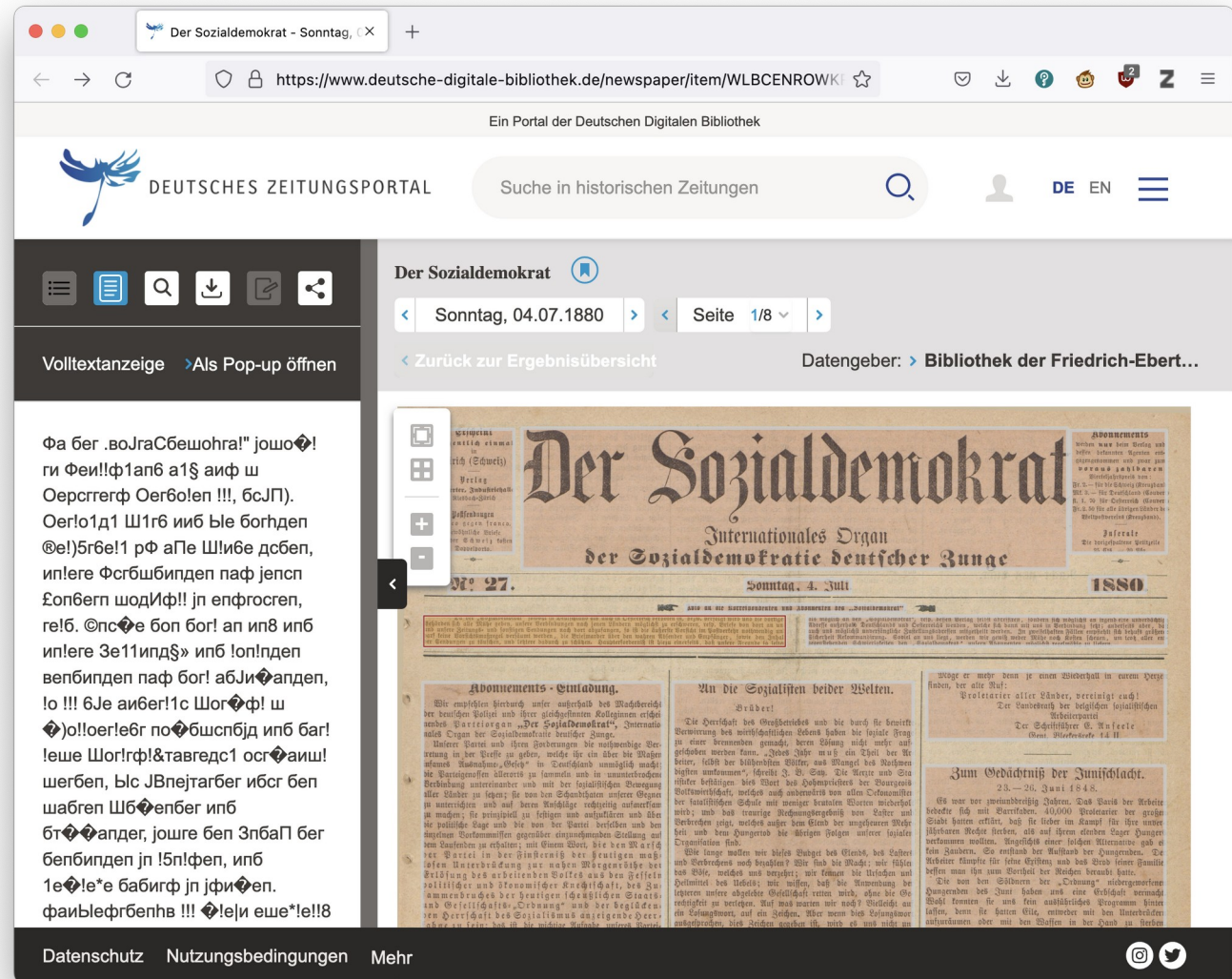
SLUB [11] --

Hier fehlt der Volltext

DFG

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

- 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)

DFGviewer DE / EN

Der furnembsten, notwendigsten, der ganzen Architectur angehorigen Mathematischen und Mechanischen kuenst eygentlicher bericht und vast klare, verständliche unterrichtung
Autor: Ryff, Walther Hermann
Einrichtung: DE-14
Signatur: Optica.31
Erscheinungsort: Nuernberg
Erscheinungsjahr: 1547

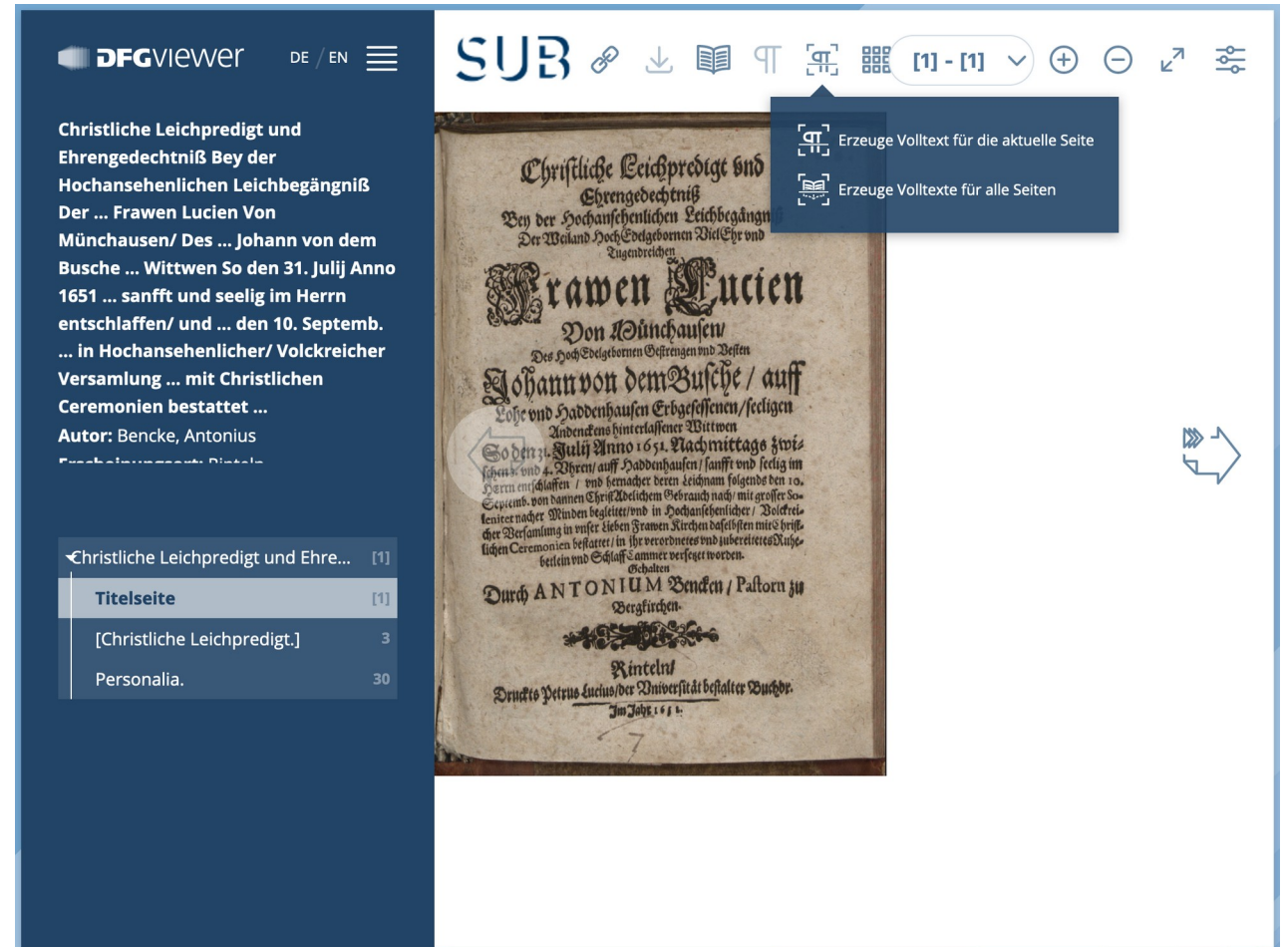
- Der furnembsten, notwendigsten...
- Titelseite
- Widmung**
- Inhaltsverzeichnis
- Das erst buch/ der neuen Per...
- Das ander buch/ der klaren u...
- Das drit Buch/ der klaren und ...
- Von rechten verstand/ Wag u...

zu handeln / vnd dem vngeübten vnd vnuerftendigen Lefer / auffß aller einfeltigt / klerlichft vnnnd verftendlichft / aus dem rechten gewiffen grund / zu gemeiner einleitung der fcharpfe fen hoch verftendigen Bücher Vitruui im Truck / gut williglichen mit zu theilen ꝛc.

Die weil aber der hoch verftendig vnd vilerfaren Architectus Vitrumus / fein Bücher vnd 8 chrifftarbeit alfo hoch gehalten / das er fie wirdig geachtet / dem Großmechtigten Keyfer Iulia erftlichen / vnd

DFG-Viewer: OCR OnDemand im Einsatz

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



The screenshot shows the DFGviewer interface. On the left, a dark blue sidebar contains the document title and a table of contents. The main area displays a scanned page of a historical document in German, featuring a large title and several columns of text. A tooltip is visible over the document, offering two options: 'Erzeuge Volltext für die aktuelle Seite' and 'Erzeuge Volltexte für alle Seiten'. The interface includes a top navigation bar with the 'SUB' logo and various icons for navigation and search.

DFGviewer DE / EN

Christliche Leichpredigt und Ehrengedächtniß Bey der Hochansehnlichen Leichbegängniß Der ... Frawen Lucien Von Münchhausen/ Des ... Johann von dem Busche ... Wittwen So den 31. Julij Anno 1651 ... sanfft und seelig im Herrn entschlaffen/ und ... den 10. Septemb. ... in Hochansehnlicher/ Volckreicher Versammlung ... mit Christlichen Ceremonien bestattet ...
Autor: Bencke, Antonius

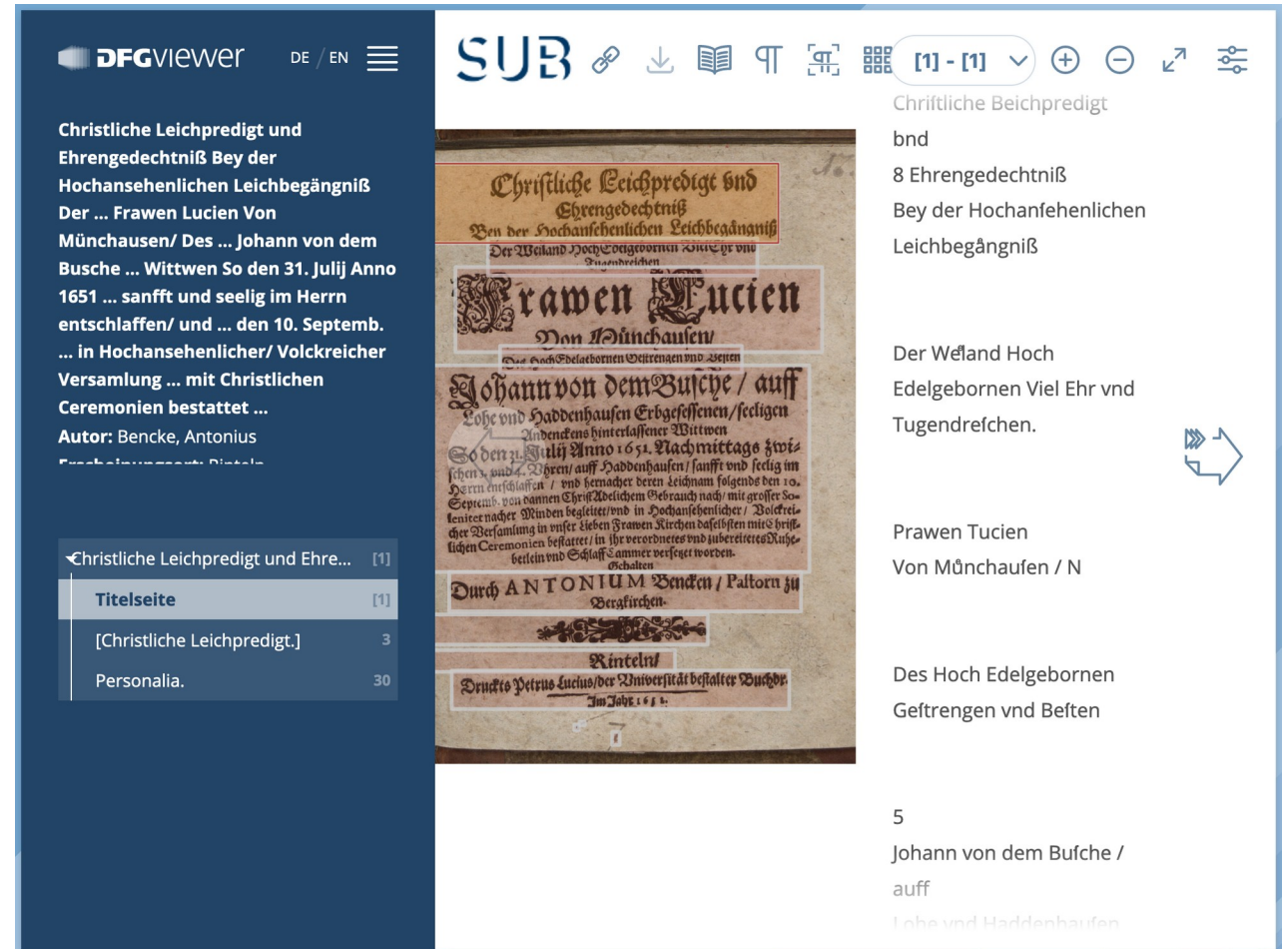
Christliche Leichpredigt und Ehre...	[1]
Titelseite	[1]
[Christliche Leichpredigt.]	3
Personalia.	30

Christliche Leichpredigt und Ehrengedächtniß
Bey der Hochansehnlichen Leichbegängniß
Der Woland hoch Edelgebornen WitEhr und
Lugendreichen
Frawen Lucien
Von Münchhausen
Des hoch Edelgebornen Vätergen und Vessen
Johann von dem Busche / auff
Lohse und Haddenhausen Erbgesessenen / seligen
Andere dem hinterlassener Wittwen
So den 31. Julij Anno 1651. Nachmittags zwis-
schen 4. Uhren auff Haddenhausen / sanfft und seelig im
Herrn entschlaffen / und hernacher deren Leichnam folgendes den 10.
Septemb. von bannen Christlichen Gestruch nach / mit großer So-
leniter nächer Münden begleitet / und in Hochansehnlicher / Doctrei-
cher Versammlung in unser Lieben Frawen Kirchen dasehesten mit christ-
lichen Ceremonien bestattet / in Ihr verordneter und zubereiteter Kiste
bestelt und Eschaffsammer versetzt worden.
Schalt
Durch **ANTONIUM Bencken /** Pastorn zu
Bergkirchen.
Kinteln!
Druckts Petrus Lucius / der Universität bestalter Buchdr.
Im Jahr 1651.

Erzeuge Volltext für die aktuelle Seite
Erzeuge Volltexte für alle Seiten

DFG-Viewer: OCR On-Demand Ergebnis

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



DFGviewer DE / EN

Christliche Leichpredigt und Ehrengedechtniß Bey der Hochansehnlichen Leichbegängniß Der ... Frawen Lucien Von Münchauen/ Des ... Johann von dem Busche ... Wittwen So den 31. Julij Anno 1651 ... sanfft und seelig im Herrn entschlaffen/ und ... den 10. Septemb. ... in Hochansehnlicher/ Volckreicher Versammlung ... mit Christlichen Ceremonien bestattet ...
Autor: Bencke, Antonius

Christliche Leichpredigt und Ehre...	[1]
Titelseite	[1]
[Christliche Leichpredigt.]	3
Personalia.	30

SUB [1] - [1] + - ↗ ⚙

Christliche Beichpredigt
 bnd
 8 Ehrengedechtniß
 Bey der Hochansehnlichen
 Leichbegängniß

Der Weiland Hoch
 Edelgebornen Viel Ehr vnd
 Tugendreichen.

Prawen Tucien
 Von Münchauen / N

Des Hoch Edelgebornen
 Gftrennen vnd Besten

5
 Johann von dem Busche /
 auff
 Lobe vnd Haddenhausen

*Christliche Leichpredigt und Ehrengedechtniß
 Bey der Hochansehnlichen Leichbegängniß
 Der Weiland HochEdelgebornen Wittwen vnd
 Prawn Tucien
 Von Münchauen
 Des Hoch Edelgebornen
 Gftrennen vnd Besten
 Johann von dem Busche / auff
 Lobe vnd Haddenhausen
 Erbgesessenen / seeligen
 Andenkens hinf erlassener Wittwen
 So den 31. Julij Anno 1651. Nachmittags zwöl-
 fhen vnd 4. Uhren/ auff Haddenhausen/ sanfft und seelig im
 Herrn entschlaffen / vnd hernacher deren Leichnam folgendes den 10.
 Septemb. von damen Christl. Adelichen Gesrauch nach / mit großer Son-
 tenarter nachher Begleitet/ vnd in Hochansehnlicher / Volckrei-
 cher Versammlung in unser lieben Frawen Kirchen darselbst mit christ-
 lichen Ceremonien bestattet / in ihre verordnete vnd zubereitete Kuch-
 schalen*

Durch ANTONIUM Bencke / Paltoru zu
 Bercksirchen.

Kirteln/
 Druckto Petrus Luchs/der Universit. befallter Buchdr.
 Im Jahr 1651.

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

- 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

- Stabilerer Betrieb des Prototypen
- Wechsel auf TYPO3 10 mit neuesten Software-releases
- 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