Landesarchiv Baden-Württemberg

Werkzeuge zur Texterkennung
Ein Blick in die digitale Werkstatt des FDMLab am Landesarchiv Baden-Württemberg

Südwestdeutscher Archivtag, 16. bis 18. Juni 2021

Benjamin Rosemann, Elisabeth Klindworth

Das Projekt FDMLab@LABW

Landesarchiv Baden-Württemberg

Laufzeit von Juli 2020 bis Juni 2022

Kontext

- Das Projekt steht im Kontext des Aufbaus einer nationalen Forschungsdateninfrastruktur (NFDI).
- Förderung durch die Baden-Württemberg Stiftung im Rahmen der Zukunftsoffensive III

Projektziele

- Aufbau einer im Landesarchiv angesiedelten Basisinfrastruktur im Bereich E-Science und Forschungsdatenmanagement (FDM)
- Verbesserung der Zugänglichkeit und Nutzbarkeit des Archivguts
  - Antworten auf die steigende Nachfrage nach standardisierten, übergreifend auswertbaren Daten geben

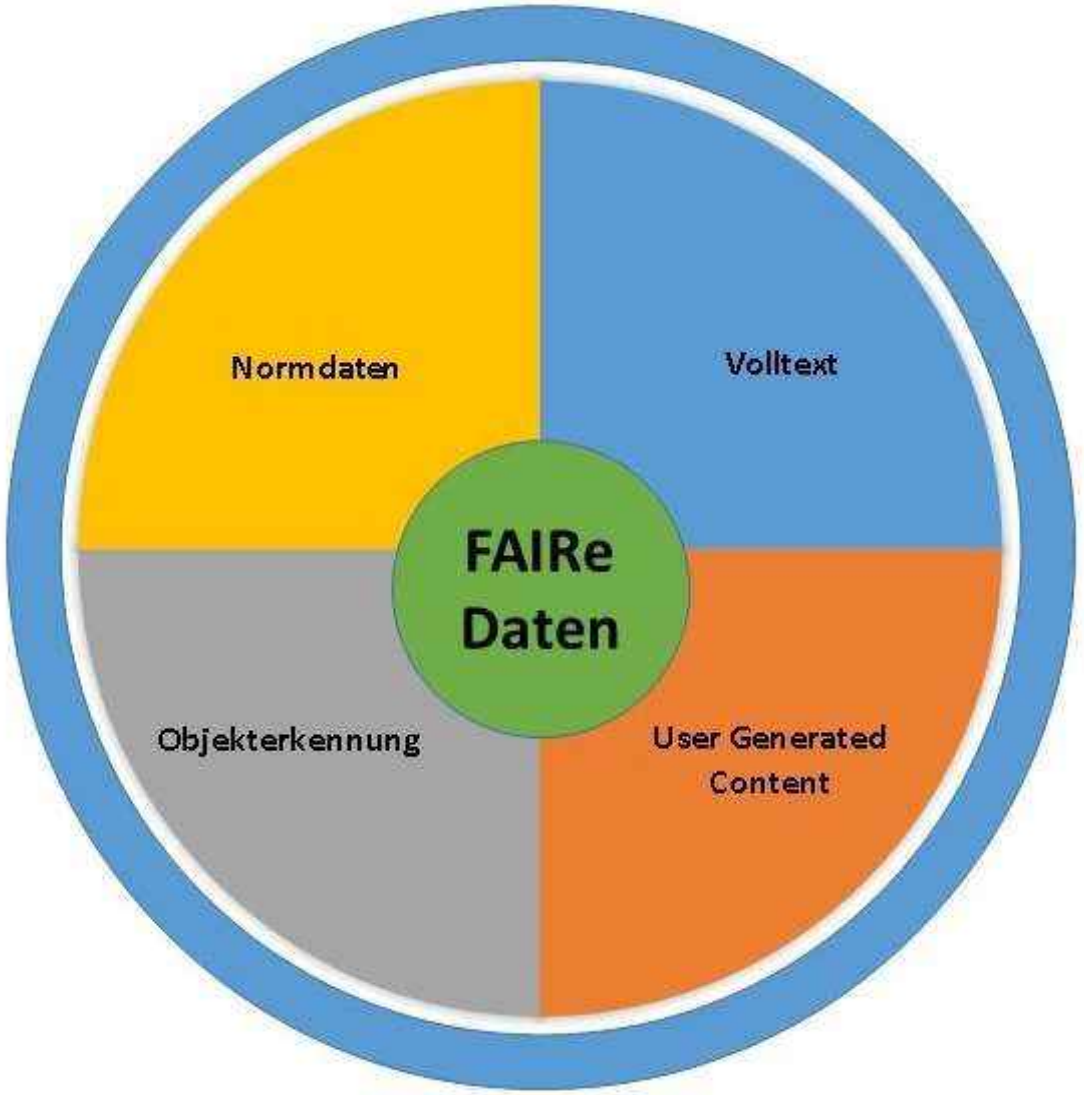
Arbeitsfelder des FDMLab

Landesarchiv Baden-Württemberg

Weitere Informationen

Projektseite auf der LABW-Homepage [1]

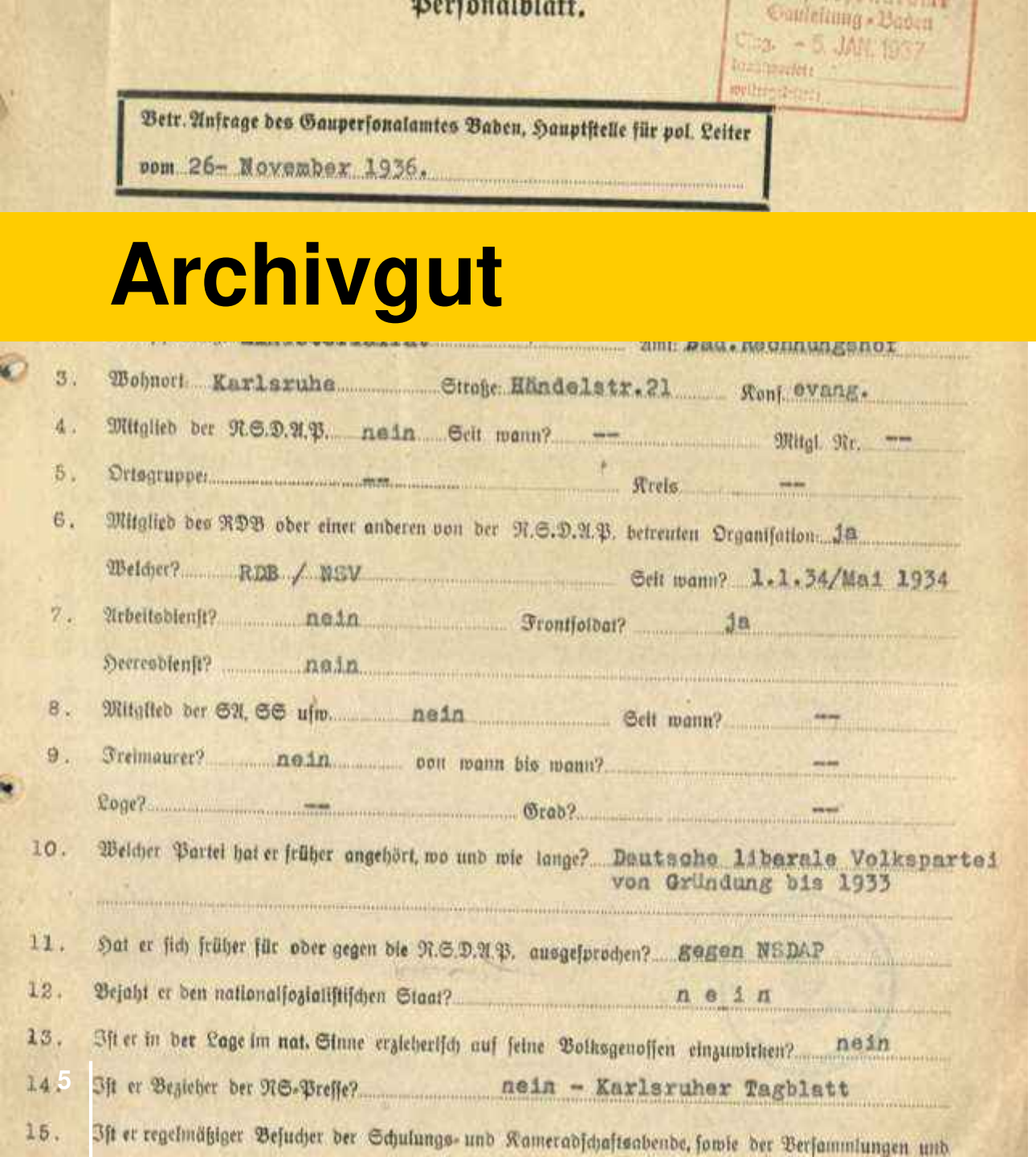
Blog mit Berichten des Projektes [2]



Landesarchiv Baden-Württemberg

1.

Was digitalisieren wir?


Personalblatt.

Gauleitung - Baden
Eing. - 5. JAN. 1937

Betr. Anfrage des Gaupersonalamtes Baden, Hauptstelle für pol. Leiter
vom 26. November 1936.

3. Wohnort Karlsruhe Straße Händelstr.21 Konf. evang.
4. Mitglied der N.S.D.A.P. nein Seit wann? -- Mitgl. Nr. --
5. Ortsgruppe: -- Kreis --
6. Mitglied des RDB oder einer anderen von der N.S.D.A.P. betreuten Organisation: ja
Welcher? RDB / NSV Seit wann? 1.1.34/Mai 1934
7. Arbeitsdienst? nein Frontsoldat? ja
Heeresdienst? nein
8. Mitglied der SA, SS usw. nein Seit wann? --
9. Freimaurer? nein von wann bis wann? --
Loge? -- Grad? --
10. Welcher Partei hat er früher angehört, wo und wie lange? Deutsche liberale Volkspartei von Gründung bis 1933
11. Hat er sich früher für oder gegen die N.S.D.A.P. ausgesprochen? gegen NSDAP
12. Bejaht er den nationalsozialistischen Staat? n e i n
13. Ist er in der Lage im nat. Sinne erzieherisch auf seine Volksgenossen einzuwirken? nein
14. Ist er Bezieher der NS-Presse? nein - Karlsruher Tagblatt
15. Ist er regelmäßiger Besucher der Schulungs- und Kameradschaftsabende, sowie der Versammlungen und

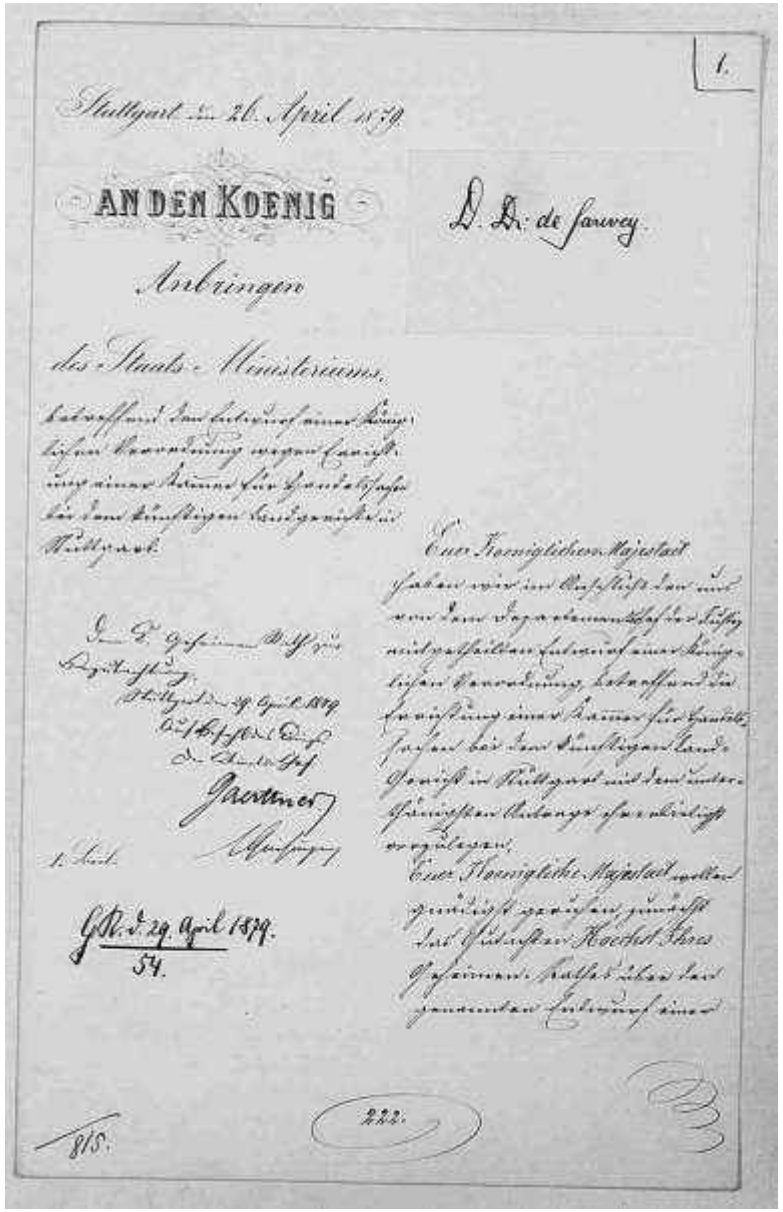
Archivgut

Landesarchiv Baden-Württemberg

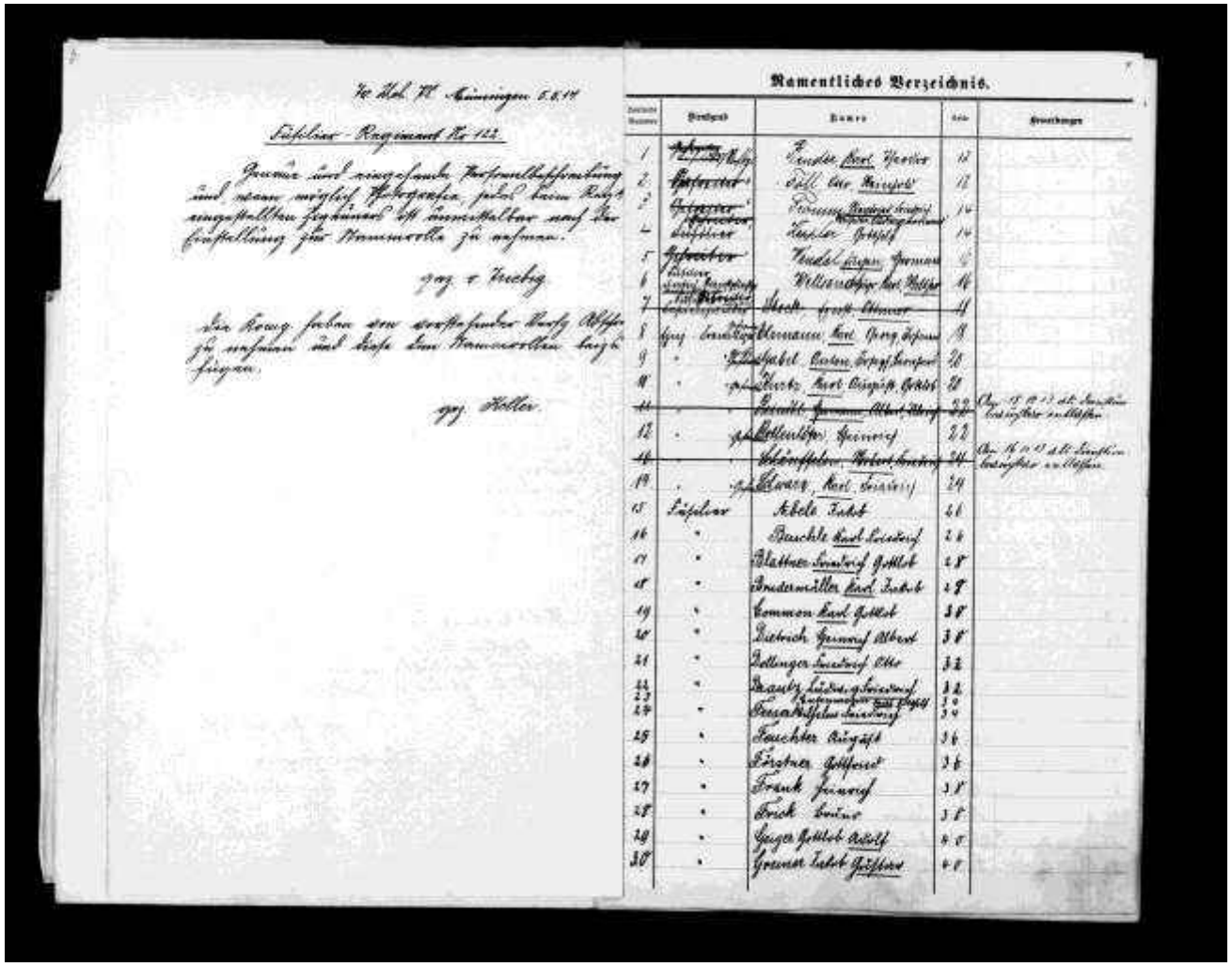
Merkmale unseres Archivguts:

- komplexe Layouts (Formulare, Tabellen)
- Mischung aus Hand-, Druck- und Schreibmaschinenschrift
- unleserliche Handschriften
- Akten mit vielen Dokumenttypen
- z. T. Abnutzungserscheinungen und Schäden am Papier
- Teilweise Mikrofilme als Grundlage

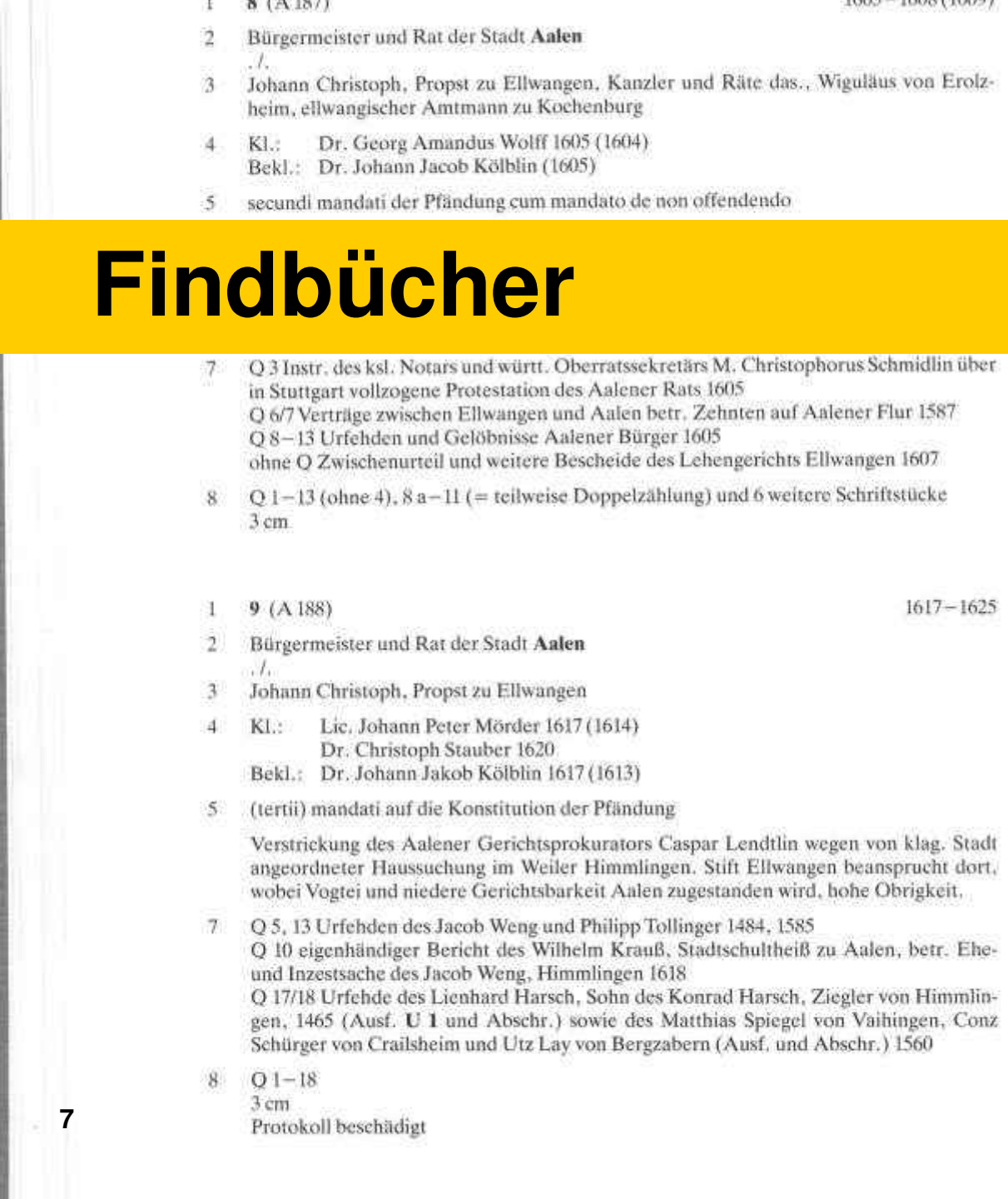
Quelle: Landesarchiv BW, GLAK 465 c Nr. 418, [3]



Landesarchiv BW, HStA E 33 Bü 354, Bild 24



Landesarchiv BW, HStA M 438 Nr. 122, Bild 6



1 **8** (A 187) 1605–1608 (1609)

2 Bürgermeister und Rat der Stadt **Aalen**
./.

3 Johann Christoph, Propst zu Ellwangen, Kanzler und Räte das., Wiguläus von Erolzheim, ellwangischer Amtmann zu Kochenburg

4 Kl.: Dr. Georg Amandus Wolff 1605 (1604)
Bekl.: Dr. Johann Jacob Kölblin (1605)

5 secundi mandati der Pfändung cum mandato de non offendendo

7 Q 3 Instr. des ksl. Notars und württ. Oberratssekretärs M. Christophorus Schmidlin über in Stuttgart vollzogene Protestation des Aalener Rats 1605
Q 6/7 Verträge zwischen Ellwangen und Aalen betr. Zehnten auf Aalener Flur 1587
Q 8–13 Urfehden und Gelöbnisse Aalener Bürger 1605
ohne Q Zwischenurteil und weitere Bescheide des Lehengerichts Ellwangen 1607

8 Q 1–13 (ohne 4), 8 a–11 (= teilweise Doppelzählung) und 6 weitere Schriftstücke
3 cm

1 **9** (A 188) 1617–1625

2 Bürgermeister und Rat der Stadt **Aalen**
./.

3 Johann Christoph, Propst zu Ellwangen

4 Kl.: Lic. Johann Peter Mörder 1617 (1614)
Dr. Christoph Stauber 1620
Bekl.: Dr. Johann Jakob Kölblin 1617 (1613)

5 (tertii) mandati auf die Konstitution der Pfändung

Verstrickung des Aalener Gerichtsprokurators Caspar Lendtlin wegen von klag. Stadt angeordneter Haussuchung im Weiler Himmlingen. Stift Ellwangen beansprucht dort, wobei Vogtei und niedere Gerichtsbarkeit Aalen zugestanden wird, hohe Obrigkeit.

7 Q 5, 13 Urfehden des Jacob Weng und Philipp Tollinger 1484, 1585
Q 10 eigenhändiger Bericht des Wilhelm Krauß, Stadtschultheiß zu Aalen, betr. Ehe- und Inzestsache des Jacob Weng, Himmlingen 1618
Q 17/18 Urfehde des Lienhard Harsch, Sohn des Konrad Harsch, Ziegler von Himmlingen, 1465 (Ausf. **U 1** und Abschr.) sowie des Matthias Spiegel von Vaihingen, Conz Schürger von Crailsheim und Utz Lay von Bergzabern (Ausf. und Abschr.) 1560

8 Q 1–18
3 cm
Protokoll beschädigt

Findbücher

Landesarchiv Baden-Württemberg

Neben Archivgut werden auch Findbücher digitalisiert, um die Erschließungsdaten über das Online-Findmittelsystem (OLF) des LABW recherchierbar zu machen.

2.

Welche Tools setzen wir ein?

Texterkennung für Druck- und Handschrift

Landesarchiv Baden-Württemberg

OCR

= *Optical Character Recognition*

– Automatische Texterkennung von **gedruckten** Texten

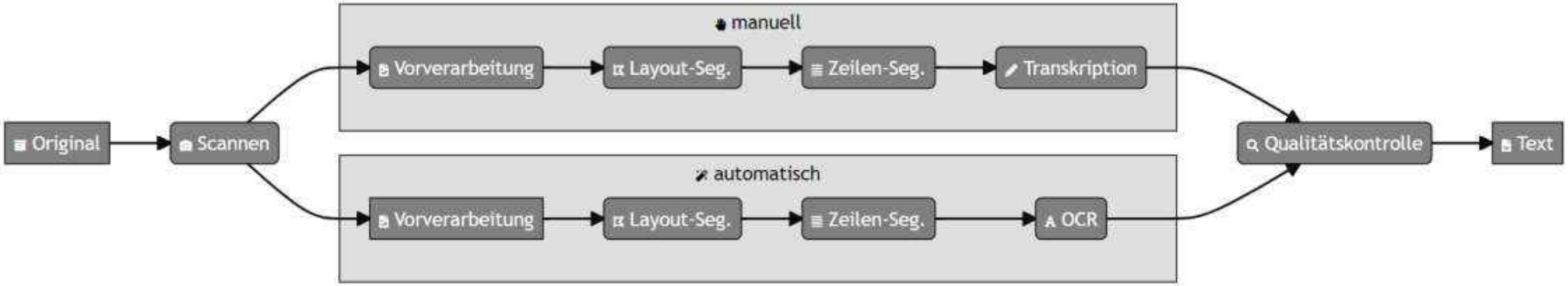
HTR

= *Handwritten Text Recognition*

– Automatische Texterkennung von **handschriftlichen** Texten

Tools und Workflow

Landesarchiv Baden-Württemberg



manuell

- OCR4All [4]
- Transkribus [5]

automatisch

- Abbyy FineReader
- OCR-D [6]
- Transkribus [5]

Qualitätskontrolle

- dinglehopper (Qurator) [7]

Landesarchiv Baden-Württemberg

3.

Kostenpflichtige Tools für OCR/HTR

OCR/HTR-Tools im Vergleich (bezahlt)

Landesarchiv Baden-Württemberg

	Abbyy FineReader	Aletheia	Transkribus
Features	• Scannen von Dokumenten • Layoutanalyse • Texterkennung	• Vorverarbeitung • Layoutanalyse • Texterkennung • Ground Truth • Textannotation	• Layoutanalyse • Texterkennung • Ground Truth • Modelltraining • Textannotation
Bedienung	Native Oberfläche	Native Oberfläche + Webbrowser	Native Oberfläche + Webbrowser
Use-Case	Automatische OCR	Manuelle/automatische OCR	Manuelle/automatische **HTR**
Sonstiges	Desktop-Version oder Server-Version.	Experimenteller Support für Schreibmaschinenschrift, nutzt Tesseract. Benötigt Java.	Startguthaben. Desktop-Version (Java) oder Web-Version.
Link	https://pdf.abbyy.com/	https://www.primaresearch.org/tools/Aletheia	https://transkribus.eu/

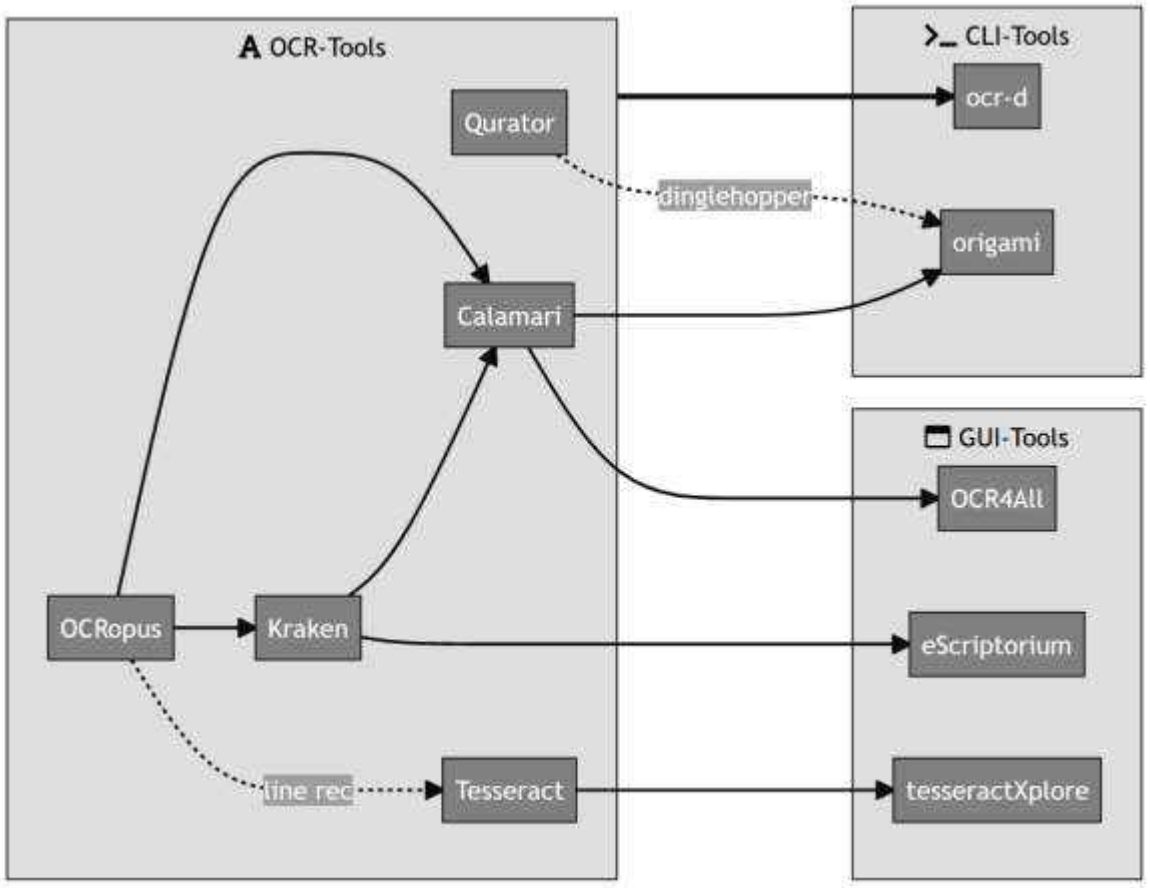
Landesarchiv Baden-Württemberg

4.

OpenSource Tools für OCR/HTR

OCR/HTR-Tools im Vergleich (OpenSource, CLI)

Landesarchiv Baden-Württemberg



	OCR-D	Origami
Features	• Vorverarbeitung • Layoutanalyse • Texterkennung	• Vorverarbeitung • Layoutanalyse • Texterkennung • Ground Truth • Modelltraining
Use-Case	Automatische Massenproduktion (OCR)	OCR für historische Zeitungen
Sonstiges	Integration vieler Tools (Tesseract, Calamari, ...)	Nutzt Calamari
Link	https://ocr-d.de	https://github.com/poke1024/origami

OCR/HTR-Tools im Vergleich (OpenSource, GUI)

Landesarchiv Baden-Württemberg

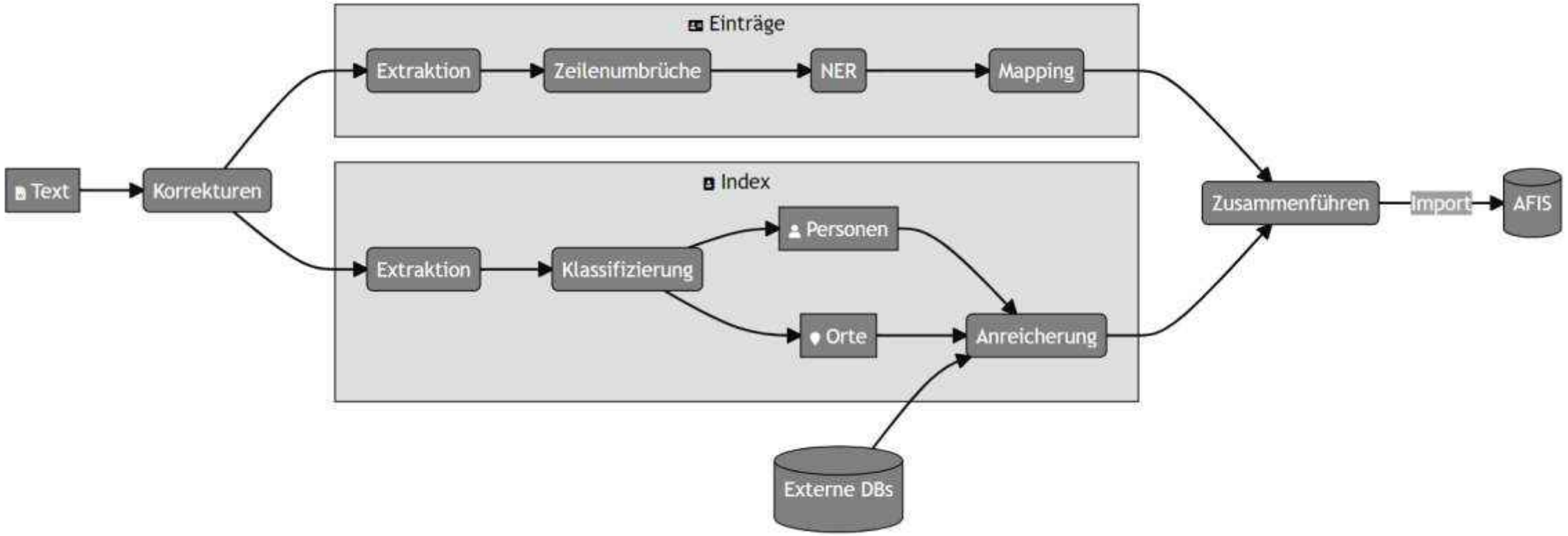
	OCR4All	eScriptorium	tesseractXplore
Features	• Vorverarbeitung • Layoutanalyse • Texterkennung • Ground Truth • Modelltraining	• Vorverarbeitung • Layoutanalyse • Texterkennung • Ground Truth • Modelltraining • Textannotation	• Vorverarbeitung • Layoutanalyse • Texterkennung
Bedienung	Webbrowser	Webbrowser	Native Oberfläche
Use-Case	Manuelle/automatische OCR	Manuelle/automatische OCR+**HTR** im Mehrbenutzermodus	Automatische OCR
Sonstiges	Nutzt Calamari. Benötigt VirtualBox oder Docker.	**Work in progress!** Nutzt Kraken. Benötigt Docker oder Linux. Infos teilweise nur auf Französisch verfügbar.	**Work in progress!** Nutzt Tesseract. Möglichkeit OCR an Server zu delegieren.
Link	http://www.ocr4all.org	https://scripta.psl.eu/en/digital-component-of-scripta/	https://github.com/JKamlah/tesseractXplore

5.

Praxisbeispiel: Digitalisierung von Findbüchern

Workflow Digitalisierung Findbücher

Landesarchiv Baden-Württemberg



Abgeschlossen: Findbuch zu Urkunden der Grafschaft Friedberg-Scheer (ca. 1 500 Urkunden) [8]

In Arbeit: Findbücher zu Akten des Reichskammergerichts (ca. 10 000 verzeichnete Einheiten)

Lessons Learned

Landesarchiv Baden-Württemberg

- Manuelle Nacharbeiten: Unregelmäßigkeiten bei Struktur, Layout und Sprache
- Analoges Findbuch stets zur Hand haben: Fehler in digitaler Version ausbessern
- Named Entity Recognition nicht ausreichend
- Extraktion transparent gestalten z.B. via Zwischenformat zur Visualisierung
- Enge Zusammenarbeit von technischem und archivfachlichem Personal
  - Granularität der Datenextraktion gemeinsam bestimmen
  - Mapping der extrahierten Daten auf Erfassungsformular abstimmen

Landesarchiv Baden-Württemberg

Vielen Dank!

Benjamin Rosemann
Landesarchiv Baden-Württemberg
Zentrale Dienste
Projekt FDMLab@LABW
URL: https://fdmlab.landesarchiv-bw.de
Tel.: +49 711 335075-512
E-Mail: benjamin.rosemann@la-bw.de
https://orcid.org/0000-0002-0780-3979

Elisabeth Klindworth
Landesarchiv Baden-Württemberg
Archivischer Grundsatz
Projekt FDMLab@LABW
URL: https://fdmlab.landesarchiv-bw.de
Tel.: +49 711 335075-513
E-Mail: elisabeth.klindworth@la-bw.de
https://orcid.org/0000-0003-1848-5870

www.landesarchiv-bw.de

Links

Landesarchiv Baden-Württemberg

[1] Projektseite FDMLab@LABW: https://www.landesarchiv-bw.de/de/landesarchiv/projekte/fdmlab%2540labw-/71653

[2] Blog FDMLab: https://fdmlab.landesarchiv-bw.de

[3] NS Überlieferung in staatlichen Archiven: https://www.leo-bw.de/themenmodul/sudwestdeutsche-archivalienkunde/querschnittsartikel/ns-uberlieferung

[4] OCR4All: http://www.ocr4all.org/

[5] Transkribus: https://transkribus.eu/

[6] OCR-D: https://ocr-d.de/

[7] Dinglehopper: https://github.com/qurator-spk/dinglehopper/

[8] Findbuch Staatsarchiv Sigmaringen Dep. 30/1 T 1, Grafschaft Friedberg-Scheer: https://www2.landesarchiv-bw.de/ofs21/olf/struktur.php?bestand=2240

[9] OCRopus: https://github.com/ocropus/

[11] Kraken: http://kraken.re/

[12] Calamari: https://github.com/Calamari-OCR/

[13] Tesseract: https://tesseract-ocr.github.io/

[14] Qurator: https://qurator.ai/

[15] Flexible character accuracy: https://doi.org/10.1016/j.patrec.2020.02.003