

KI im Archiv

Konferenz "Offene Archive" am 13.06.2022
Elisabeth Klindworth, Benjamin Rosemann

Was erwartet Sie in diesem Workshop?

01 Das FDMLab

02 Was kann KI?

03 Was ist KI?

04 Praktische
Beispiele aus
dem FDMLab

05 KI Infrastruktur?

06 Diskussion



Bild von [Gerd Altmann](#) auf [Pixabay](#)

01

Das FDMLab

Laufzeit von Juli 2020 bis Juli 2022

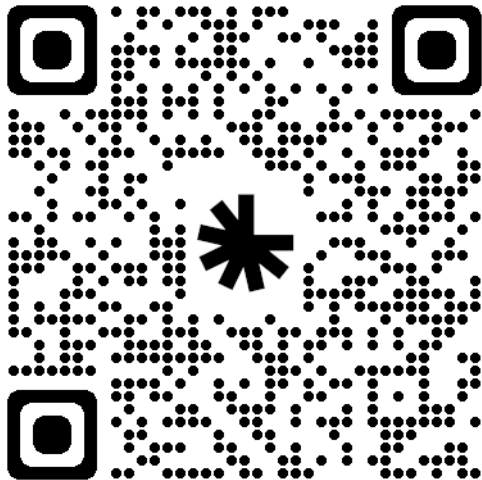
Kontext

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

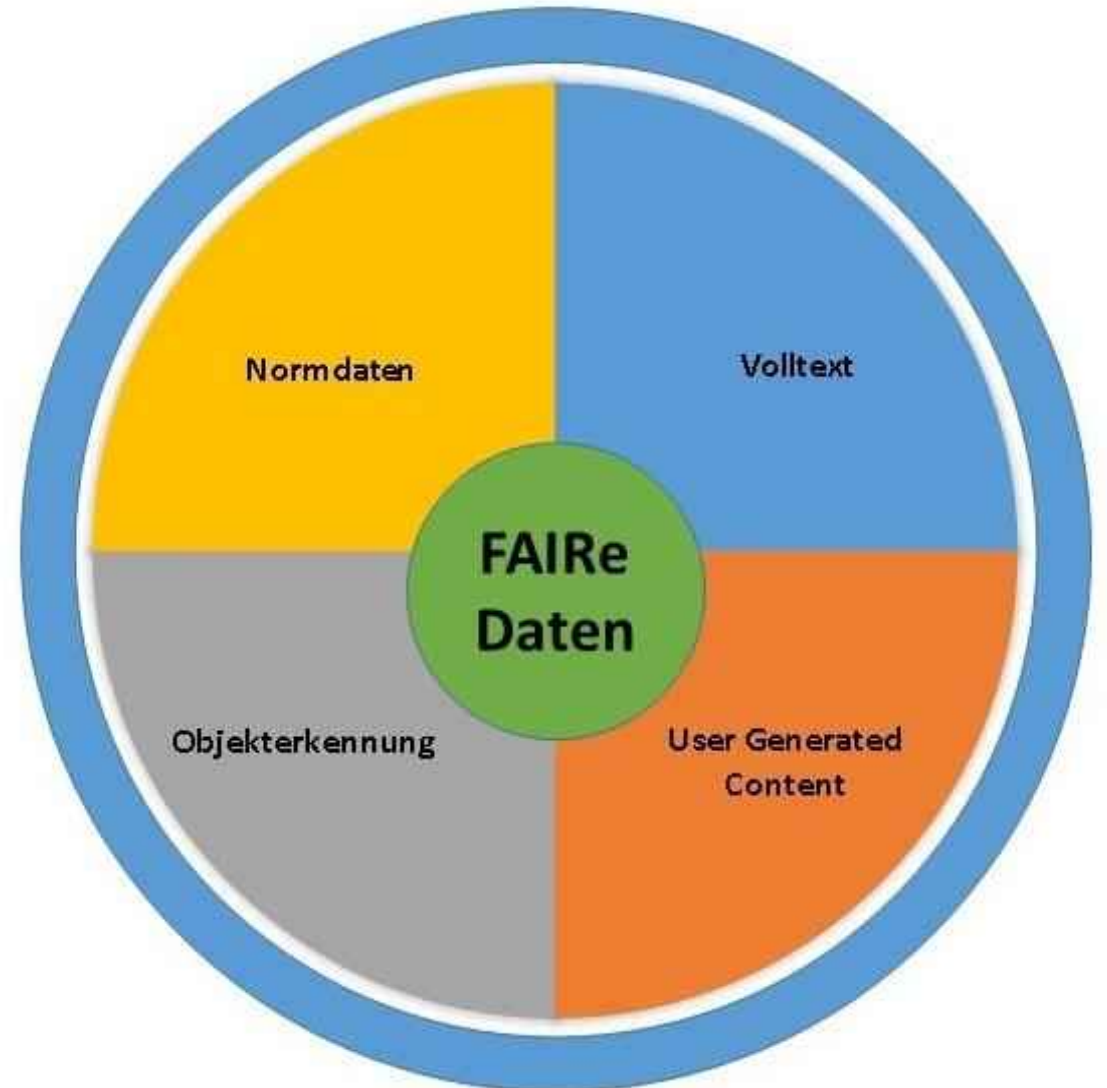
Projektziele

- 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

Weitere Informationen



<https://fdmlab.landesarchiv-bw.de/>
Blog mit Berichten des Projektes.



02

Was kann Künstliche Intelligenz?

KI kann...

- lesen, schreiben, malen, komponieren
- zuhören, sprechen, gehen, rennen
- Autos, LKW und Schiffe fahren
- Flugzeuge fliegen
- mehr als 100 Sprachen übersetzen
- Schach, Go und Jeopardy gewinnen
- Chemische Kampfstoffe entwickeln
- ...

03

Was ist Künstliche Intelligenz?

Was ist Künstliche Intelligenz?



Training



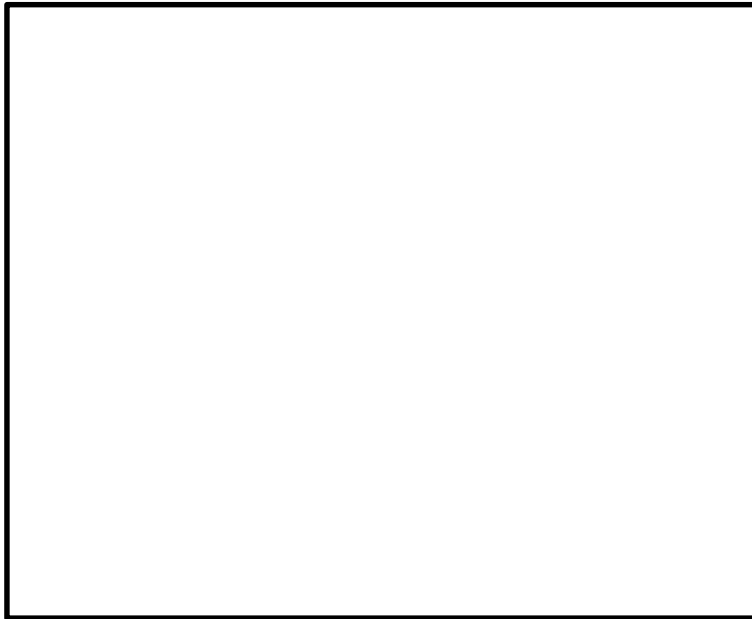
Was ist Künstliche Intelligenz?



Training



Leere Seite



KI

Sehr geehrte Damen und Herren,

Mit freundlichen Grüßen

Was ist Künstliche Intelligenz?



Training



Ethische und moralische Fragen (symbolisch)

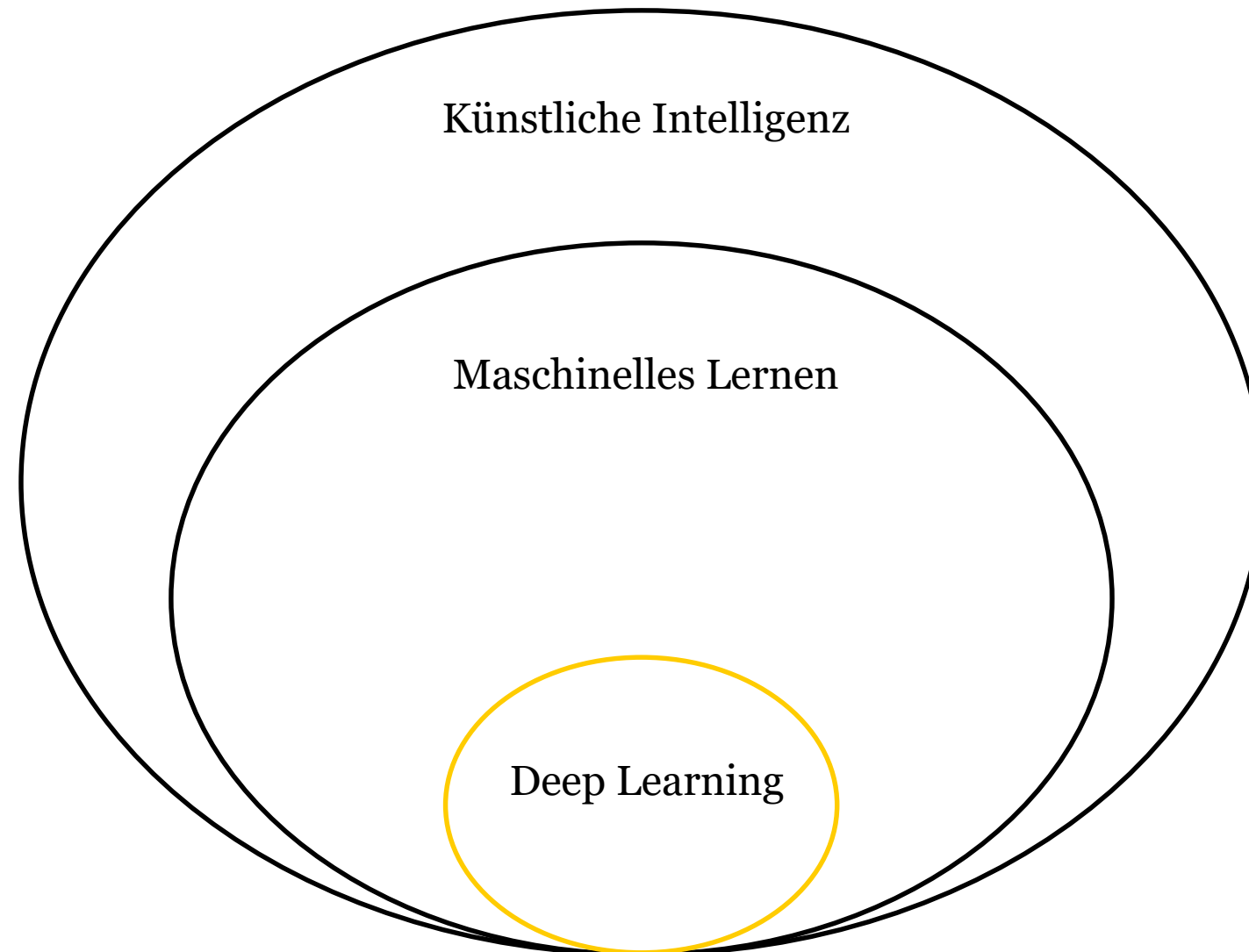


Was ist Künstliche Intelligenz?



Training





Natural
Language
Processing

Knowledge
Representation

Automated
Reasoning

Machine
Learning

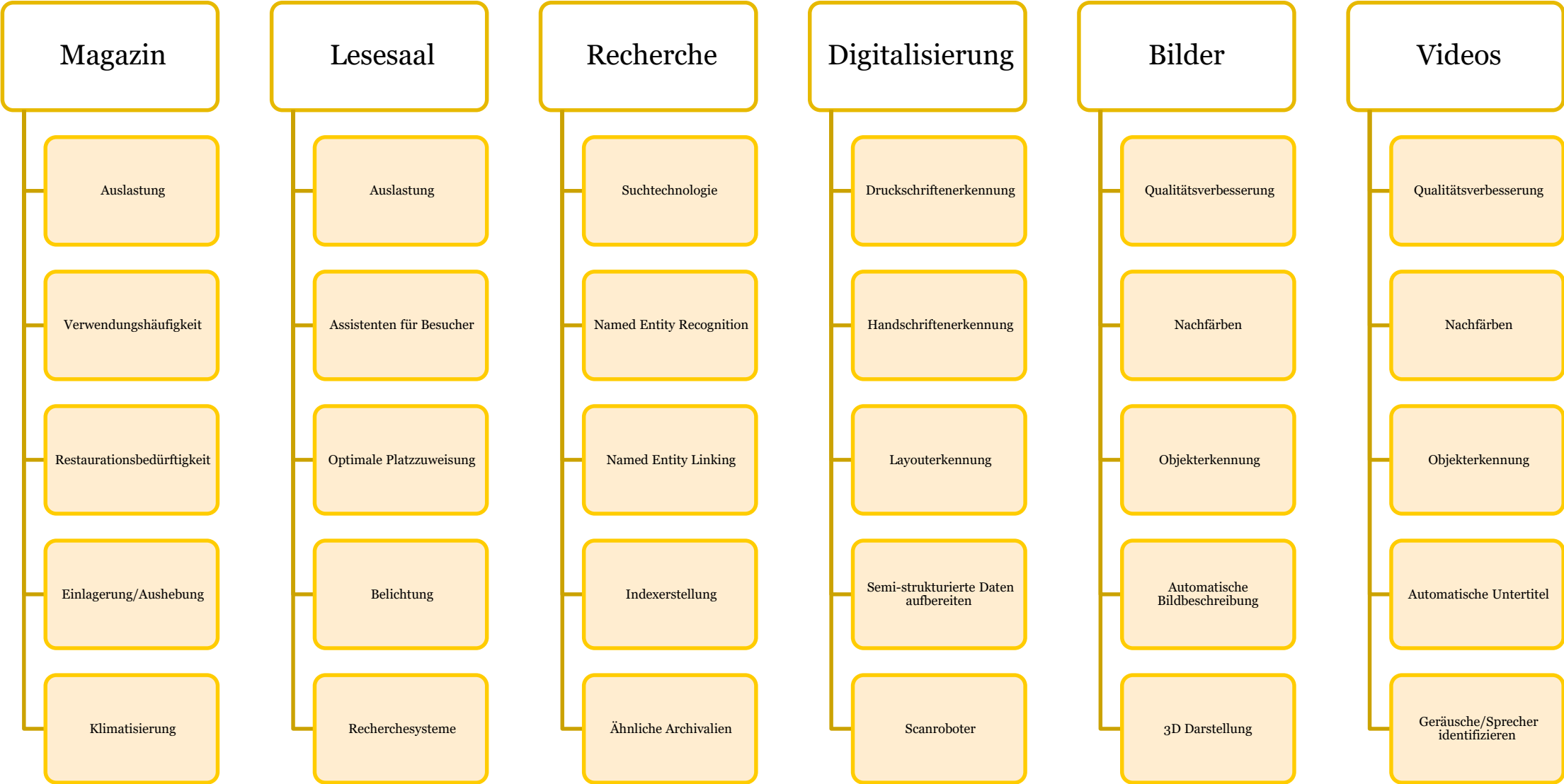
Computer
Vision

Robotic

Quellen:

- Stuart Russell und Peter Norvig, "What is AI?", in *Artificial Intelligence: A Modern Approach* (Upper Saddle River, N.J.: Prentice Hall, 2010), 2-3.

Ideen für KI im Archiv



04

Praktische Beispiele aus dem FDMLab

OCR

= *Optical Character Recognition*

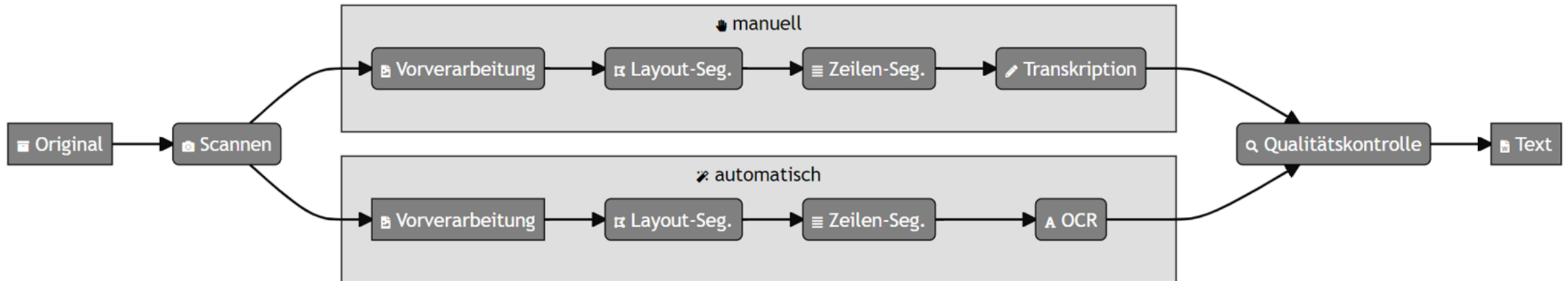
Automatische Texterkennung von
gedruckten Texten

HTR

= *Handwritten Text Recognition*

Automatische Texterkennung von
handschriftlichen Texten

Workflow der Texterkennung



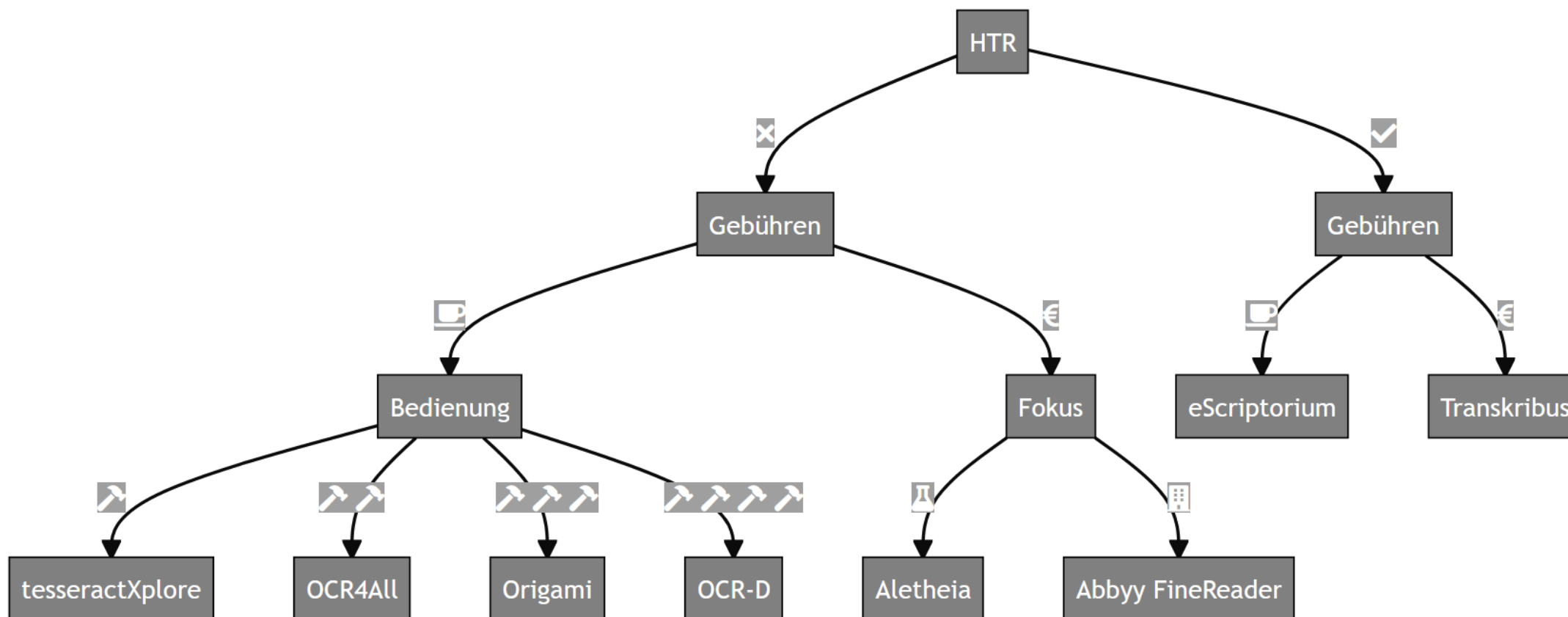
Transkribus – ein Tool für historische Handschriften

The screenshot displays the Transkribus Expert Client interface. The top window shows a document viewer with a handwritten table. The table has columns for names and dates, with handwritten entries in German. The bottom window shows a list of documents with columns for ID, Title, and Pages.

ID	Title	Pages
9065	GLAK_456C_Nr_0001	14
906382	HSA_M547_Bd_0001	10
906379	GLAK_456C_Nr_4967	9
906292	GLAK_456C_Nr_3809	14
902621	HSA_M535_Bd_0029	14
902620	HSA_M513_Bd_0003	14
902619	HSA_M479_Bd_0032	13
902618	HSA_M459_Bd_0006	14
902617	GLAK_456C_Nr_4283	15
902616	GLAK_456C_Nr_4252	14
902615	GLAK_456C_Nr_0093	14
902614	GLAK_456C_Nr_0060	13
902613	GLAK_456C_Nr_0030	14
902612	GLAK_456C_Nr_0003	14
902610	HSA_M549_Bd_0001	13
902608	HSA_M545_Bd_0005	14
902607	HSA_M543_Bd_0042	13
902606	HSA_M543_Bd_0016	13
902605	HSA_M528_Bd_0019	14
902602	HSA_M518_Bd_0001	14
902601	HSA_M510_Bd_0016	13
902598	HSA_M510_Bd_0014	13
902597	HSA_M470_Bd_0036	14
902596	HSA_M462_Bd_46	12
902595	HSA_M460_Bd_0002	16
902593	HSA_M438_Bd_0002	13
902592	HSA_M457_Bd_0005	14
902591	HSA_M457_Bd_0001	14
902590	GLAK_456C_Nr_3038	14

Below the document viewer, there is a list of documents with checkboxes and labels:

- 4-3 Christoph
- 4-4 Bockemann
- 5-1 Karl
- 5-2 Ferdinand
- 5-3 Piper
- 6-1 Adolf
- 6-2 Reinhold



Texterkennung: Wo stehen wir heute?

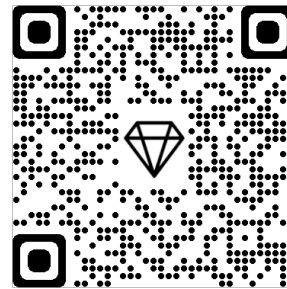


- Tools (kostenpflichtig & Open Source) für OCR und HTR sind verfügbar
 - z.T. ohne grafische Benutzeroberfläche
- Training eigener Erkennungsmodelle zur Erhöhung der Qualität der Ergebnisse unter erheblichem Ressourcenaufwand
- Technisch nicht vollständig gelöste Probleme:
 - Erkennung von Tabellenstrukturen
 - Qualitätskontrolle der Layoutsegmentierung



OpenRefine is like Excel on steroids!

Quelle unbekannt



<https://openrefine.org/>

OpenRefine Gurs

Facet / Filter

Undo / Redo 10 / 10

Refresh

Reset All

Remove All

Geschlecht (m/w)

change

17 choices Sort by: name count

Cluster

m 1758

maennlich 9

man 3

man 10

manlich 8

mannlich 9

mannlich 4

mannlich 22

w 1973

weblich 1

wei 2

Familienverhältnis

change

19 choices Sort by: name count

Cluster

alleinstehend 1

gesch. 1

geschieden 3

ledig 29

single 2

verehatet 1

verh. 6

verheirat 1

verheiratet 1575

verhieratet 1

verhiratet 1

Todesdatum (Jahr)

change reset

1,940 — 2,022

☒ Numeric

☐ Non-numeric

☐ Blank

☐ Error

3906

0

14

0

3986 matching rows (4000 total)

Extensions: Wikidata

Show as: rows records

Show: 1 5 10 25 50 100 250 500 1000 rows

« first

« previous

3 of 160 pages

next »

last »

All	Geschlecht (m/w)	akad. Grad/Titel	Familienverhältnis	Geburtsdatum (exakt)	Geburtsort	Geburtsort GND-ID	Religion	Todesdatum (Tag)	Todesdatum (Monat)	Todesdatum (Jahr)	Todesdatum (beschreibend)
51.	w			16.09.1869	Sankt Goar	4051602-7	israelitisch	1	9	1943	ermordet
52.	m			26.12.1878	Tauberbischofsheim	4059128-1	israelitisch	19	12	1943	verschollen
53.	m			31.03.1863	Schopfloch	4204796-1	israelitisch	8	8	1945	ermordet
54.	w			09.08.1855	Altdorf bei Ettenheim	4260757-7	israelitisch	8	5	1958	für tot erklärt, ermordet
55.	w			19.06.1892	Worms	4068942-7	israelitisch	19	8	1941	verschollen
56.	w		verheiratet	25-09-1885	Lichtenau	4690984-9	israelitisch	18	1	1955	
57.	m		verheiratet	30-03-1882	Schwegenheim	4235625-8	israelitisch	8	2	1941	
58.	m		verwitwet	16. August 1867	Stuttgart	4058282-6	 Stuttgart (4058282-6) Gebietskörperschaft oder Verwaltungseinheit		2	1945	
59.	m		verheiratet	15.07.1879	Königsbach	4132021-9			5	1945	ermordet
60.	m			29.11.1879	Karlsruhe	4029713-5			5	1940	ermordet
61.	m			30-05-1868	Schriesheim	4053296-3			11	1948	ermordet
62.	m		verheiratet	01.08.1884	Gemünden im Hunsrück	2001763-7	israelitisch	8	5	1942	verschollen
63.	w		verwitwet	12.12.1876	Berwangen	4280920-4	israelitisch	8	10	1941	ermordet
64.	m			06-08-1874	Richen	7830876-8	israelitisch	31	12	1942	ermordet
65.	w			15.10.1877	Sierck-les-Bains (Frankreich)	4054905-7	israelitisch	8	8	1945	ermordet
66.	w			01.08.1921	Tauberbischofsheim	4059128-1	israelitisch	8	5	1945	
67.	w		verehatet	08.04.1872	Dreibrunnen (Deutsches Reich heute Frankreich)	4817592-4	israelitisch	5	6	1945	für tot erklärt, ermordet
68.	w		verheiratet	12-03-1878	Kiel	4030481-4	israelitisch	22	1	1945	
69.	m			16-03-1872	Lichtenau	4690984-9	israelitisch	8	3	1942	verschollen
70.	m			05.11.1887	Bonfeld	4593548-8	israelitisch	8	5	1941	für tot erklärt, ermordet
71.	w		verheiratet	06-06-1879	Eschenau	4509954-6	israelitisch	28	9	1962	für tot erklärt, ermordet
72.	m			09-04-1867	Bendzin bei Pless (Russisches Kaiserreich heute Polen)	4559259-7	israelitisch	8	8	1945	für tot erklärt, ermordet
73.	m		verheiratet	19.10.1896	Reichenbach (Lautertal-Reichenbach)	7564415-0	israelitisch	8	1	1945	verschollen
74.	m			29.06.1878	Sinsheim	4055123-4	israelitisch	21	10	1956	für tot erklärt, ermordet
75.	m	Dr. Med.		26.04.1887	Bad Neustadt a. d. Saale	4041939-3	israelitisch	8	9	1952	für tot erklärt, ermordet

25

KI im Archiv 2022

Benjamin Rosemann, Elisabeth Klindworth

Clustering am Beispiel mit OpenRefine

Integrierte Algorithmen

- FP: Fingerprinting
- FP: N-Gram Fingerprinting
- Phon: Cologne-phonetic
- Phon: Methaphone3
- Phon: Daitsch-Moktoff
- Phon: Baider-Morse
- NN: Levensthein
- NN: Prediction by Partial Matching

Cluster & Edit column "Name_original"

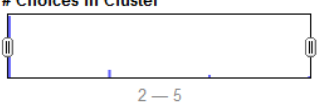
This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. [Find out more...](#)

Method: **key collision** Keying Function: **metaphone3** 113 clusters found

Cluster	Count	Members	Preview
4	10	- Steinhausen, Konrad Franz (von) (5 rows) - Steinhausen, Konrad Franz von (3 rows) - Steinhausen, Conrad Franz - Steinhausen, Conrad Franz von	Steinhausen, Konrad Franz (von)
4	10	- Hofmann, Caspar Friedrich (von) (4 rows) - Hofmann, Caspar Friedrich von (3 rows) - Hofmann, Caspar Friedrich (2 rows) - Hofmann, Kaspar Friedrich	Hofmann, Caspar Friedrich (von)
4	14	- Brandt, Johann Adolph (6 rows) - Brandt, Johann Adolph Georg (5 rows) - Brandt, Johann Adolph Georg von (2 rows) - Brandt, Johann Adolph Georg von	Brandt, Johann Adolph
4	7	- Brandt, (Johann) Ferdinand Wilhelm (2 rows) - Brandt, (Johann) Ferdinand Wilhelm von (2 rows) - Brandt, Johann Ferdinand Wilhelm (2 rows) - Brandt, (Johann) Ferdinand Wilhelm (von)	Brandt, (Johann) Ferdinand Wilhelm
3	5	Leweßau (Liebesau) von Rostock,	Leweßau (Liebesau) von Rostock,

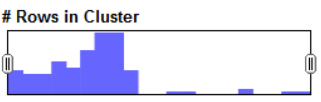
Select All Unselect All Export Clusters Merge Selected & Re-Cluster Merge Selected & Close Close

Choices in Cluster



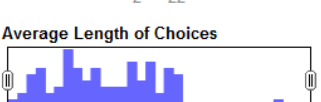
2 — 5

Rows in Cluster



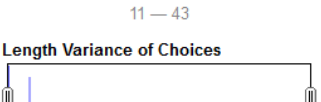
2 — 22

Average Length of Choices



11 — 43

Length Variance of Choices



0 — 7.26

Suche in GND via lobid

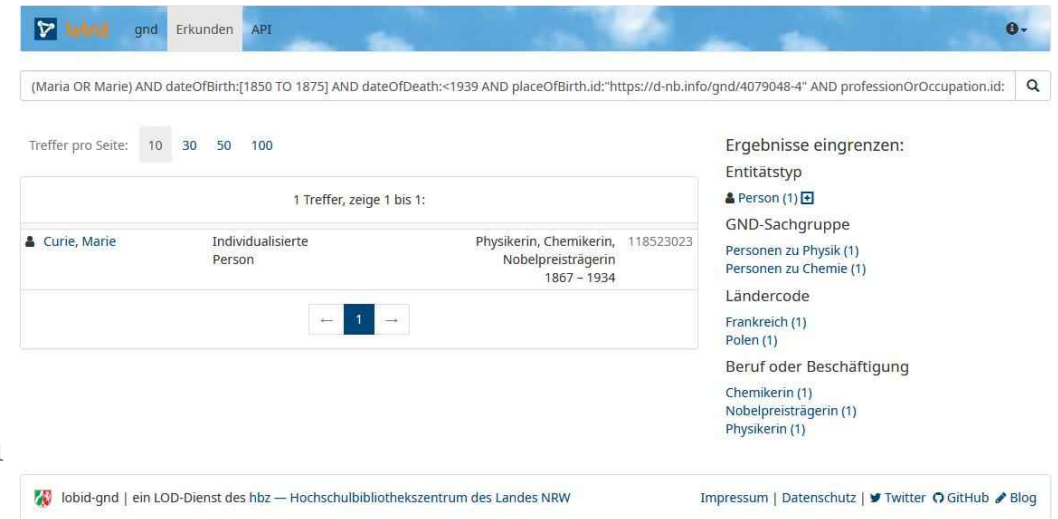
(Maria OR Marie) AND
dateOfBirth:[1850 TO 1875] AND
dateOfDeath:<1939 AND
placeOfBirth.id:

"https://d-nb.info/gnd/4079048-4" AND # Warschau
professionOrOccupation.id:(

"https://d-nb.info/gnd/4219058-7" OR # Chemikerin

"https://d-nb.info/gnd/4337175-9" OR # Physikerin

"https://d-nb.info/gnd/4202451-1") # Naturwissenschaftlerin



The screenshot shows the search results for the query "(Maria OR Marie) AND dateOfBirth:[1850 TO 1875] AND dateOfDeath:<1939 AND placeOfBirth.id:'https://d-nb.info/gnd/4079048-4' AND professionOrOccupation.id:'https://d-nb.info/gnd/4219058-7' OR 'https://d-nb.info/gnd/4337175-9' OR 'https://d-nb.info/gnd/4202451-1'". The results are displayed in a table with one entry: "Curie, Marie". The table shows the individual's name, their professions (Physikerin, Chemikerin, Nobelpreisträgerin), and their birth and death dates (1867 – 1934). The results are filtered by "Person (1)" and "GND-Sachgruppe". The "Ergebnisse eingrenzen:" section on the right shows the following filters: "Entitätstyp" (Person (1)), "GND-Sachgruppe" (Personen zu Physik (1), Personen zu Chemie (1)), "Ländercode" (Frankreich (1), Polen (1)), and "Beruf oder Beschäftigung" (Chemikerin (1), Nobelpreisträgerin (1), Physikerin (1)).

Bildschirmfoto von

<https://lobid.org/gnd/search?q=%28Maria+OR+Marie%29+AND+dateOfBirth%3A%5B1850+TO+1875%5D+AND+dateOfDeath%3A%3C1939+AND+placeOfBirth.id%3A%22https%3A%2F%2Fd-nb.info%2Fgnd%2F4079048-4%22+AND+professionOrOccupation.id%3A%28%22https%3A%2F%2Fd-nb.info%2Fgnd%2F4219058-7%22+OR+%22https%3A%2F%2Fd-nb.info%2Fgnd%2F4337175-9%22+OR+%22https%3A%2F%2Fd-nb.info%2Fgnd%2F4202451-1%22%29> (aufgerufen am 27. Mai 2022).

Suchtechnologien am Beispiel GND

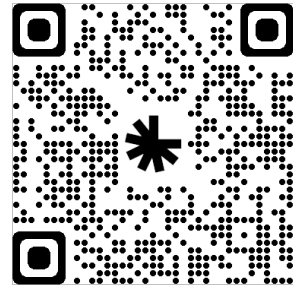
Ungenauere Namen mit OpenRefine abgleichen



[https://fdmlab.landesarchiv-bw.de/
workshop/openrefine-fortgeschrittene/
15-erweiterter-gnd-abgleich-mit-lobid/](https://fdmlab.landesarchiv-bw.de/workshop/openrefine-fortgeschrittene/15-erweiterter-gnd-abgleich-mit-lobid/)

Name	Geb.datum	Geburtsort
Winfried Kretschmann	1948-05-17	Spaichingen
Winfried Hemmann	1952	Rottenburg am Neckar
Manfred Lucka	1961-03-13	Garching a. d. Alz
Thekla Wallcer	1969	Dülmen
Tleresla Bauen	1965-04-06	Zweibrücken
Tlenesa Schoqger	1961	Füssen
Danvai Bayas	1983	Heidelberg
Tlnomas Strolai	1960-03-17	Heilbronn
Nicoie Hollmeister-Kraul	1972	Balingen
Peler Haula	1960-12-24	Walldürn
Marlon Genlgez	1971	Haslach im Kinzigtal
Nicoie Rasayi	1965-05-20	Hongkong
Baaloara Bosch	1958-09-05	Stuttgart

OpenRefine Workshops mit Fokus aufs Archiv



<https://fdmlab.landesarchiv-bw.de/workshops/>



Workshop OpenRefine

Named-entity recognition: Englisch

The German Federal Archives or **Bundesarchiv ORG** (**BArch ORG**) are **the National Archives of Germany ORG** .
They are subordinated to the Federal Commissioner for Culture and the Media (**BKM ORG**).
The **Bundesarchiv ORG** was established at the current location in **Koblenz GPE** after **World War II EVENT** .
The current President of the **Bundesarchiv ORG** has been **Michael Hollmann PERSON** since **May 3rd, 2011 DATE** .
It employs **around 2,200 CARDINAL** people at **23 CARDINAL** locations.

Das Model *en_core_web_lg* hat einen **f-score** von **0,85** und unterscheidet **18** Arten von **Entitäten**:
CARDINAL, DATE, EVENT, FAC, GPE, LANGUAGE, LAW, LOC, MONEY, NORP, ORDINAL, ORG,
PERCENT, PERSON, PRODUCT, QUANTITY, TIME und WORK_OF_ART.

Quellen:

- Erstellt mit spaCy 3.3 *en_core_web_lg* (https://spacy.io/models/en#en_core_web_lg).
- Wikipedia contributors, "German Federal Archives", *Wikipedia, The Free Encyclopedia*, https://en.wikipedia.org/w/index.php?title=German_Federal_Archives&oldid=1069206774 (accessed May 27, 2022).
- Wikipedia Autoren, „Bundesarchiv (Deutschland)“, *Wikipedia – Die freie Enzyklopädie*, [https://de.wikipedia.org/w/index.php?title=Bundesarchiv_\(Deutschland\)&oldid=222166628](https://de.wikipedia.org/w/index.php?title=Bundesarchiv_(Deutschland)&oldid=222166628) (aufgerufen am 27. Mai 2022).

Named-entity recognition: Deutsch

Das Bundesarchiv (BArch) ist eine Bundesoberbehörde der **Bundesrepublik Deutschland. LOC**

Es ist der Beauftragten der Bundesregierung für Kultur und Medien (**BKM ORG**) unterstellt.

Das Bundesarchiv wurde nach dem **Zweiten Weltkrieg MISC** an seinem heutigen Standort in **Koblenz LOC** gegründet.

Der amtierende Präsident des **Bundesarchives ORG** ist seit dem 03. Mai 2011 **Michael Hollmann. PER**

Es beschäftigt circa 2.200 Mitarbeiter an 23 Standorten.

Das Model *de_core_news_lg* hat einen **f-score** von **0,85** und unterscheidet **4** Arten von **Entitäten**: LOC, MISC, ORG, PER.

Quellen:

- Erstellt mit spaCy 3.3 *de_core_news_lg* (https://spacy.io/models/de#de_core_news_lg).
- Wikipedia Autoren, „Bundesarchiv (Deutschland)“, Wikipedia – Die freie Enzyklopädie, [https://de.wikipedia.org/w/index.php?title=Bundesarchiv_\(Deutschland\)&oldid=222166628](https://de.wikipedia.org/w/index.php?title=Bundesarchiv_(Deutschland)&oldid=222166628) (aufgerufen am 27. Mai 2022).

Named-entity recognition: Deutsch im Archiv

Besitzhinweis auf Mobiliar (hauptsächlich gekauft bei der **Möbelfirma LOC** **Winfried Mueller, Stuttgart PER**),

1 vollständigen Silberkasten und mindestens 4 Silber-Becher.

Zahlung von 11100 fl. samt rückständigen neunjährigen Zinsen in Höhe von 4999 fl. oder Einsetzung in das adelige Gut Werenwag.

Gen. Forderung stammt aus Streit über Erbe des im November 1595 verst. K[C]aspar (von) **Laubenberg. MISC**

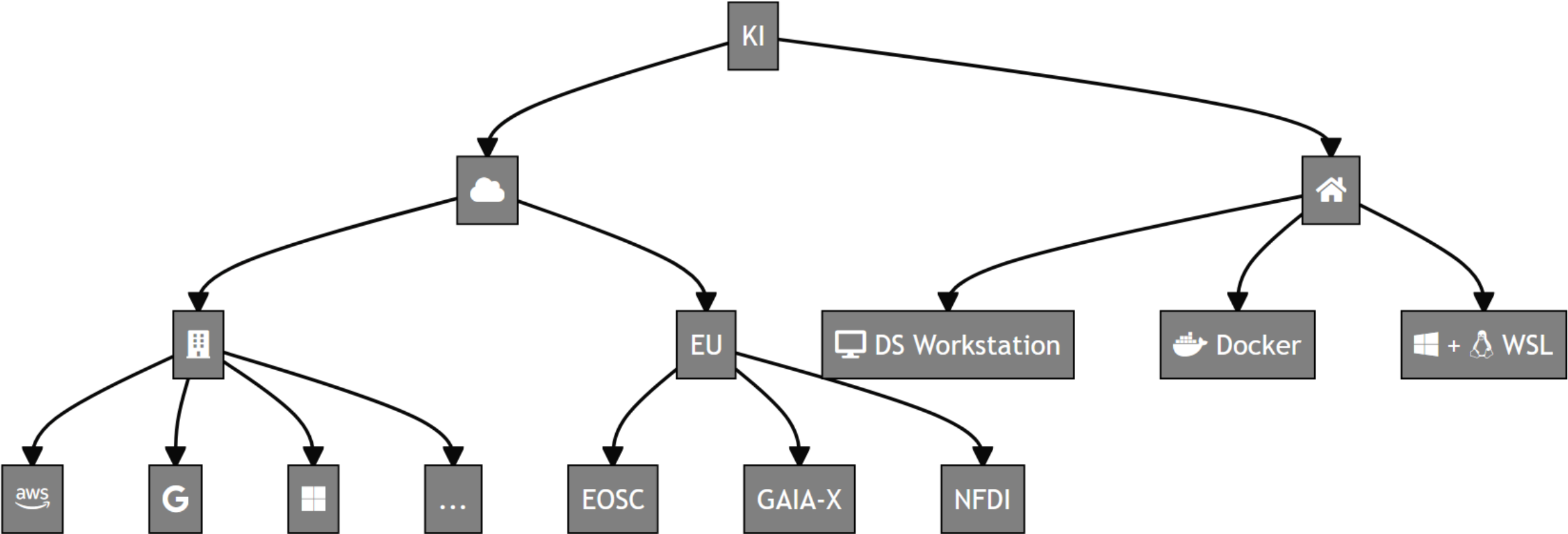
Entzug des ordentlichen Richters und Verstoß gegen Haigerlocher **Evokationsprivileg PER** durch Verschleppung

kläg. Sohns gen. "junger Lenz" nach Sigmaringen, wo gegen ihn **Kriminalprozeß PER** angestrengt wurde.

05

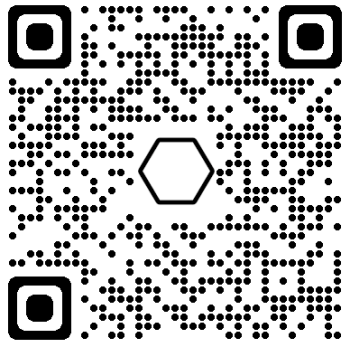
KI-Infrastruktur für Archive?

Möglichkeiten KI Anwendungen zu verwenden

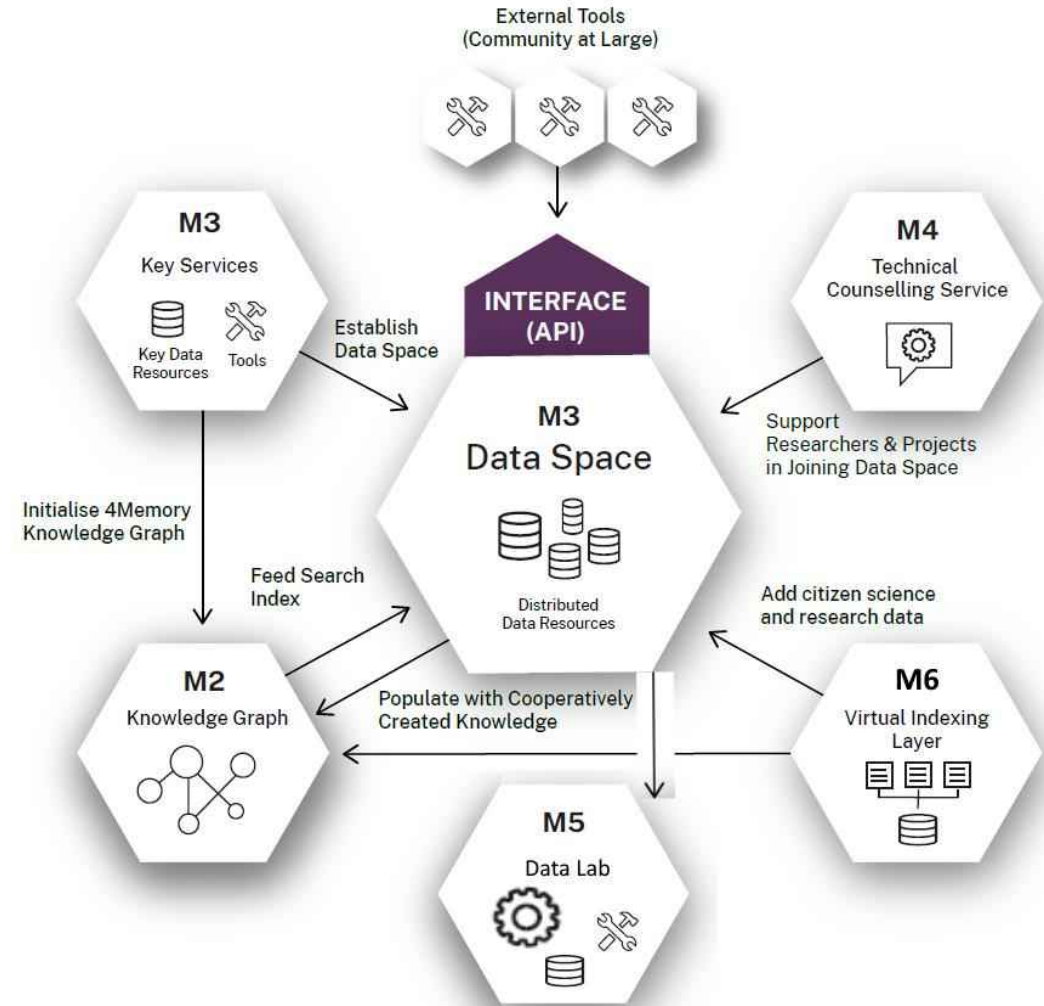


Task Area 3: Data Services

- **M5: Data Lab**
- High Performance Cluster
- Passende (online) Werkzeuge



<https://4memory.de/>



Quellen:

- Daniel Fähle, Harald Sack, "Perspektive „digitaler Werkzeugkasten“ für historische Forschung mit Archivgut" (Präsentation), 53. Deutscher Historikertag, München, 06. Oktober 2021.

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>