

KI im Archiv? Aber wie!

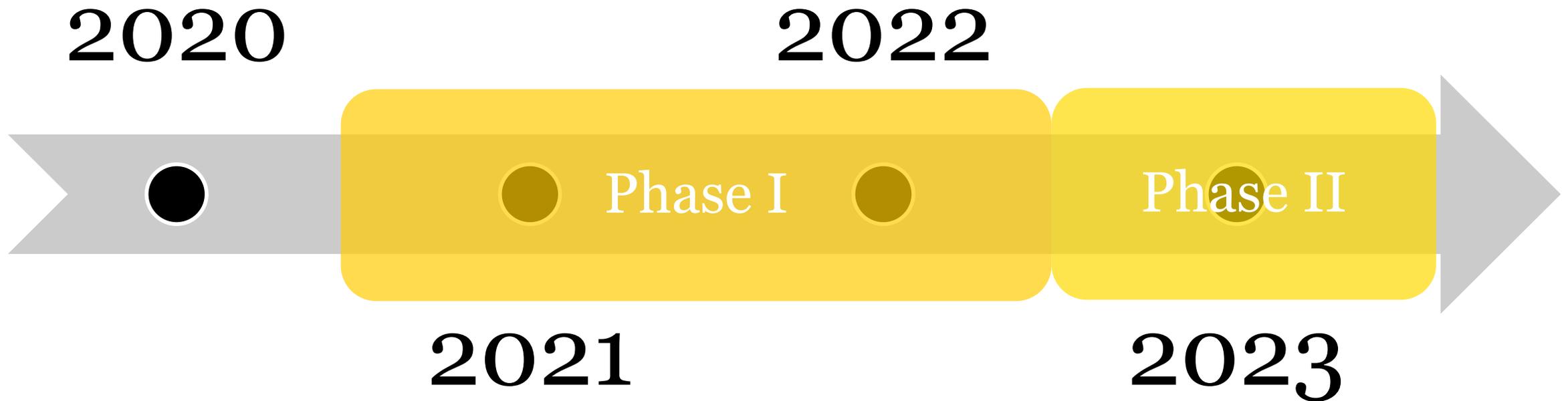
Werkzeuge für die Erschließung
und darüber hinaus

EDV-Tage 2022 am 29.09.2022
Benjamin Rosemann

01

Das FDMLab

Laufzeit des FDMLab@LABW



Förderung durch die **Baden-Württemberg Stiftung** im Rahmen der **Zukunftsoffensive III**.

Vorgehen im FDMLab@LABW



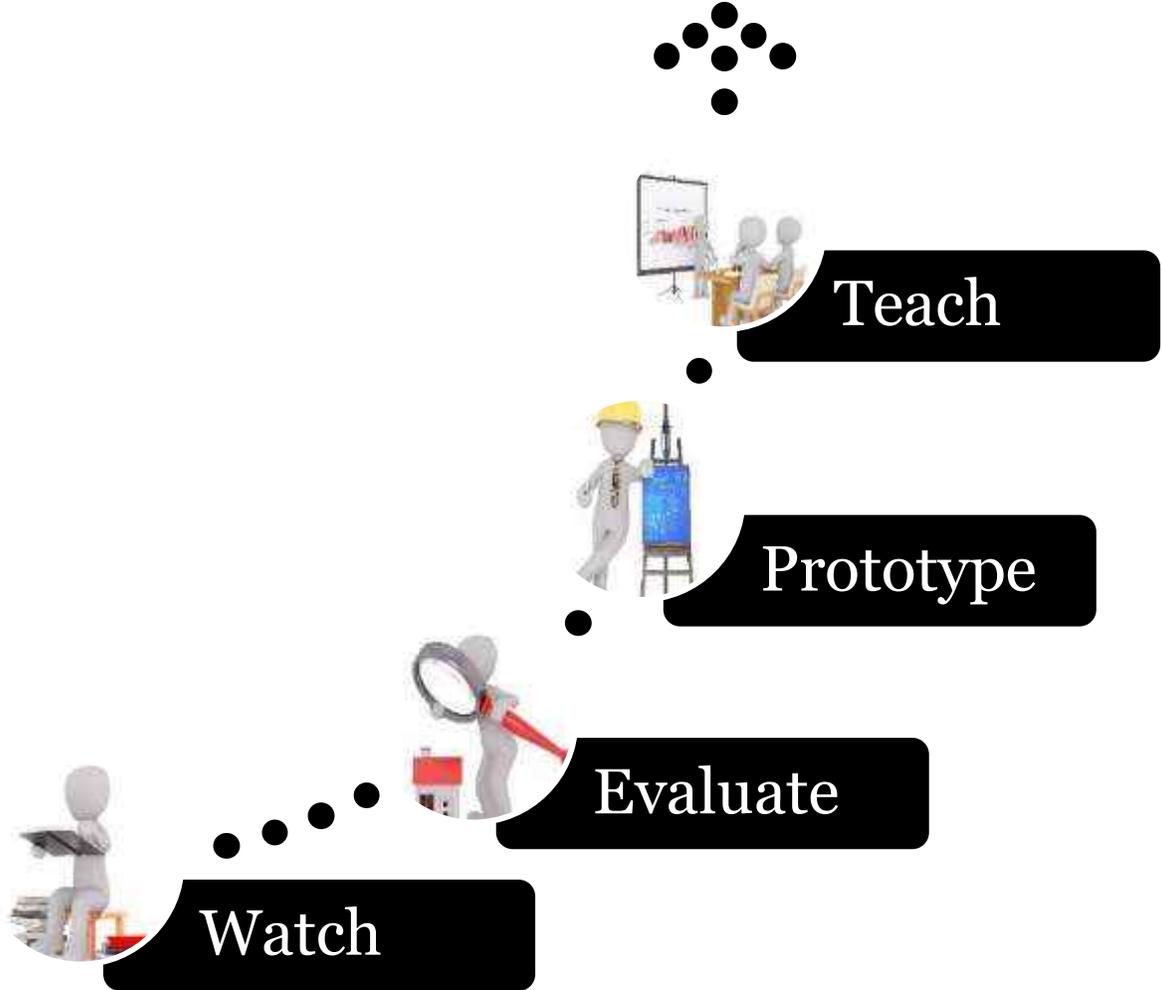
Quantity



Quality



Questions



02

Was ist Künstliche Intelligenz?

Die Rechenmaschine



Es ist Sonntag Nachmittag und es regnet. Laut Werbeprospekt kosten 1kg Äpfel im Supermarkt 2 Euro.

Ich möchte 5kg Äpfel. Wie viel Geld benötige ich zum Einkaufen?

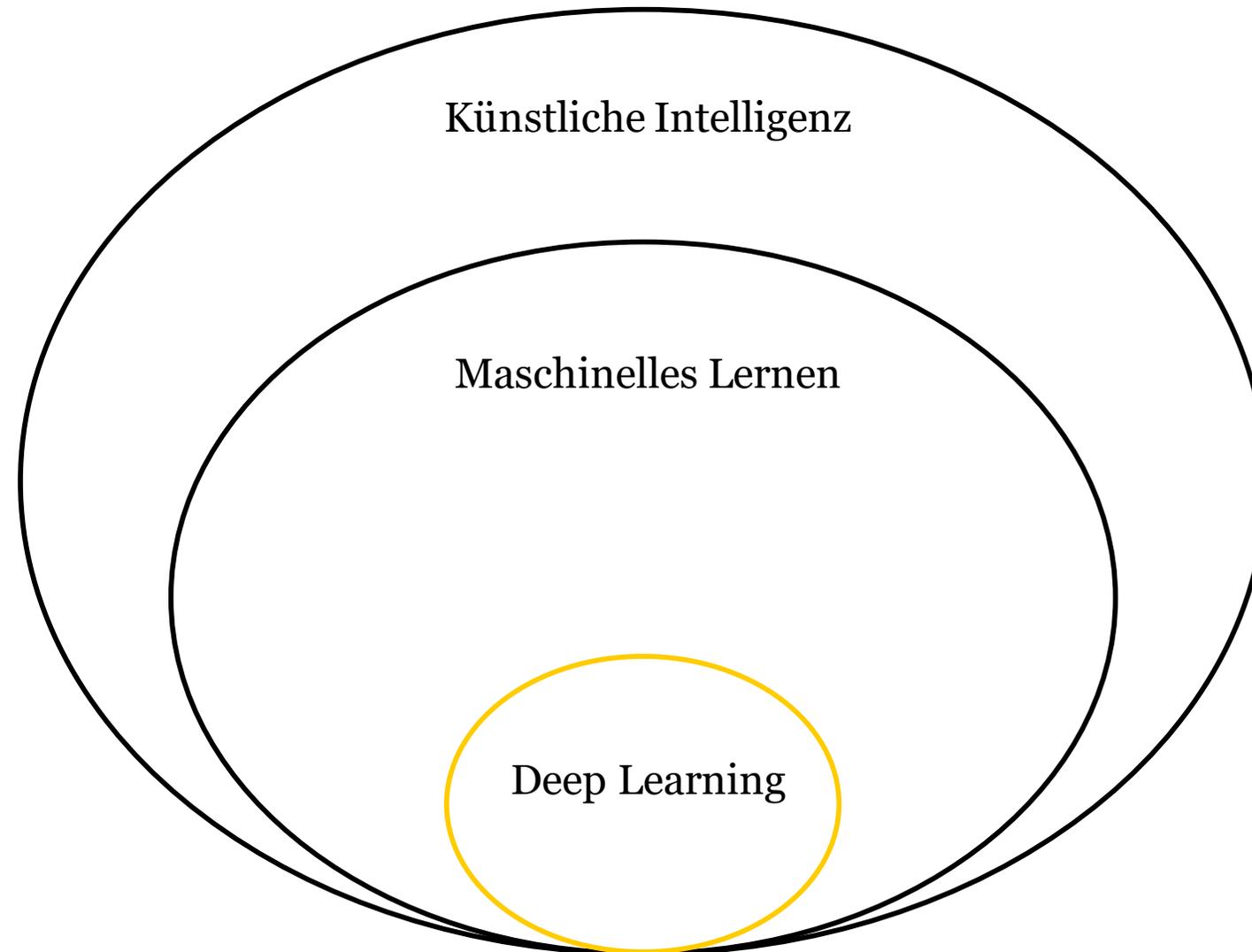
Was ist Künstliche Intelligenz?



Training



KI und Teilbereiche I



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.

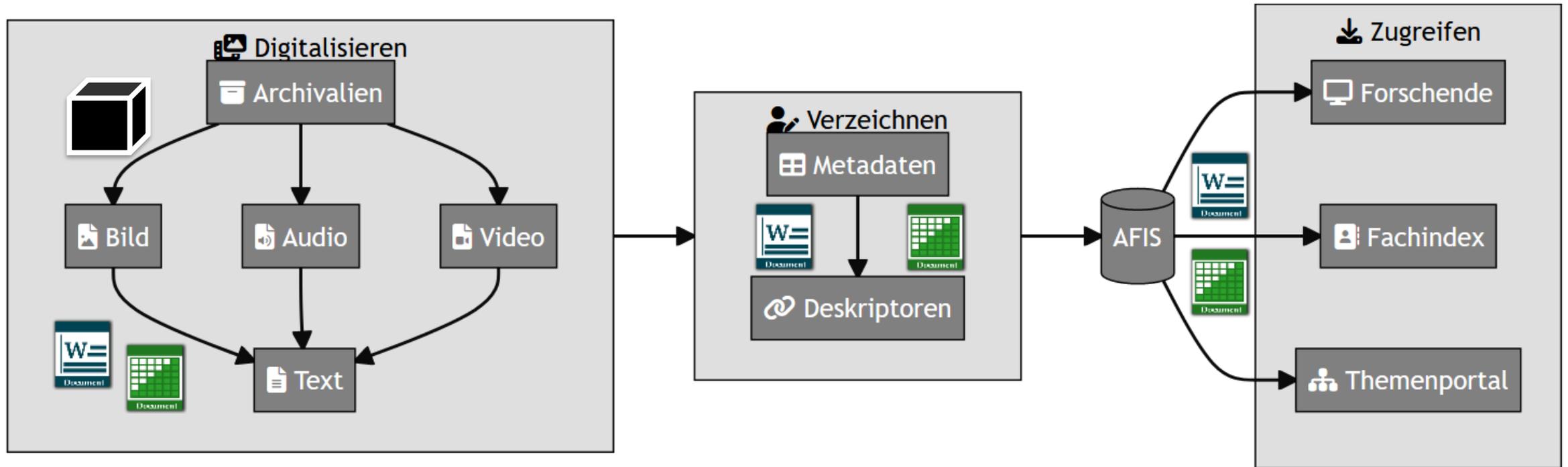
5 Zutaten für KI- und Daten-Projekte



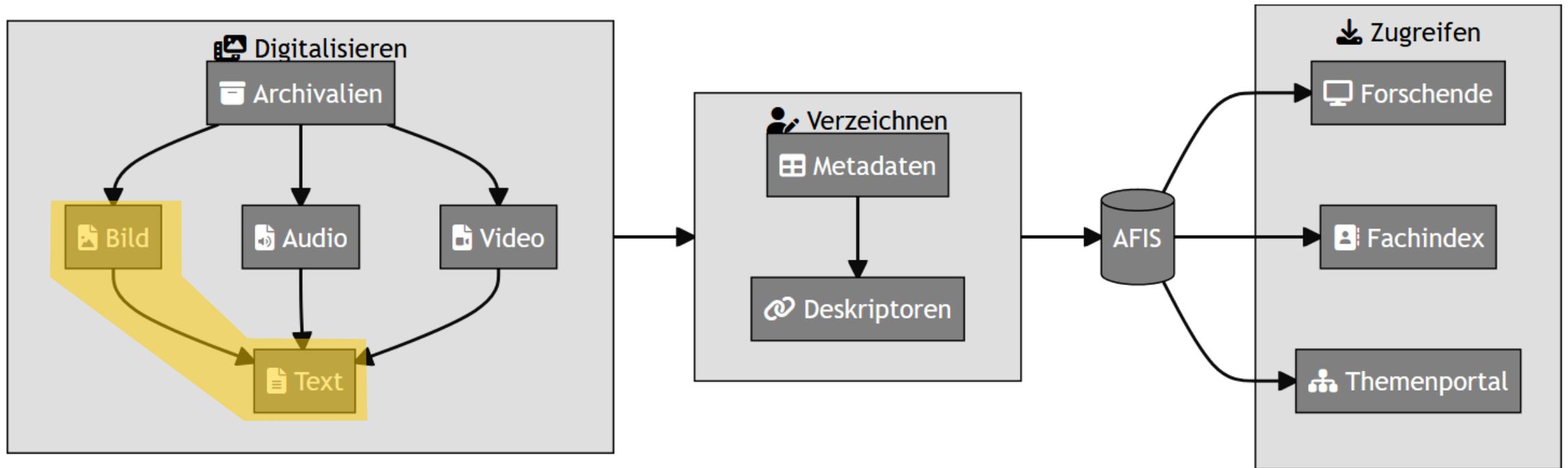
03

Praktische Beispiele aus dem FDMLab

Workflow Digitalisierung



Workflow Digitalisierung



Volltexte erstellen mit KI Tools



OCR-D

<https://ocr-d.de/>



Transkribus

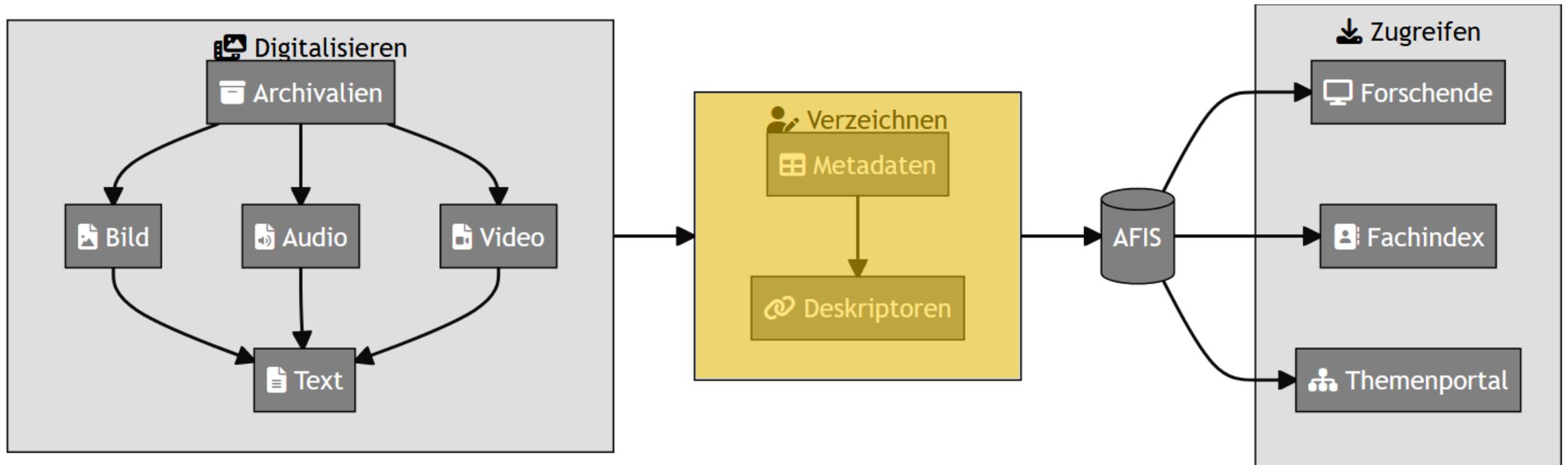
<https://transkribus.eu/>



eScriptorium

<https://gitlab.com/scripta/escriptorium>

Workflow Digitalisierung



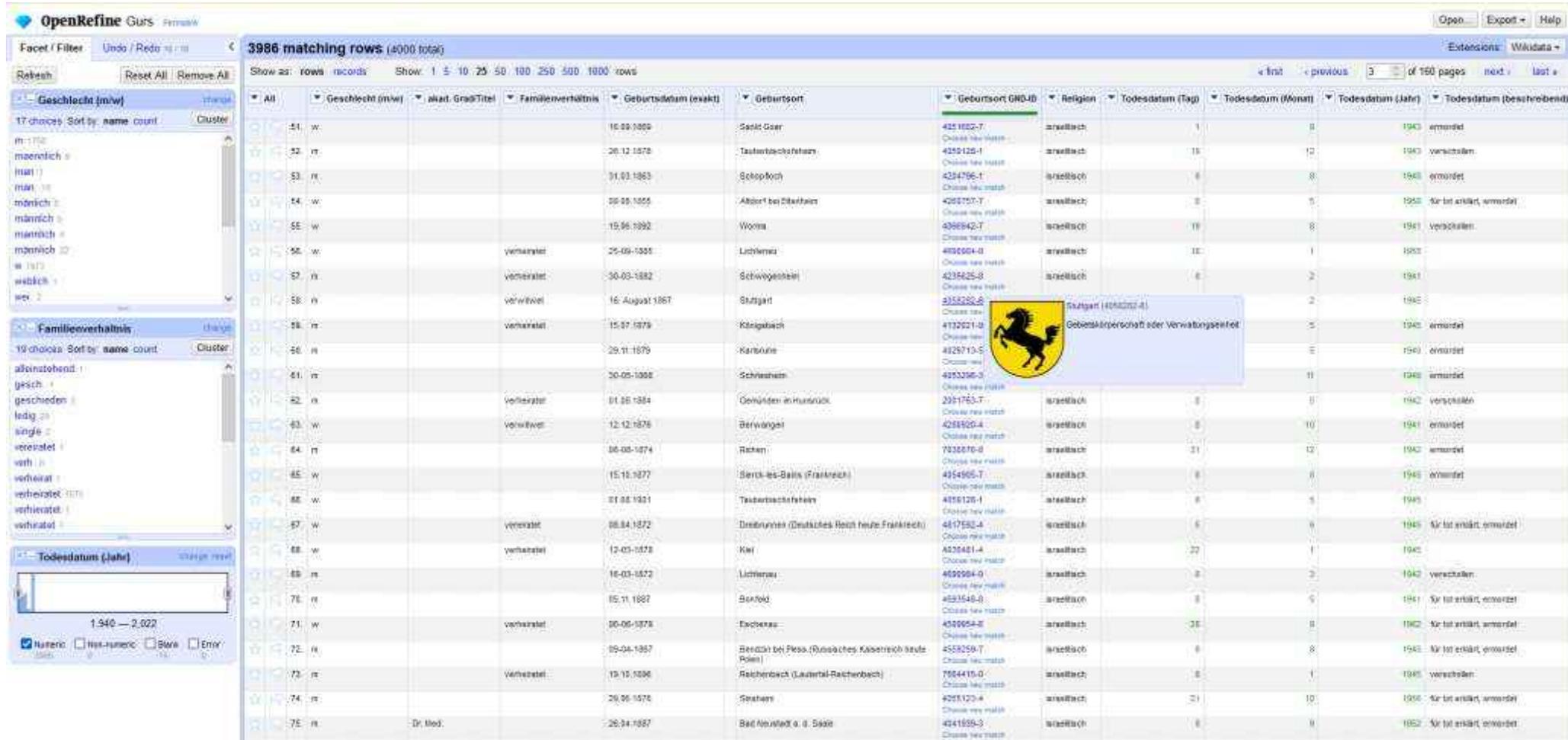


OpenRefine is like Excel on steroids!

Quelle unbekannt



<https://openrefine.org/>



OpenRefine Gurs Facet / Filter Undo / Redo 3 / 18 Open Export Help

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 150 pages next last

All	Geschlecht (m/w)	akad. Grad/Titel	Familienverhältnis	Geburtsdatum (exakt)	Geburtsort	Geburtsort GRD-ID	Religion	Todesdatum (Tag)	Todesdatum (Monat)	Todesdatum (Jahr)	Todesdatum (beschreibend)	
<input type="checkbox"/>	51. w.			16.09.1869	Sankt Gaer	4251832-7	evangelisch	1	8	1943	ermordet	
<input type="checkbox"/>	52. m.			26.12.1878	Taubenbäckersheim	425120-1	evangelisch	15	12	1943	verschollen	
<input type="checkbox"/>	53. m.			31.03.1863	Schopflach	4254706-1	evangelisch	6	8	1943	ermordet	
<input type="checkbox"/>	54. w.			29.05.1855	Albort bei Bittenheim	425757-7	evangelisch	8	5	1952	Sir ist erkrankt, ermordet	
<input type="checkbox"/>	55. w.			15.06.1892	Worms	4258942-7	evangelisch	19	8	1941	verschollen	
<input type="checkbox"/>	56. w.		verheiratet	25.09.1888	Lichtenau	492004-0	evangelisch	10	1	1932		
<input type="checkbox"/>	57. m.		verheiratet	30.03.1882	Schwägenstein	425625-8	evangelisch	6	2	1941		
<input type="checkbox"/>	58. m.		verwitwet	16. August 1867	Stützgart	425220-8				2	1942	
<input type="checkbox"/>	58. m.		verheiratet	15.07.1878	Königsbach	412221-0				5	1942	ermordet
<input type="checkbox"/>	56. m.			29.11.1879	Karlsruhe	425713-5				5	1943	ermordet
<input type="checkbox"/>	61. m.			30.05.1888	Schwanheim	425326-3				11	1940	ermordet
<input type="checkbox"/>	62. m.		verheiratet	01.05.1884	Göhndorf an Hunsrück	2011763-7	evangelisch	8	8	1942	verschollen	
<input type="checkbox"/>	63. w.		verwitwet	12.12.1876	Berwanger	425920-4	evangelisch	8	10	1941	ermordet	
<input type="checkbox"/>	64. m.			06.05.1874	Rachen	732076-0	evangelisch	31	12	1942	ermordet	
<input type="checkbox"/>	65. w.			15.10.1877	Sers-les-Bains (Frankreich)	425465-7	evangelisch	8	8	1943	ermordet	
<input type="checkbox"/>	66. w.			01.02.1921	Taubenbäckersheim	4256126-1	evangelisch	8	5	1945		
<input type="checkbox"/>	67. w.		verheiratet	08.04.1872	Dreibrunn (Deutsches Reich heute Frankreich)	4417582-4	evangelisch	5	9	1943	Sir ist erkrankt, ermordet	
<input type="checkbox"/>	68. w.		verheiratet	12.03.1878	Karl	423481-4	evangelisch	27	1	1942		
<input type="checkbox"/>	69. m.			16.03.1872	Lichtenau	492004-0	evangelisch	8	2	1942	verschollen	
<input type="checkbox"/>	70. m.			05.11.1887	Berfeld	4251548-0	evangelisch	8	5	1941	Sir ist erkrankt, ermordet	
<input type="checkbox"/>	71. w.		verheiratet	06.05.1878	Eichenau	4200054-0	evangelisch	20	8	1942	Sir ist erkrankt, ermordet	
<input type="checkbox"/>	72. m.			09.04.1867	Bendin bei Pless (Deutsches Kaiserreich heute Polen)	4258259-7	evangelisch	8	8	1943	Sir ist erkrankt, ermordet	
<input type="checkbox"/>	73. m.		verheiratet	19.10.1896	Reichenbach (Lautertal-Reichenbach)	7064415-0	evangelisch	8	1	1945	verschollen	
<input type="checkbox"/>	74. m.			26.06.1878	Sersheim	425120-4	evangelisch	21	10	1956	Sir ist erkrankt, ermordet	
<input type="checkbox"/>	75. m.	Dr. Med.		26.04.1867	Bad Nauendorf o. B. Seeb	4241859-3	evangelisch	8	9	1952	Sir ist erkrankt, ermordet	

Facets:

- Geschlecht (m/w)**: 17 choices. Sorted by name count. Cluster.
- Familienverhältnis**: 19 choices. Sorted by name count. Cluster.
- Todesdatum (Jahr)**: 1.940 — 2.022. Includes filters for Numeric, Non-Numeric, Blank, Empty.

Tooltip:  Stützgart (406007-4) Geburtskörperchaft oder Verwaltungsstelle

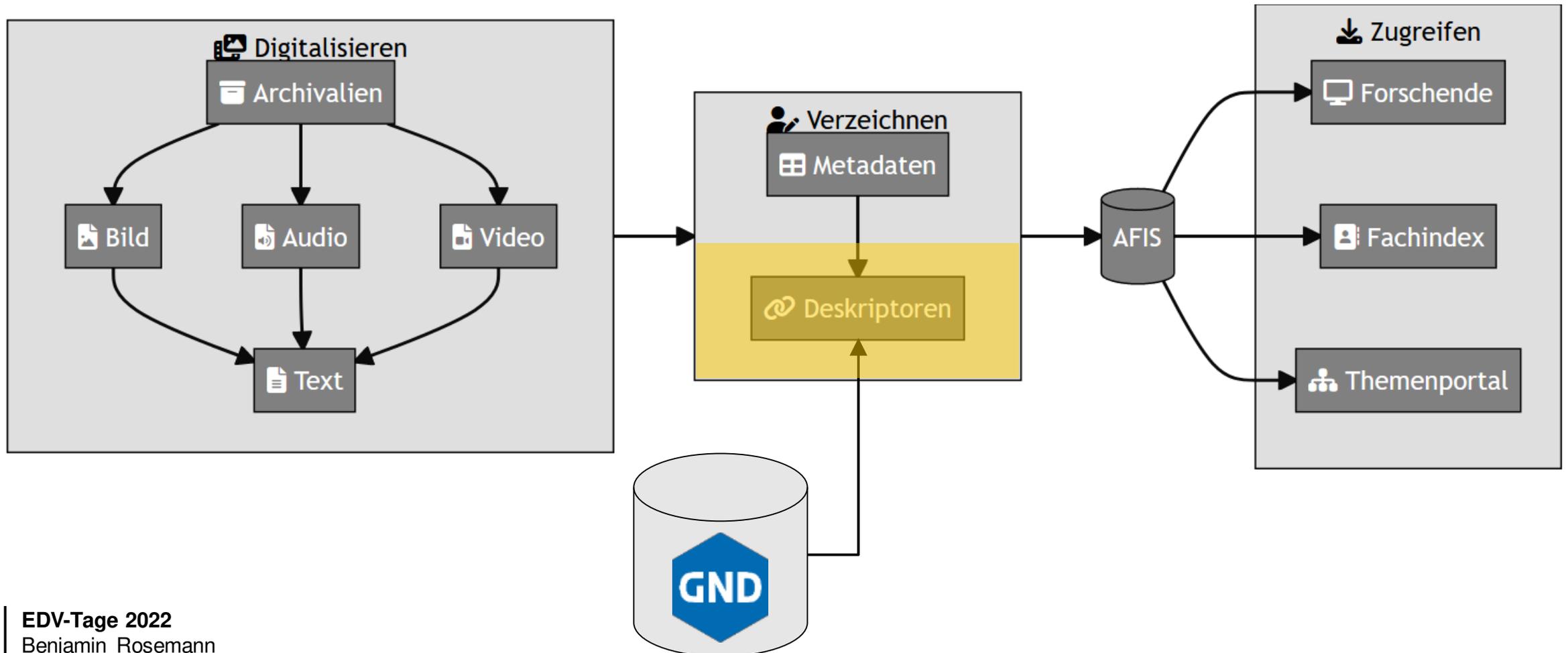
Integrierte Algorithmen

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

Name
Benjamin Rosemann
Benjamin Rosemann
Rosemann Benjamin
Beniamjn Rosemann
Benjamin Roseman
Benjamin Rosemam
Benjamin R. Rosemann

Orte
Nordschwaben (Rheinfelden)
Nordschwaben bei Rheinfelden
Nordschwaben-Rheinfelden
Rheinfelden-Nordschwaben
Nordschwaben / Rheinfelden
Rheinfelden Nordschwaben
Rheinfelden, Nordschwaben

Workflow Digitalisierung



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



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

Ungenaue Namen mit OpenRefine abgleichen



<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

Ungenauere Namen mit OpenRefine abgleichen



<https://fdmlab.landesarchiv-bw.de/workshop/openrefine-fortgeschrittene/15-erweiterter-gnd-abgleich-mit-lobid/>

Reconcile column "Name"

[Access Service API](#)

Reconcile each cell to an entity of one of these types:

- Normdatenressource
AuthorityResource
- Individualisierte Person
DifferentiatedPerson
- Geografikum
PlaceOrGeographicName
- Schlagwort
SubjectHeading
- Werk
Work

Also use relevant details from other columns:

Column	Include?	As Property
GND-ID	<input type="checkbox"/>	<input type="text"/>
Geburtsdatum	<input checked="" type="checkbox"/>	Geburtsdatum
Geburtsort	<input type="checkbox"/>	<input type="text"/>
Geburtsort (GND-ID)	<input checked="" type="checkbox"/>	Geburtsort

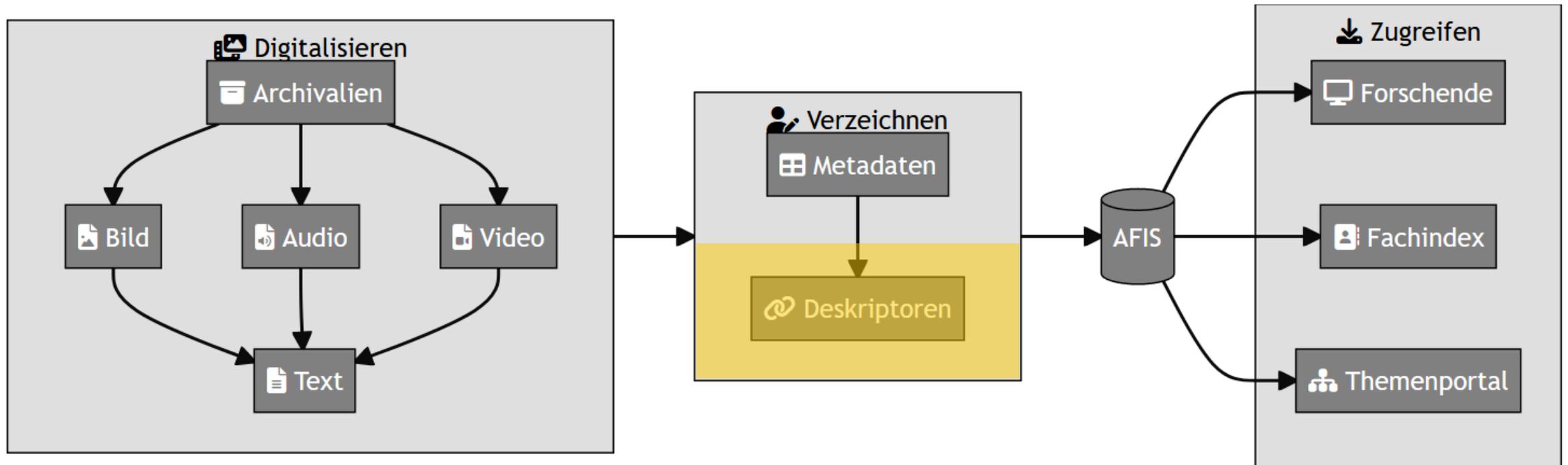
Reconcile against type:

Reconcile against no particular type

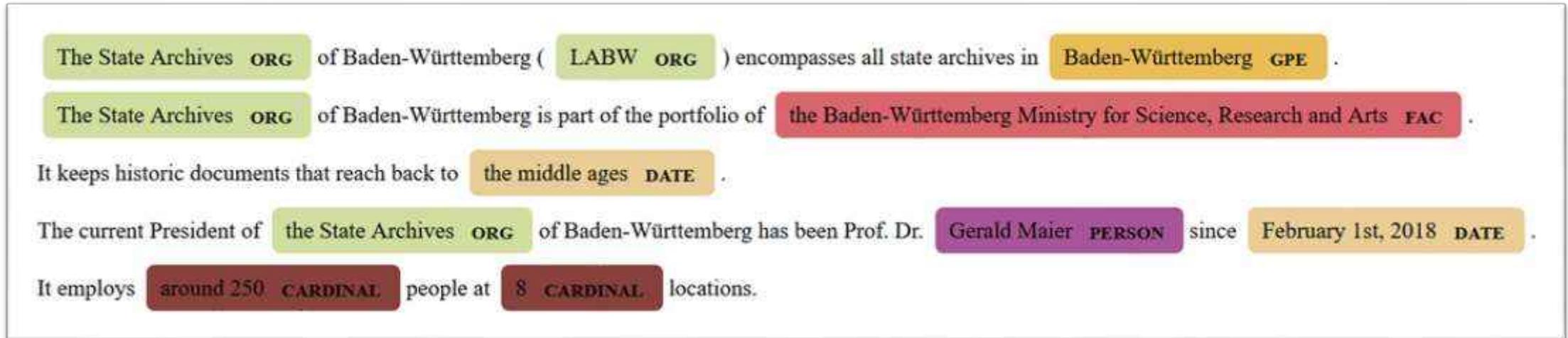
Auto-match candidates with high confidence

Maximum number of candidates to return

Workflow Digitalisierung



Named-entity recognition: Englisch



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 Autoren, „Landesarchiv Baden-Württemberg“, Wikipedia – Die freie Enzyklopädie, https://de.wikipedia.org/w/index.php?title=Landesarchiv_Baden-W%C3%BCrttemberg&oldid=223510351 (aufgerufen am 20. Juli 2022).

Named-entity recognition: Deutsch

Das **Landesarchiv Baden-Württemberg** **LOC** (LABW) umfasst alle **baden-württembergischen** **MISC** Staatsarchive.

Es gehört zum Geschäftsbereich des **Ministeriums für Wissenschaft, Forschung und Kunst Baden-Württemberg.** **ORG**

Es verwahrt historische Dokumente, die bis zum Mittelalter zurückreichen.

Der amtierende Präsident des **Landesarchives** **PER** ist seit dem 01. Februar 2018 Prof. Dr. **Gerald Maier.** **PER**

Es beschäftigt circa 250 Mitarbeiter an 8 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, „Landesarchiv Baden-Württemberg“, Wikipedia – Die freie Enzyklopädie, https://de.wikipedia.org/w/index.php?title=Landesarchiv_Baden-W%C3%BCrttemberg&oldid=223510351 (aufgerufen am 20. Juli 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.

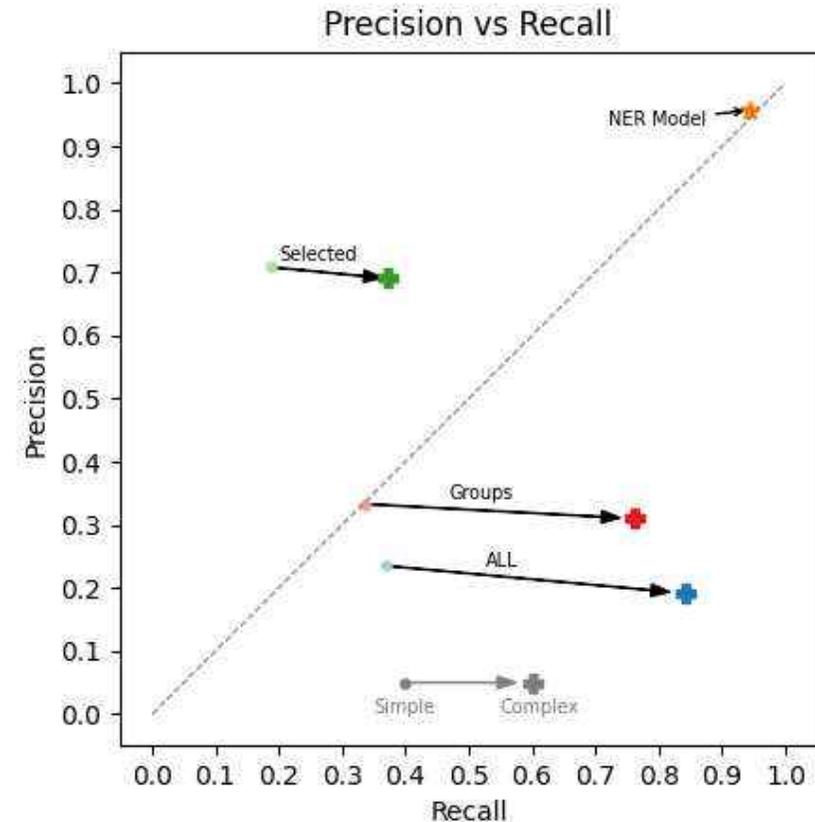
Gen. Forderung stammt aus Streit über Erbe des im November 1595 verst. K[C]aspar (von) **Laubenberg. MISC**

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.

NER Modelle für Objekte

- Erkennung von Objekten
- Beispiel: Ludwigsburger Porzellantasse
- Starthilfe mit Sachbegriffen aus der GND
- Mit GND-Daten: **f-score 0,44-0,48**
- Aktives Lernverfahren
- Eigenes Modell: **f-score 0,95**

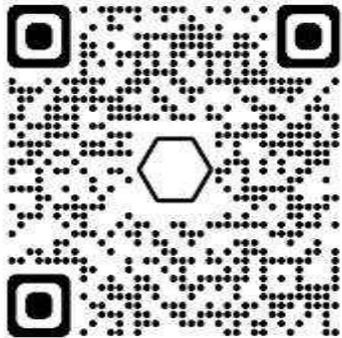


04

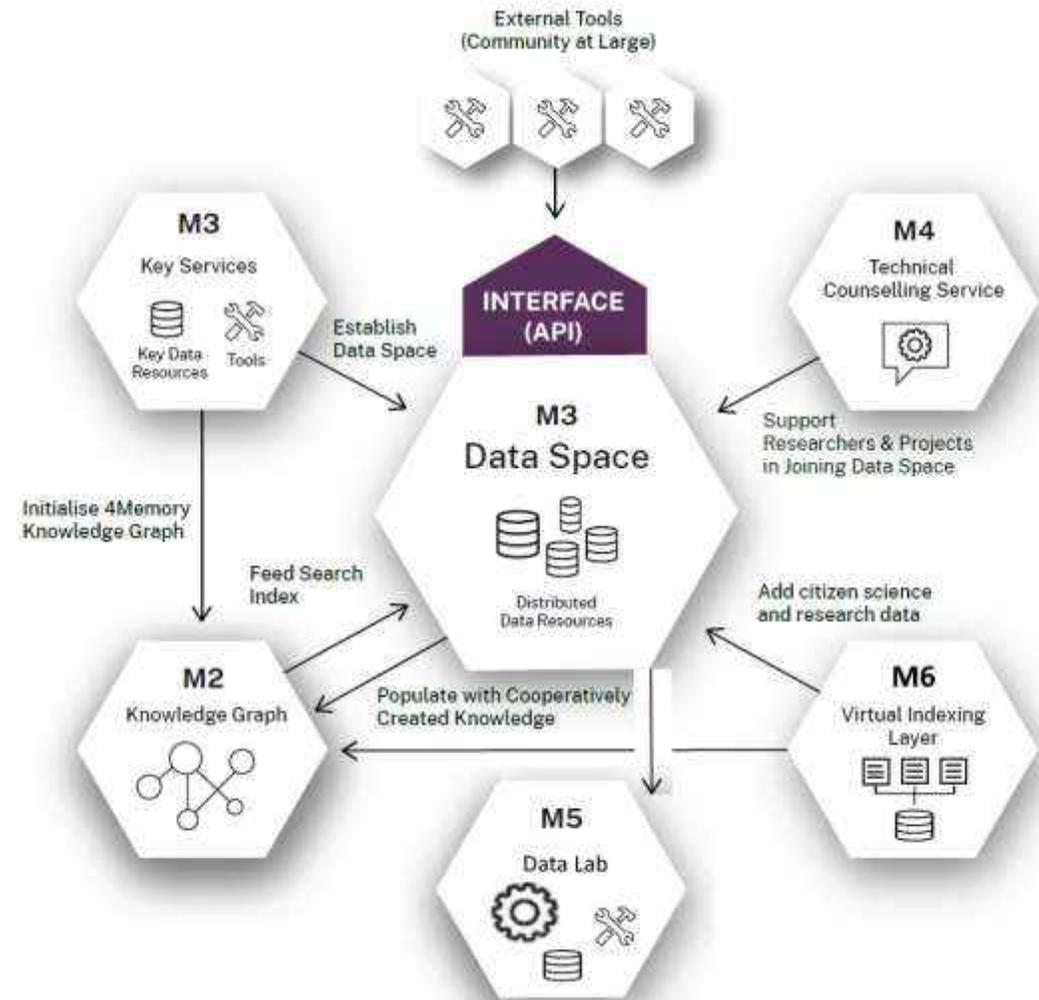
KI-Infrastruktur für Archive?

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>

