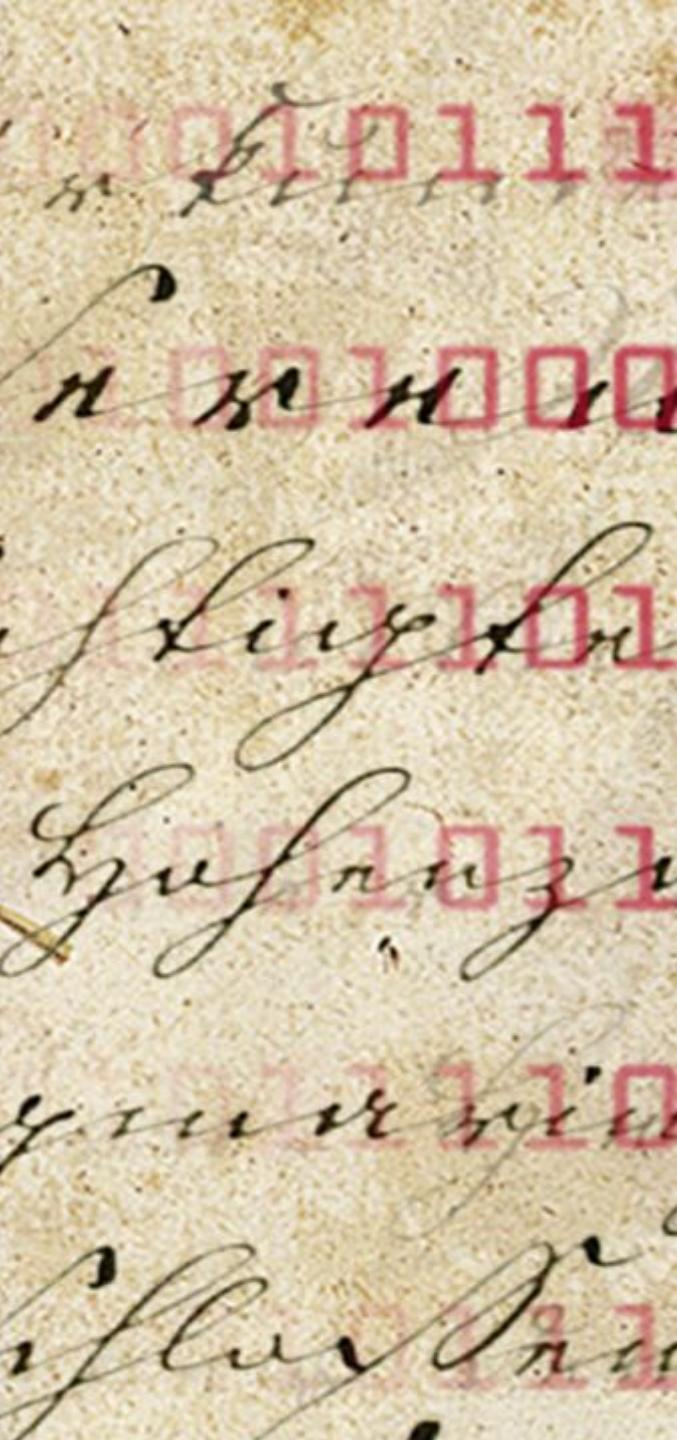


Ein digitaler Werkzeugkasten für historische Forschung mit Archivgut Status quo und Perspektiven

53. Deutscher Historikertag, Fachsektion 6.10.2021

Prof. Dr. Harald Sack (FIZ Karlsruhe), Daniel Fähle (Landesarchiv BW)



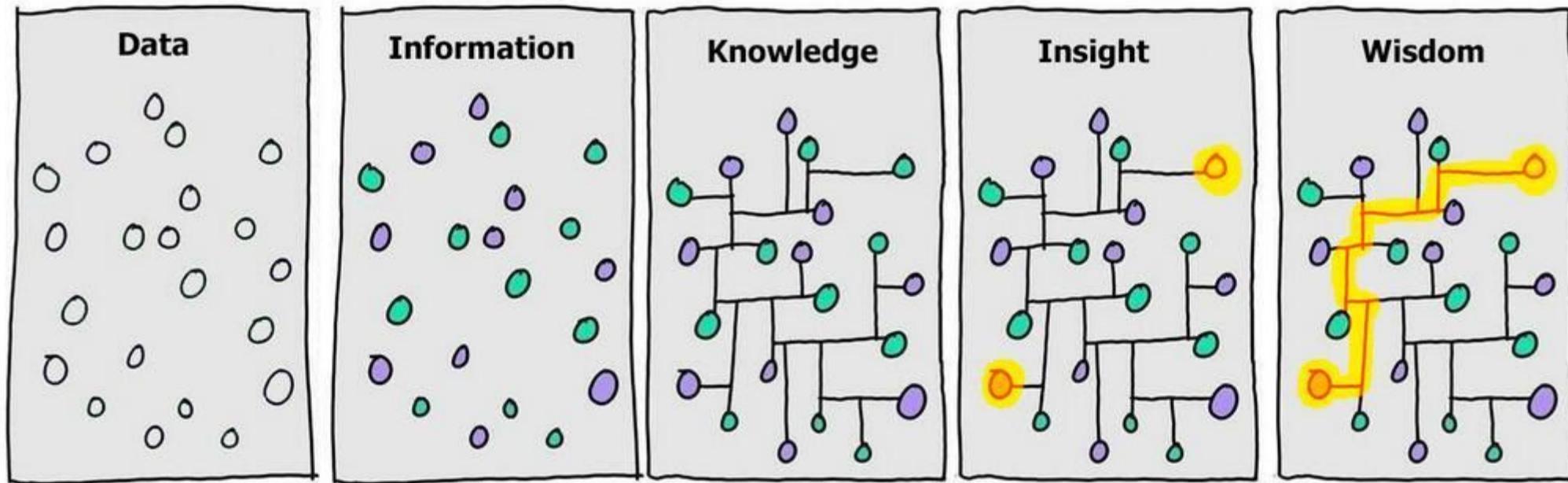
01 Archivportal-D und
4Memory Data Space

02 Datenqualität und
Schnittstellen

03 Anwendungen und
Kompetenzen

04 Wissensextraktion
aus historischen
Forschungsdaten

05 Angewandte
Wissensgraphen



Quelle: Internet Meme, Herkunft: ?

01

Datenangebot: Archivportal-D und 4Memory Data Space



*Historiker*innen brauchen für eine effiziente Forschung aggregierte Datensammlungen, die eine effiziente Suche und einen Zugriff auf Inhalte aus der Ferne erlauben.«*

Jörg Wettlaufer, Welche Services braucht die digitale Geschichtswissenschaft von Bibliotheken, Archiven, Museen und Datenzentren?, 5.10.2021

Archivportal-D

- seit 2014 “Data Hub” der Archive, eng angelehnt an Deutsche Digitale Bibliothek
- Ziel: vollständiges Nachweissystem für Archivgut in Deutschland
- sachthematische Zugänge (Weimar, Wiedergutmachung)
- zentraler Andockpunkt für Services (z.B. Normdatenanreicherung)
- API der DDB nutzbar

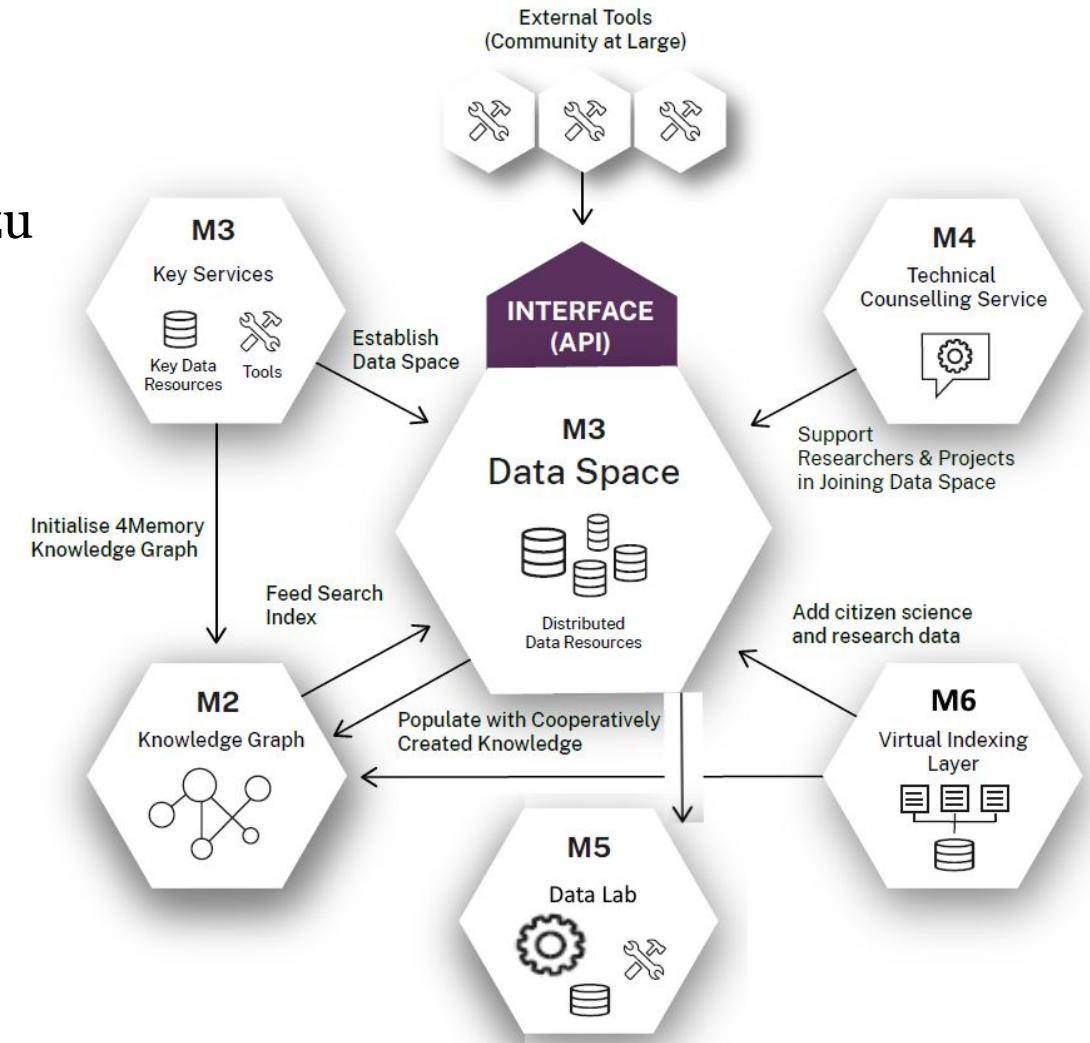


Quelle: pixabay

NFDI4Memory - Data Space



- Bestehende relevante Repositorien, DH-Tools und Datenangebote interoperabel und FAIR machen
- Vernetzung bestehender Dienste/Infrastrukturen zu einem übergreifenden »Ökosystem« (Data Space)
- Bedarfsgerechte Optimierung und Erweiterung zentraler Angebote
- Services: übergreifende Suche auf Basis eines Knowledge Graph; kooperative Erschließung
- Data Lab: Erprobung von Anwendungen und Technologien



Erweiterung der Datengrundlage

Grundproblem:

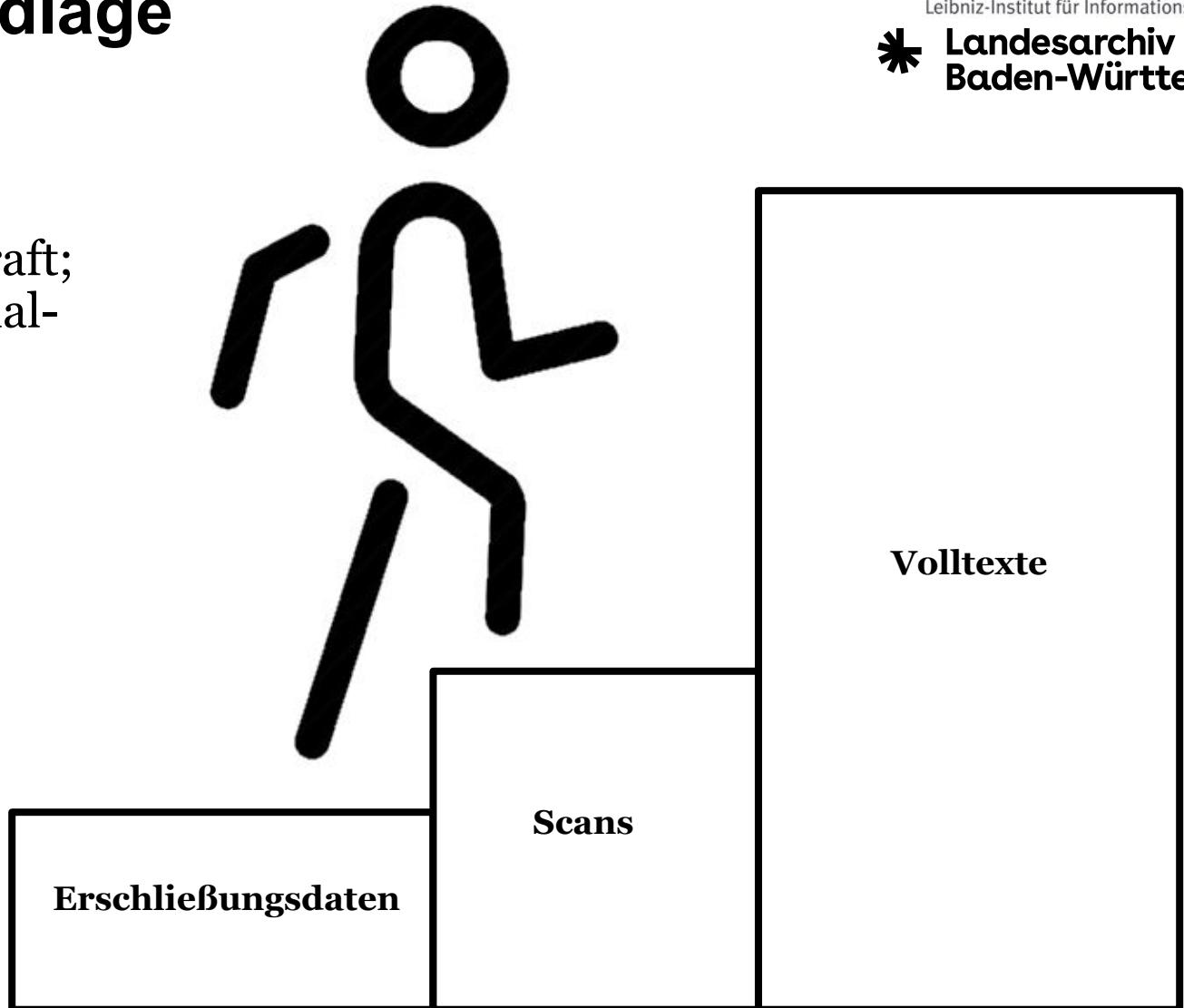
Erschließungsdaten mit wenig Aussagekraft;
flache Erschließung aufgrund von Personal-
und Ressourcenmangel

Abhilfe:

(semi-)automatisierte Verfahren zur
Generierung von Metadaten durch
Datenextraktion und Mining-Verfahren

Quantensprung:

Bereitstellung von Volltexten



02

Bedarfe: Datenqualität und Schnittstellen

FAIR-Prinzipien für Datenqualität

Zentrale Leitplanken

- stabile Identifikation/Zitierfähigkeit
- standardisierte und maschinenlesbare Metadaten
- Verwendung von Normdaten
- rechtssichere Nachnutzung durch Einsatz gängiger Lizenzen
- Bereitstellung von Schnittstellen/API
- Images in ausreichender Qualität



Quelle: pixabay

Datenbereitstellung - Schnittstellen

Vorhanden:

- Bereitstellung von Metadaten und Primärdaten im Download sowie Exporte im EAD(DDB)- und METS/MODS-Format

Bedarfe:

- Datenbereitstellung für individuelle Datensets und Corpora
- RDF-Exporte für semantische Nutzungsszenarien

Lösungsansätze:

- Erweiterung der DDB-API
- Records in Contexts (RiC): neuer Erschließungsstandard unterstützt semantische Datenmodellierung (Umsetzung im DFG-Projekt EEZU)

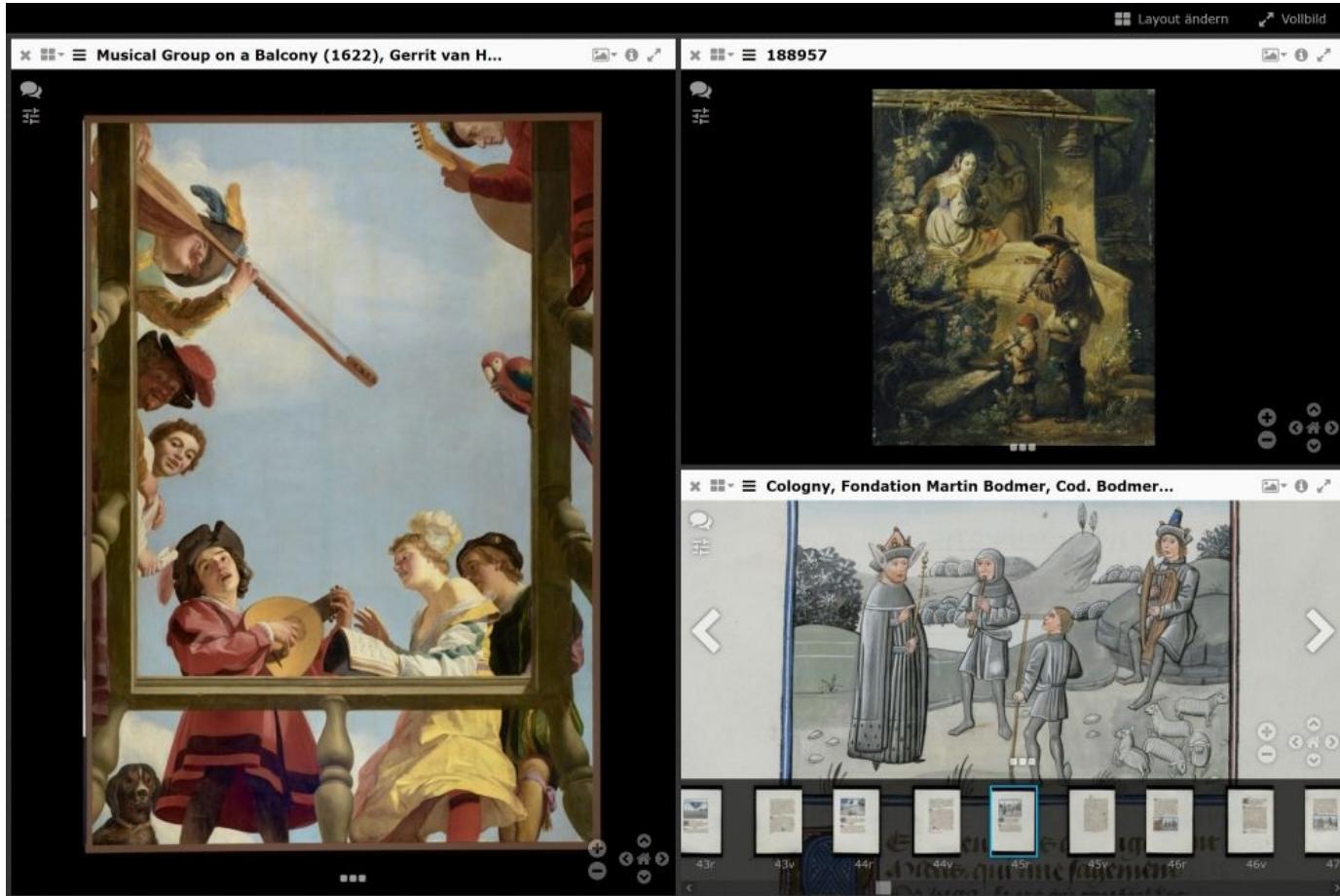


Quelle: Twitter; @hapyresearchers

Schnittstellen



- hochperformant
- standardisierte Schnittstellen für Metadaten, Images und Präsentation
- Zusammenführung von Bildern aus unterschiedlichen Quellen
- ermöglicht Vielzahl an Nutzungsmöglichkeiten in eigenem Viewer (Annotation etc.)



Quelle: Sonja Gasser; Mirador-Viewer
<https://www.artlog.net/de/notebook/iiif-kleines-experiment-mit-einem-bildstandard>

03

Werkzeuge: Anwendungen und Kompetenzen

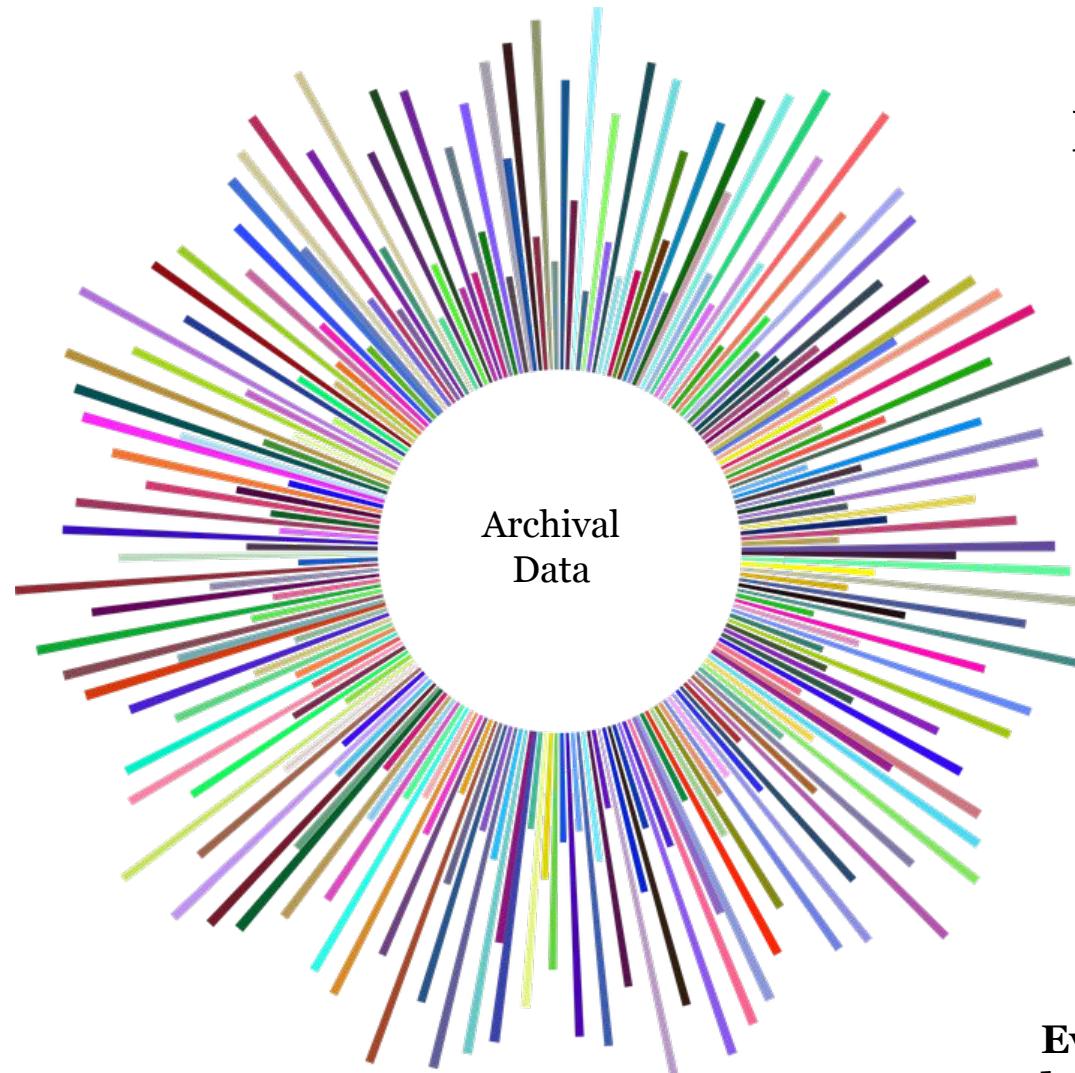
Anwendungen

Bereinigung

Transkription

Annotation

Geoinformationssystem



Datenanalyse

Datenstrukturierung

Visualisierung

Extraktion/Mining

Evaluierung u.a. im FDMLab:
<https://fdmlab.landesarchiv-bw.de/>

Kompetenzen

Data Literacy auf der Nutzerseite:
“the ability to collect, manage, evaluate, and apply data in a critical manner” (C. Ridsdale)

Aufgabenverteilung:
Archive stellen Daten über Schnittstellen bereit, Forscher*innen setzen passende Tools ein

Pflege des Werkzeugkastens:
gemeinsame Aufgabe von Archiven, Forschung, Digital Humanities und Informatik



E-Mail und Telefon

Ein größerer Anteil an Findmitteln und der größte Teil des Archivguts wird auch künftig nur analog zur Verfügung stehen.

Archivarinnen und Archivare beraten und unterstützen gerne!



04

Wissensextraktion von historischen Forschungsdaten

OCR und Texttranskription

Landratsamt Kehl

Fernsprecher 295/297
Nr. 1064 Abt. I a

23. Juni 1952

47 35

Kehl a. Rh. den

Wiedergutmachung nat.soz.Unrechts an Angehörigen des öffentlichen Dienstes

-Auf die Erlasse vom 3.4.52 Nr. 14 531/III und vom 29.5.52 Nr. 24 808/III -
Anl.: 2 Hefte Akten

hier:
Antrag des Ratschreibers Karl in Legelshurst.

Blatt 1 37
33 Blatt I.
Legelshurst, 16. 5. 1952.
des Innern
JUL 1952
452/111

Niedergutmachung nat.soz. Unrechts an Angehörigen des öffentlichen Dienstes

Antrag des Ratschreibers Karl in Legelshurst.

024435/III 19. MAI 1952

An Ministerium des Finanz., Freiburg.

Zum Ausklang an den bereits in Bearbeitung befindlichen Antrag.

1. Von interessierter Seite wurde ich darauf aufmerksam gemacht, daß auch der durch die Dienstklassung 1933 abschließend verliehene Ruhegehalt zur Wiedergutmachung kommt. Ich bezog 1932/33 im Gehalt von 1500 RM. jährlich für 2 Dienststunden, 3 Normale nach neuem Entlassung - im Februar 1933 - teils mit dem

04

Eine Wiedereinstellung als Ratschreiber der Gemeinde Legelshurst erfolgte erst wieder auf Grund der Bestellung durch den Herrn Landrat des Kreises Kehl mit Schreiben vom 4. September 1945. Infolge der tatsächlichen Sachlage befürwortet das Bürgermeisteramt Legelshurst die Antragstellung auf Wiedergutmachung".

Landesamt für die Wiedergutmachung Stuttgart

Herrn
Landgerichtspräsidenten a.D. Marx beim Landgericht Stuttgart Stuttgart.

Betr.: Entschädigungsanschrift

Bezug: Dortiges Schreiben vom 13.5.65

Anl.: 1 Bd. Akten 11737-1 bis 11737-11

14 8.9.65

7 Stuttgart-1, den
Theodor-Heuss-Str. 26
Tel.: 296051/52
Az.: ET - 11737 u. 11737-1 - I Wei
8. 9. 65

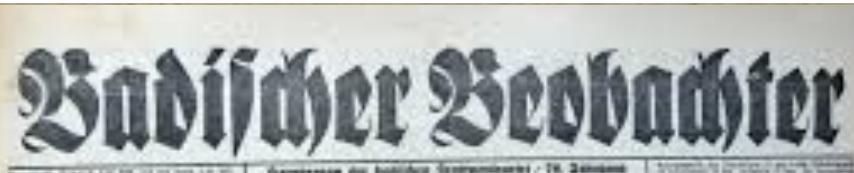
Auskunft aus dem Strafregister der Staatsanwaltschaft zu Mannheim

22. JUL. 1917 5

Familienname: [REDACTED]	Vorname: Jakob (Rufname unterstreichen)					
Geburtsangaben: (Tag, Monat, Jahr) 17. 1. 1878	Gemeinde: Mannheim (entl. Stadtteil) Stadt: Mannheim Verwaltungsbezirk: Mannheim					
Landgerichtsbehörde: Mannheim						
Familienstand: ledig — verheiratet — verwitwet — geschieden						
Vor- und Familien-(Geburts-)Name des (bzw. früheren) Ehegatten:						
Vor- und Familienname des Vaters:	Vor- und Geburtsname der Mutter: Eing 18. JAN 1947					
Stand (Beruf): Schneider ggf. des Ehemanns:						
Wohnort ggf. letzter Aufenthaltsort: Straße und Hausnummer:	Jugen, Blauehardstr. 8/91					
Staatsangehörigkeit:	Heimatgemeinde: Mannheim					
d.	Heimatbezirk: Mannheim					
Verurteilung(en) vermerkt:						
Mr.	am	durch	wegen	auf Grund von	zu	Bemerkungen
Keine Vorstrafen Die Richtigkeit der Personalen ist nicht nachgeprüft. Mannheim, d. n. 20. Jan. 1947. Staatsanwaltschaft Der Strafregisterdienst a. L. K. Laut Justiz - Oberinspektor Angestellte						

Str. 9

OCR und Texttranskription



Der Lehrer der Menschheit

Sam 10. November der Theodoritische des Baptist. Band XI.

Paulus Christian Christoph Laufener Valerij Maximiliane urbis urbi et inimico memoratu digna que apud tuis diffusa sit ut quambre gnoxi possint ab illustribus auctoribus diligere constituta sunt euolentibus longe magnificis labores mihi cuncta complectendi cupido incessit. omnis eius gesta modico voluminu numero considerite Aut quis compos mentis domestice per historie seriem feliciori stilo conditam ciore cura vel prestanciori facundia traditur ueritatem Teigitur huic cepto penes quem homini



foloweth the Recapitulacion
pitulo
hier endeth the table of the Ru-
esent book.

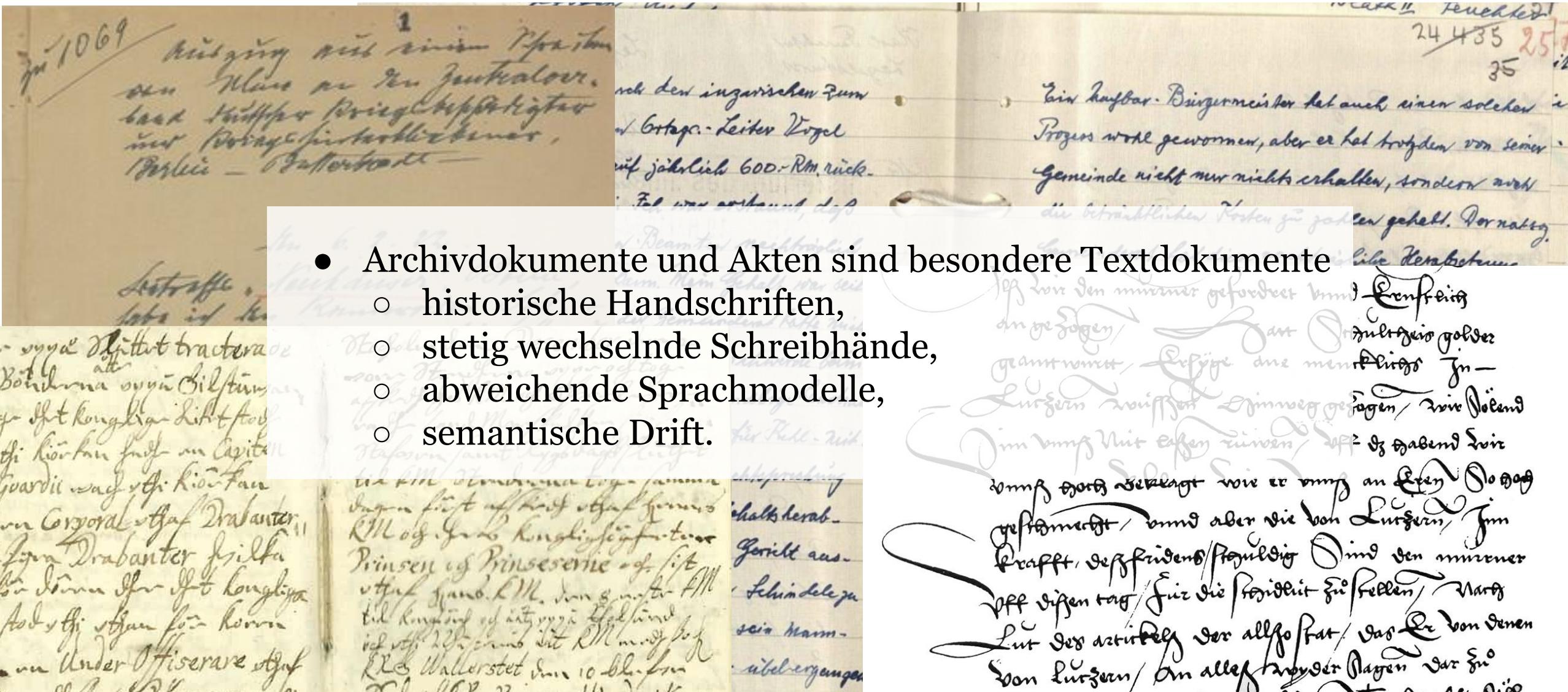


TRESEXCEL-
LENT ET ANCIEN
ieu Pythagorique, dict Rythmoma-
chie, fort propre & tresvtil à la re-
creation de

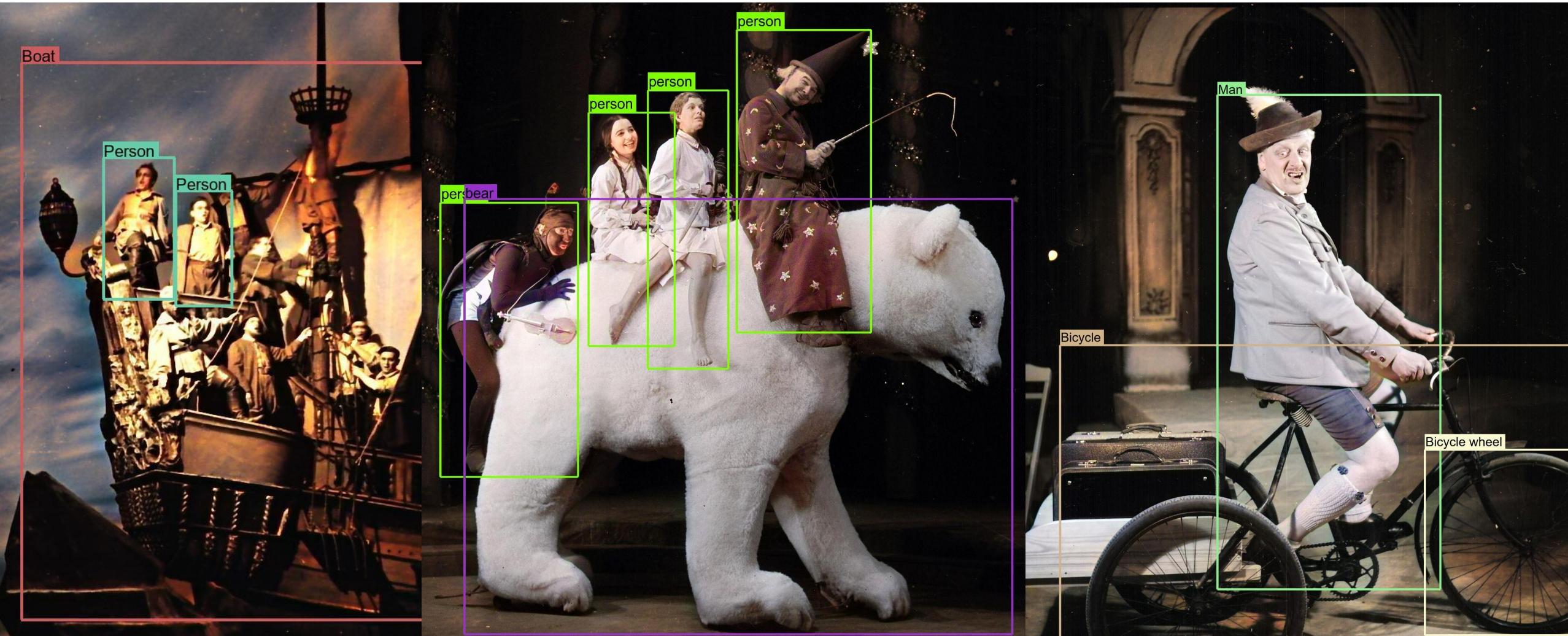
- Archivdokumente und Akten sind besondere Textdokumente
 - historische Schriftformen,
 - ungewöhnliche Formatierung und Layout,
 - abweichende Sprachmodelle,
 - semantische Drift.

obtenuir vra
Graff zu Sausenberg Graff zu Spanheim und
Lahr und Mahlberg ic. Ob zwahren auf Unser
befindenden und bey vorgeweihten seitzen laetigen Kriegs-Wese, gänglich dev artier un
fer und Wohnungen auferbaet werden. Wir aber gnädigst gemeint sind und gern sch
völlig aufgebracht und in guten Stand gesetzte werde. So haben Wir nicht allein be
sage Unser Residenz-Stadt mit der bereits angesangnen Fortification gänglichen zu sch
eher völlig erbaut werde, die so wol Unsern eingessenen Burgern als auch denen Fremden
neue Häuser zu erauen vermö ermeidet Patentes ertheile Freheiten hinc wiederhol
welt fremde/ was Nation unter denen deren im Röm. Reich/ vermö Frieden Schlus
fändische/ welche in mehr ermeidet Unserer Residenz Model-mäßige Häuser bauen wer
blicationis dren Jahr gegeben/ denen Fremden aber weder Zeit noch gewisser Termin p
gebauet und bezogen sijn wird/ von ihren Häusern und Gütern/ von der Frohn in allem a
rer Hauss-Batter von Unsern eingessenen Burgern/ so solches Haus erbauet mit tod ab
solchen Kindern Jeniges/ deme sothans Haus an seinem Erbtheit zufallere/ frey ingleid
ein vor allem nicht das geringste an Herrschafft. Geldter/ worunter aber Reichs- und G
Bandshäusern nicht verstanden/ außer einer Stuben/ und Cammer zu Logierung Unse
dig seyn. Da aber ermeidet Unser eingessene Burgere ihre Häuser in den angezeigten
den/ alsdann die oder derselbe obbenandte Freiheit nicht allein nicht zu geniesen haben/ so
bauen angegeben/ überlassen werden solle/ zumahnen Wir öfters ersage Unser Residenz-E
seln Wir auch ferners/ so viel die Leibhagenshaft betrifft/ die Fremden nicht allein für si
tuum exempt, nicht weniger auch das zur Erbauung solcher neuen Häuser benötigte B
fogeln lassen. Soleten aber die Fremde einige burgische Güter an sich erhandlen/ so v
gewöhnlichen Anlagen/ wie vorher/ underwerffen: Wie es mit den Inhabitischen dñss
sionem gemasum explicitiert worden. Welches zu Mänlichens Wissenschaft under
affigiert wird. So geschehen Rastatt/ den 5. Decemb. 1701.

Handschriften und OCR



Automatisierte visuelle Analyse



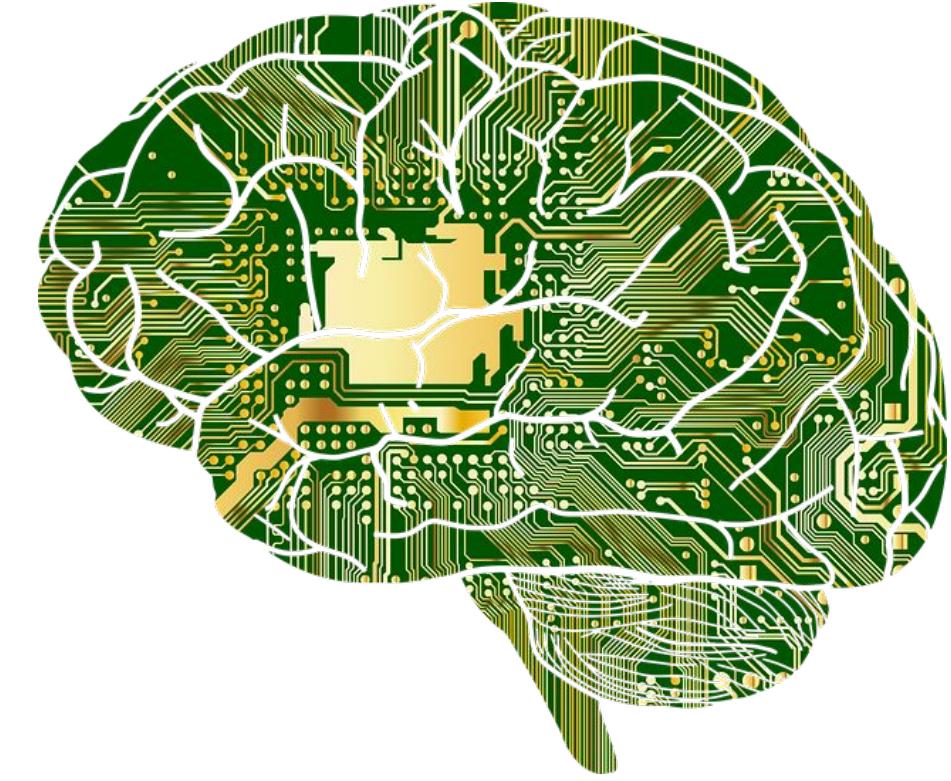
Automatisierte visuelle Analyse

- Archivdokumente sind besondere Bilddokumente
 - historische Fotografien
 - historische Gemälde und Zeichnungen
 - mit oftmals unklarem Kontext
- Deep Learning basierte Bild- und Objekterkennungsverfahren verfügen oft über einen “modernen” Bias



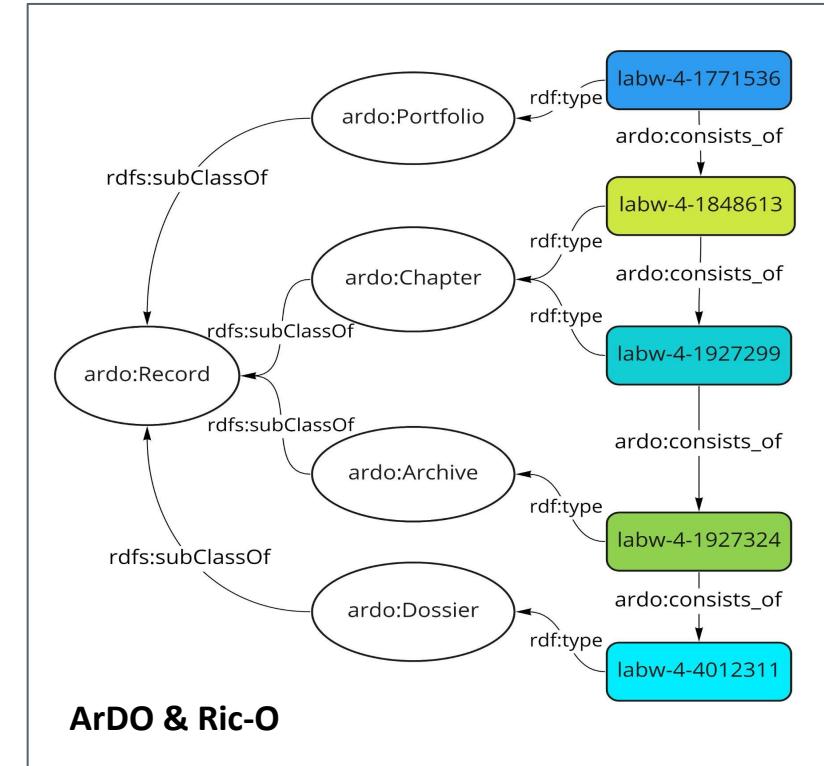
Language Models

- Repräsentation von natürlicher Sprache als statistisches Model
- Nahezu grenzenlos verfügbare Trainingsdaten
- Repräsentation der Semantik natürlicher Sprache durch Einbettung (Embedding) in einen (dichten) Vektorraum
- Potenzielle Einsatzfelder in Archiven:
 - Qualitätsverbesserung von OCR Transkripten
 - Ähnlichkeitsbasierte Suche
 - automatisierte Verschlagwortung
 - automatisierte inhaltliche Zusammenfassung
 - automatisierte Bild- und Objektbeschreibungen
 - automatisierte Übersetzungen
 - inhaltsbasierte Empfehlungssysteme



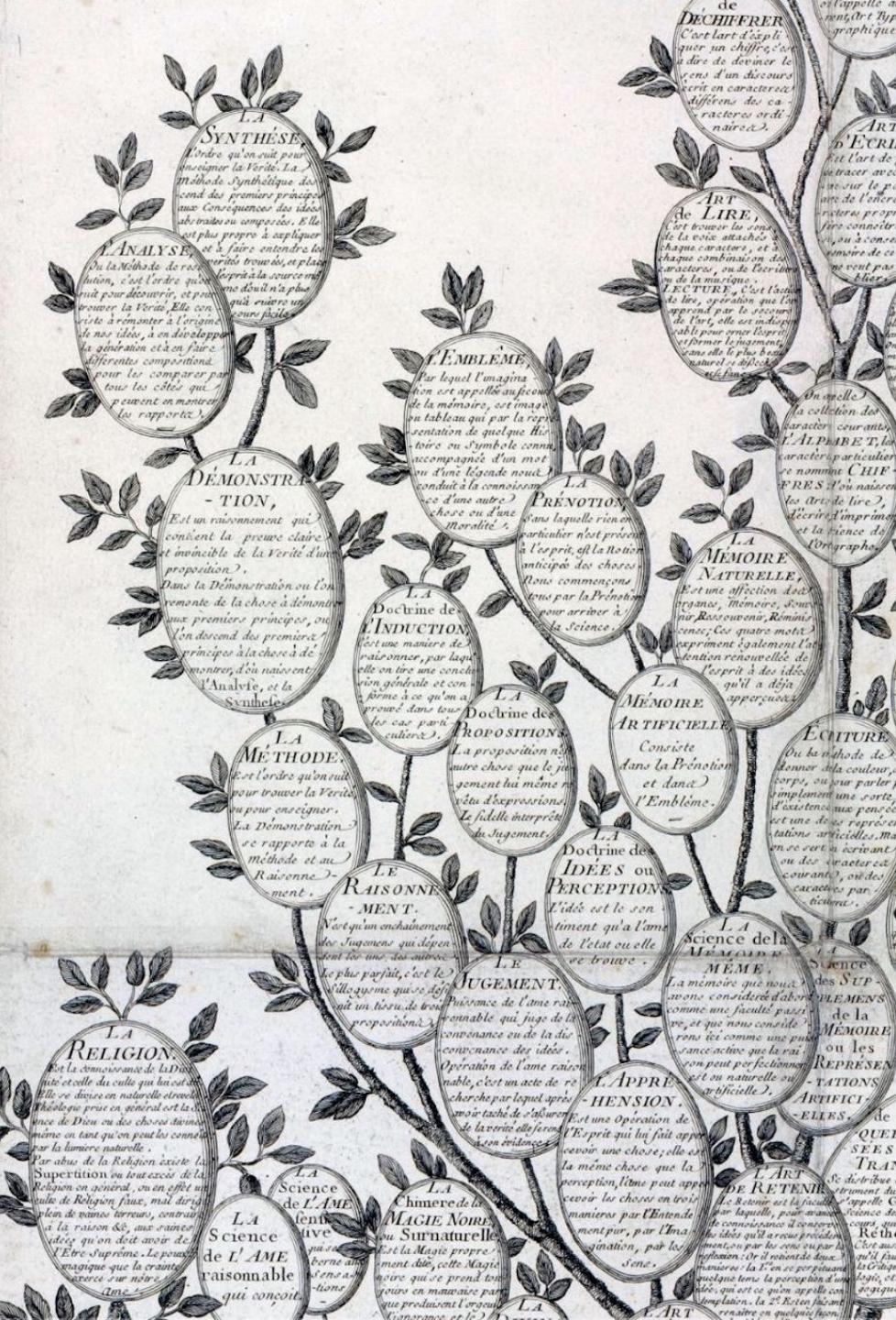
Semantische Repräsentation von Archivwissen

- Ontologien zur semantischen Repräsentation von Archivstrukturen, -vorgängen und -inhalten
 - (statische) strukturelle Organisation
 - Vorgänge und Prozesse
 - zeitliche Dynamik und Entwicklungen
 - Provenienz und Kontext
 - Klassifikation und Klassifikationsdynamik



Semantische Analyse und Erschließung

- Wissensgraphen zur Erschließung und Anreicherung von Archivinhalten
- Semantische Annotation von Archivdokumenten
- Potenzielle Einsatzfelder in Archiven:
 - semantische Suche
 - Exploration des Archivbestandes
 - inhaltsbasierte Empfehlungen
 - Visualisierungen



Ontologien, Wissensgraphen und Normdaten

- **Ontologien** formalisieren die Bedeutung der durch sie repräsentierten Konzepte und Relationen
- **Wissensgraphen** verknüpfen Bedeutungsinhalte auf der Basis von Ontologien
- **Normdaten** bilden eine (traditionelle) Datenreferenz
- Miteinander verknüpft bilden sie die Basis zur effizienten Umsetzung der **FAIR Prinzipien** für Archivdaten



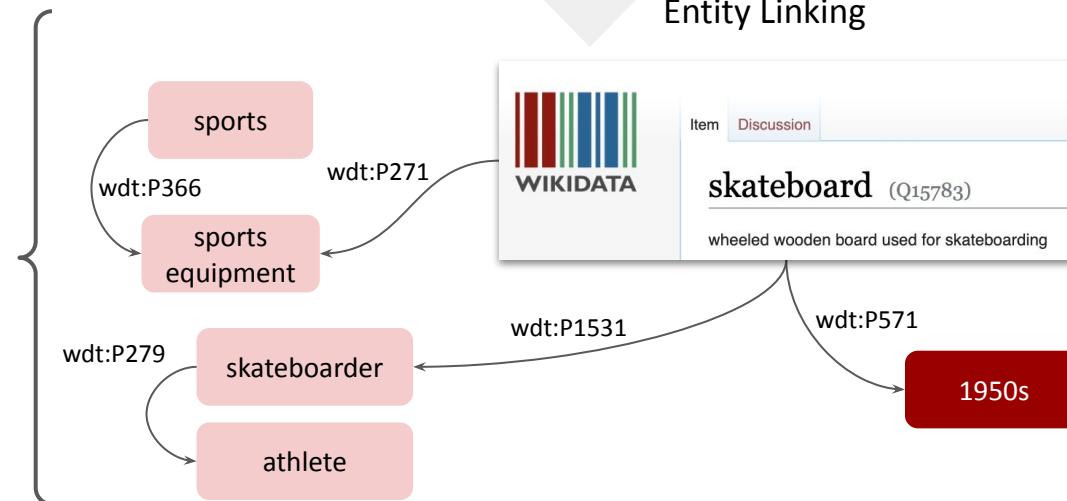
05

Angewandte Wissensgraphen

Zusammenspiel symbolischer und subsymbolischer KI

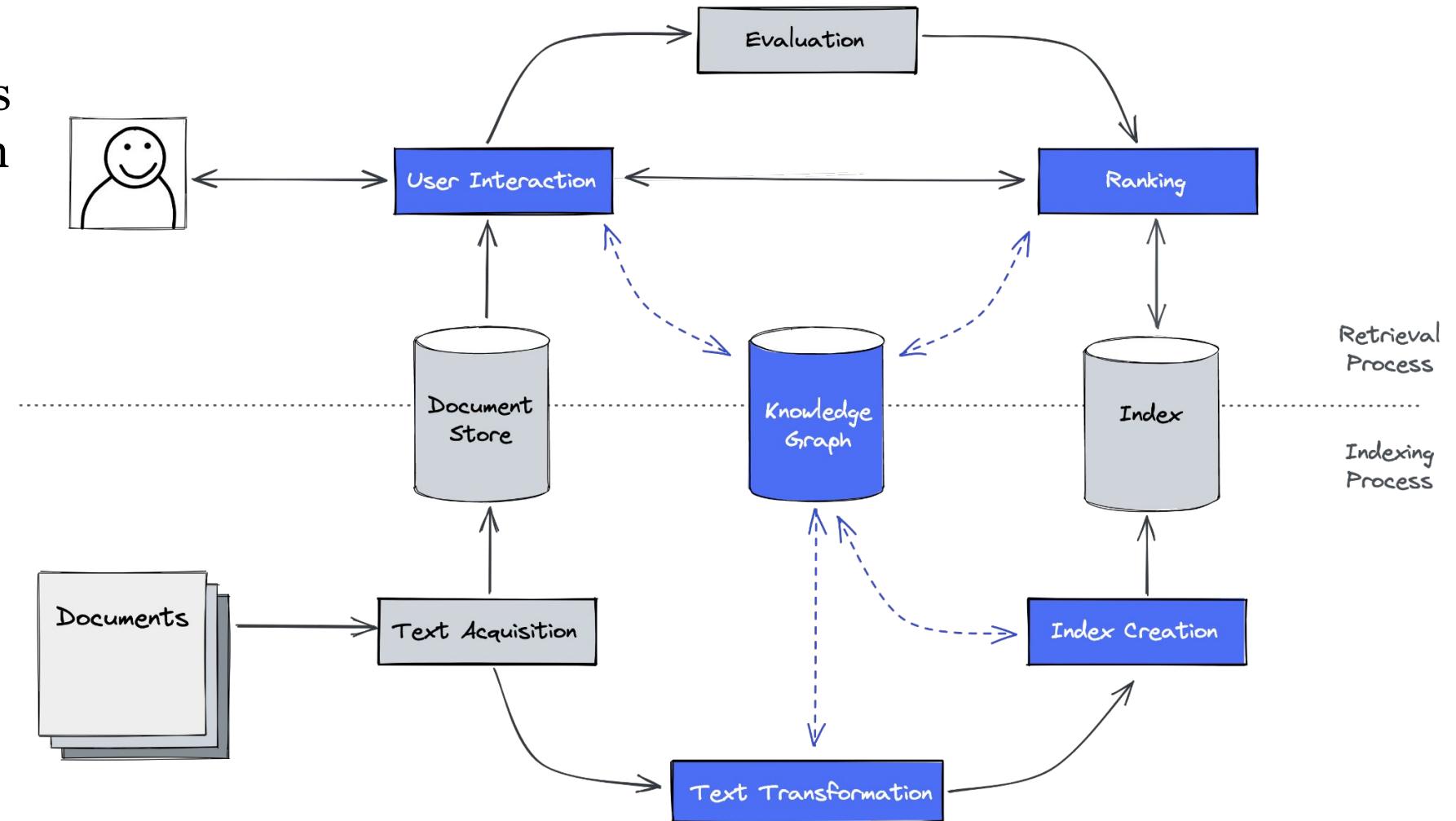
- Automatisierte Erschließungsverfahren erzeugen fehlerhafte Ergebnisse (machine learning = subsymbolische KI)
- Via Ontologien und Wissensgraphen (symbolische KI) können solche Fehler erkannt werden

externes Wissen
(symbolische KI)



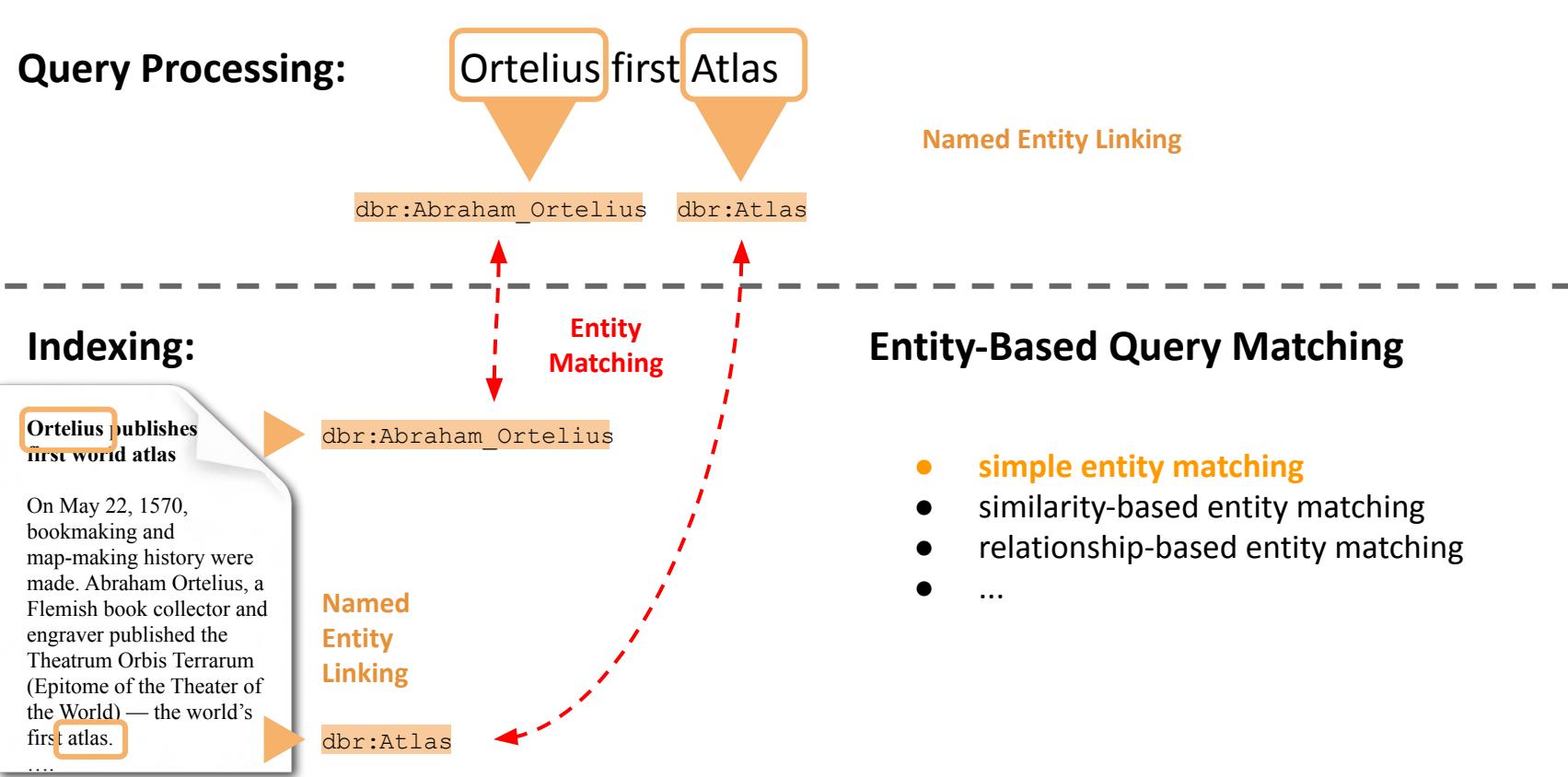
Semantisches Information Retrieval

- Knowledge Graph als zentrales Element im Retrieval Prozess
- steuert Indexerzeugung und -nutzung, Ranking und Benutzer Interaktion
- ermöglicht semantische und explorative Suche



Entitäten-zentrierte Suche

- Suche nach bedeutungstragenden Entitäten
- “Things, not strings”



Entitäten-zentrierte Suche

- Suche nach bedeutungstragenden Entitäten
- “Things, not strings”
- **Einbeziehung semantischer Ähnlichkeiten**

Query Processing:

Ortelius first Atlas

Named Entity Linking

dbr:Abraham_Ortelius dbr:Atlas

Indexing:

Mercator's atlas of Europe
Gerardus Mercator was perhaps the best-known mapmaker of all time. Rather than undertaking expeditions and mapping on the ground himself, however, he was an armchair cartographer.
....
Named Entity Linking

dbr:Atlas
dbr:Gerardus_Mercator

dbr:Abraham_Ortelius

Entity-Based Query Matching

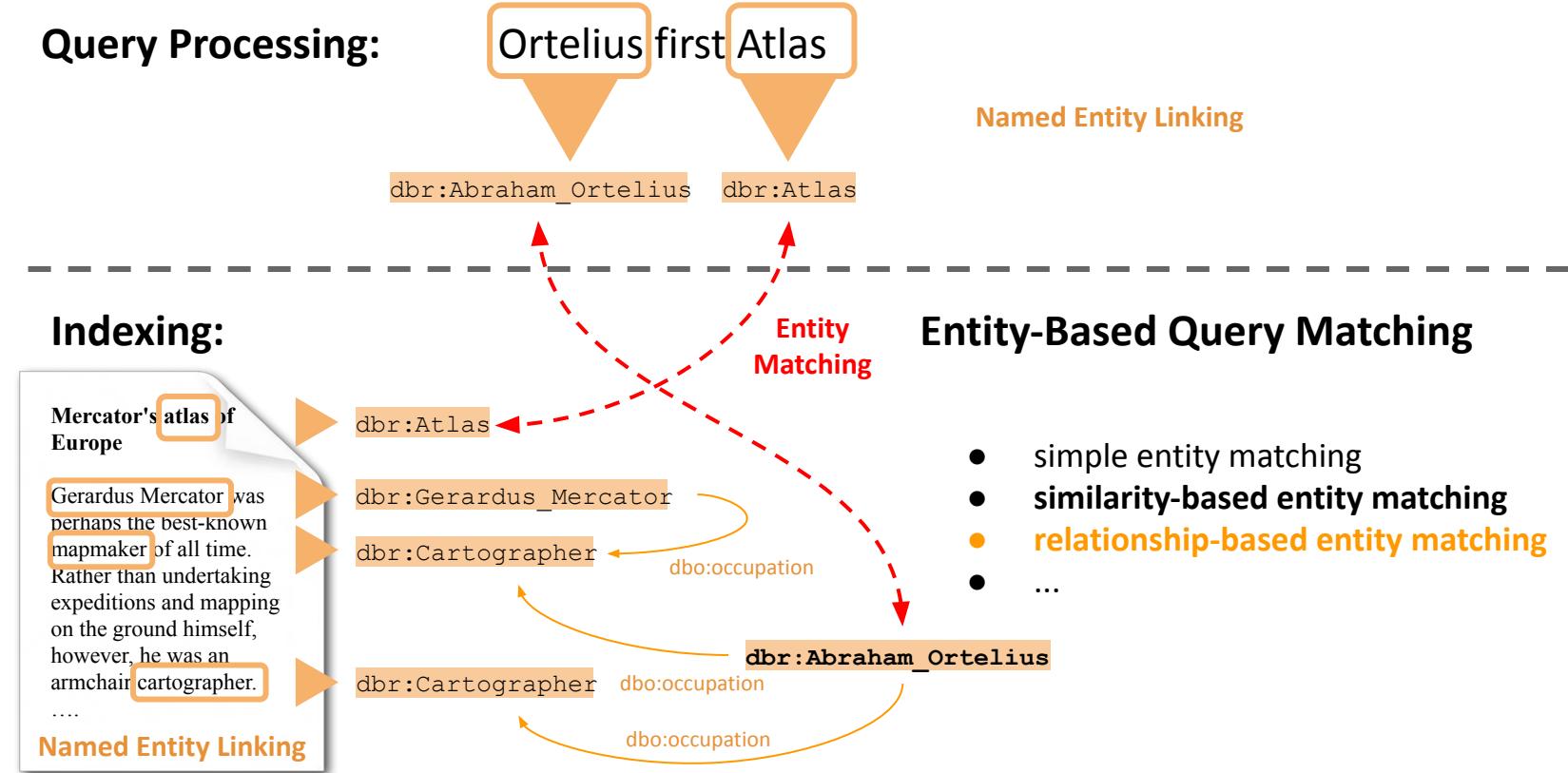
- simple entity matching
- **similarity-based entity matching**
- relationship-based entity matching
- ...

two entities are considered **semantically similar**,

- if they share property-value pairs
- if they share properties with similar values

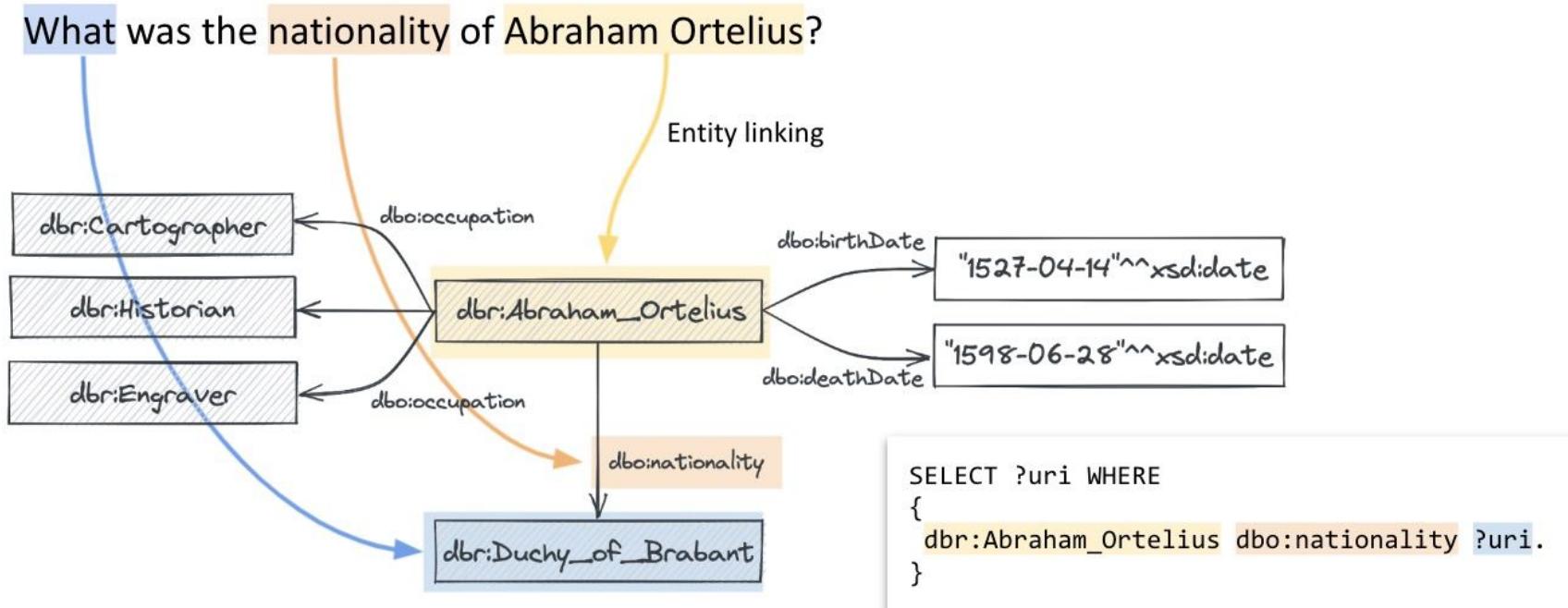
Entitäten-zentrierte Suche

- Suche nach bedeutungstragenden Entitäten
- “Things, not strings”
- Einbeziehung semantischer Ähnlichkeiten
- **Einbeziehung von semantischen Beziehungen und Relationen**



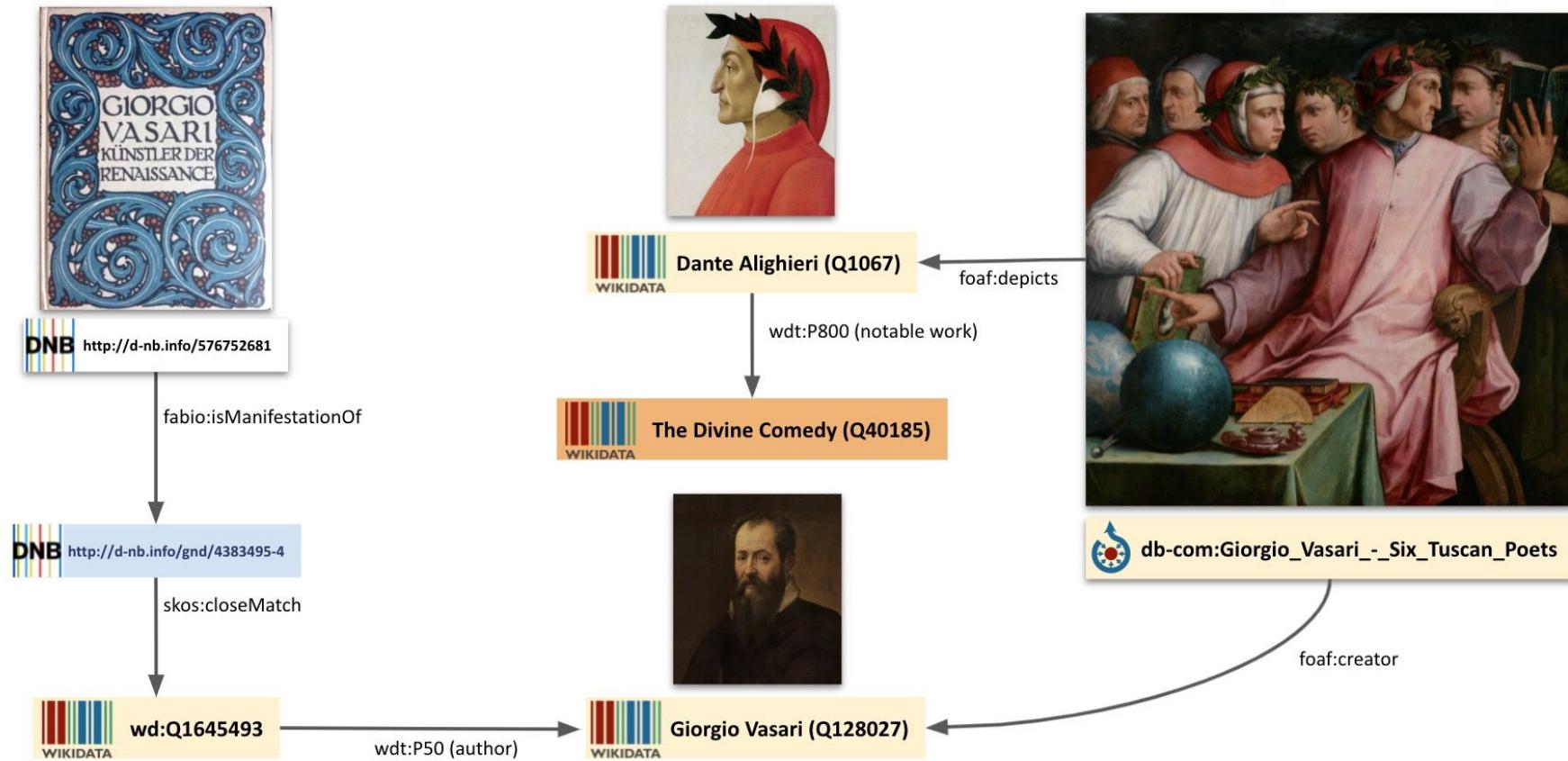
Question Answering

- Beantwortung sachbezogener Fragen
- semantisch Analyse (NLP, Entity Linking, Relation Extraction)
- Formulierung von Knowledge Graph Abfragen (SPARQL)



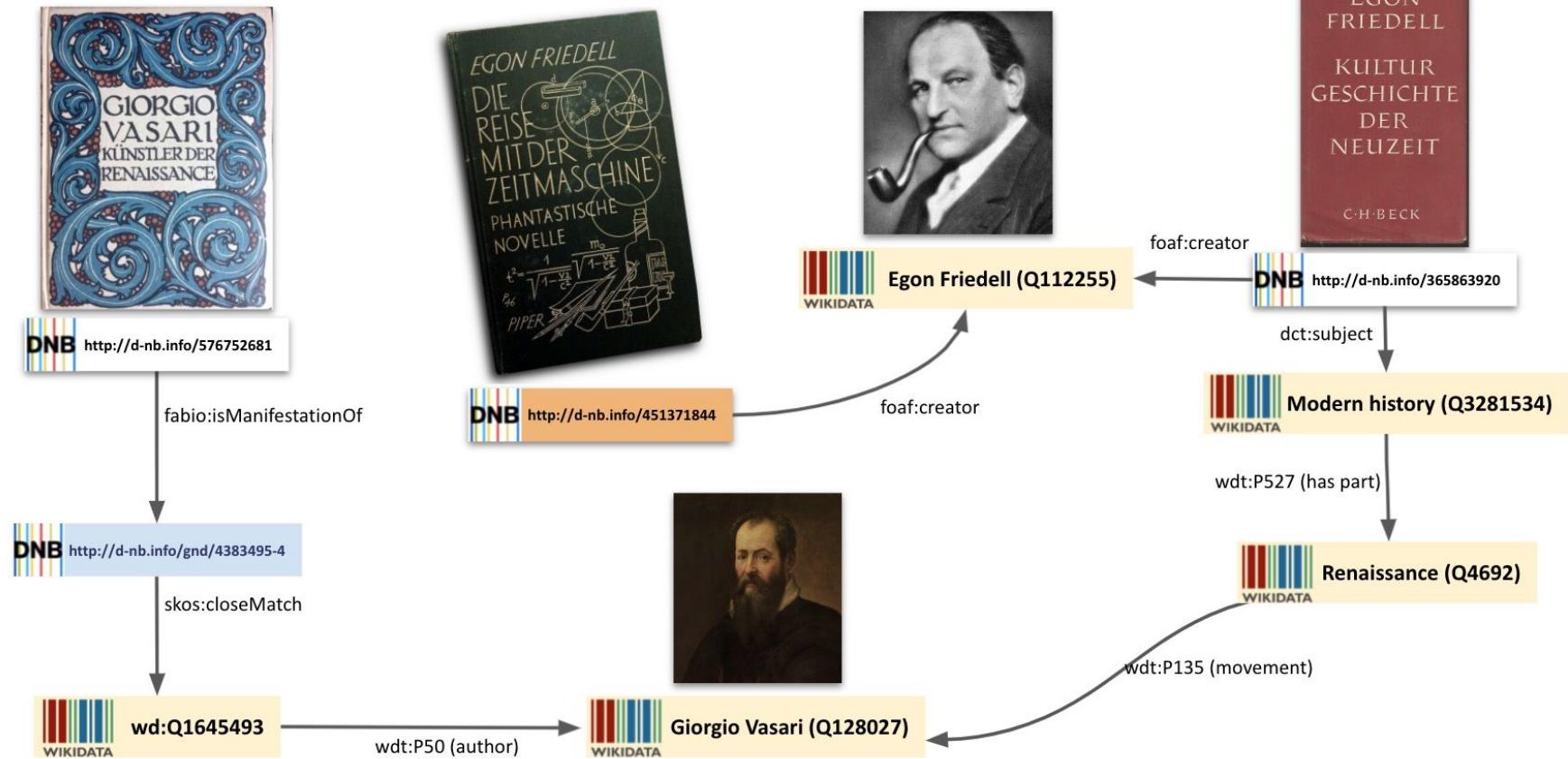
Explorative Suche

- Suche in unbekanntem Terrain
- unklares Suchziel
- komplexe Informationsbedürfnisse
- “Stöbern” statt zielgerichtetes Suchen



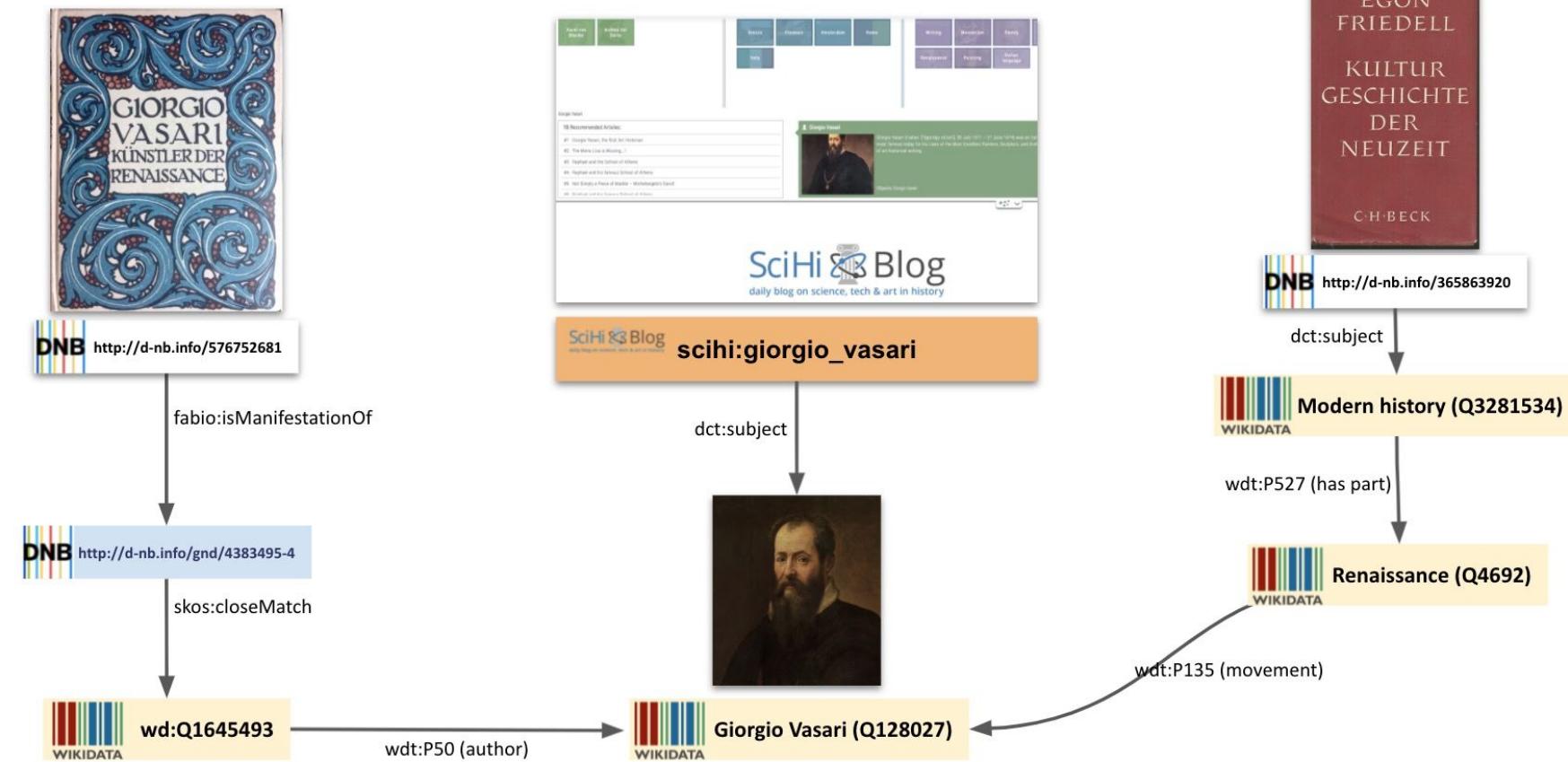
Explorative Suche

- überraschende Funde (“Serendipity”)



Explorative Suche

- überraschende Funde (“Serendipity”)
- Problem der Visualisierung (“Google-Effekt”)



Visualisierung und Exploration

<http://scih.org/>



Giorgio Vasari (1511-1574)

On July 30, 1511, Italian Renaissance painter, architect, writer and historian was born. He is best known today for his *Lives of the Most Eminent Painters* considered the ideological foundation of art-historical writing.

One of the Lesser Known Renaissance Artists

From all the great Renaissance artist, Giorgio Vasari might be one of the less this might be that although an artist of considerable repute, there were so many Renaissance artists that Vasari's art didn't stand out among them. But, on the Vasari's major merits that we know about so many Renaissance artists and authored and published the very first collection of biographies of famous artists

Born in Arezzo

Vasari was born in Arezzo, Tuscany. Recommended at an early age by his cousin Luca Signorelli, he became a pupil of Guglielmo da Marsiglia, a skillful painter of stained glass. Sent to Florence at the age of sixteen by Cardinal Silvio Passerini, where he enjoyed the friendship and patronage of the Medici

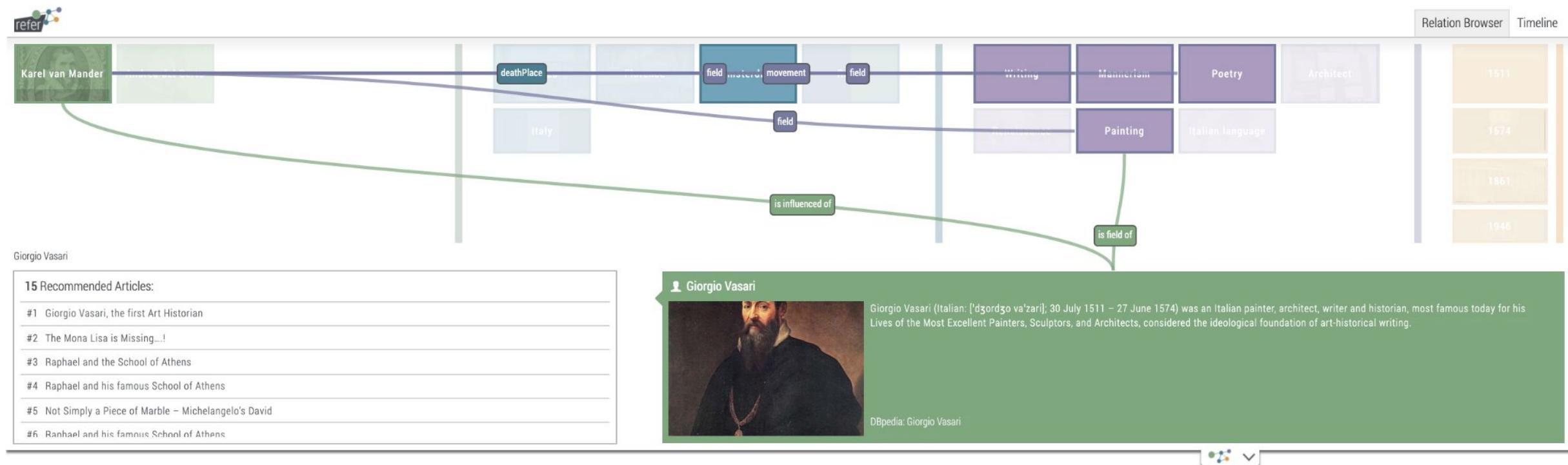
Giorgio Vasari	
Giorgio Vasari	(Italian: [dʒɔrdʒo va'zari]; 30 July 1511 – 27 June 1574)
was an Italian painter, architect, writer and historian, most famous	
birth year	1511
death year	1574
is influenced of	Andrea del Sarto
field	Architect
birth place	Arezzo
death place	Florence

ivgut
BW)



- Notwendige Voraussetzungen:
 - semantische Annotation
 - geeignete Werkzeuge zur Visualisierung

Visualisierung und Exploration



- z.B. via refer.cx WordPress PlugIn at <http://scihi.org/>

Vielen Dank!

Prof. Dr. Harald Sack
Bereichsleiter Information Service
Engineering
FIZ Karlsruhe
harald.sack@fiz-karlsruhe.de

Daniel Fähle M.A.
Referatsleiter IT, digitale Dienste
Landesarchiv Baden-Württemberg
Abteilung Zentrale Dienste
daniel.faehle@la-bw.de