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

Von Daniel Fähle und Harald Sack

Archive stellen bereits ein breites inhaltliches Angebot an interdisziplinär relevanten Forschungsdaten in Form von digitalisierten und originär digitalen Quellen samt zugehöriger Erschließungsinformationen zur Verfügung. Um innovative Forschungsmethoden einsetzen zu können, werden Schnittstellen und Werkzeuge benötigt, die eine Analyse, Anreicherung und Auswertung von *Archiv-Big-Data* ermöglichen. Ein Portfolio an verwendbaren Tools und Diensten gibt es schon heute – aber wie könnte ein dezidierter *digitaler Werkzeugkasten* für die Forschung mit Archivgut in Zukunft aussehen? Wie müssen die entsprechenden archivischen Angebote ausgebaut werden, um aktuelle und künftige Anforderungen der digitalen Forschung an Informationsinfrastrukturen erfüllen zu können? In diesem Beitrag werden dahingehend die Handlungsfelder Erweiterung des Datenangebots, Verbesserung der Datenqualität und Optimierung der Bereitstellungskanäle eingehender betrachtet. Es folgt ein Überblick zu einschlägigen Methoden und Werkzeugen sowie die Auseinandersetzung mit der Frage nach den erforderlichen Kompetenzen, um den *Werkzeugkasten* einsetzen zu können. Der zweite Teil des Beitrags veranschaulicht anhand konkreter Beispiele, welche Methoden und Tools bei der Extraktion von Wissen zum Einsatz kommen können und zeigt Anwendungsszenarien von Wissensgraphen auf.

Archivportal-D, 4Memory Data Space und die Bedeutung von Volltexten

*Historiker*innen brauchen für eine effiziente Forschung aggregierte Datensammlungen, die eine effiziente Suche und einen Zugriff auf Inhalte aus der Ferne erlauben.*[1] So hat Jörg Wettlaufer auf diesem Historikertag die Frage beantwortet, welche Services die digitale Geschichtswissenschaft von den Gedächtnis- und Infrastruktureinrichtungen benötigt. Die Archive sind dank der

[1] Jörg *Wettlaufer*: Abstract zum Vortrag Welche Services braucht die digitale Geschichtswissenschaft von Bibliotheken, Archiven, Museen und Datenzentren? 2021. https://www.historikertag.de/Muenchen2021/sektionen/gedaechtnisinstitutionen-in-der-digitalen-welt-bibliotheken-museen-archive-und-die-geschichtswissenschaft (aufgerufen am 05. 10. 2022).

Etablierung des Archivportals-D[2] als zentralem Nachweissystem bzw. *Data-Hub* hier bereits gut aufgestellt. Aufbauend auf den beachtlichen Fortschritten bei der Digitalisierung von Findmitteln und auch bereits erzielter substanzieller Ergebnisse bei der Digitalisierung der Quellen selbst, kann die historische Forschung auf ein umfangreiches Datenangebot gebündelt zugreifen. Seit der Freischaltung des Archivportals-D 2014 wächst dieses Angebot kontinuierlich, sodass inzwischen ein die Archivsparten übergreifender deutschlandweiter Forschungsdatenpool bereitsteht. Aktuell finden sich 23,7 Millionen Erschließungsdatensätze aus 220 Archiven (Stand: Oktober 2021). Erklärtes Ziel ist es, perspektivisch einen möglichst vollständigen Nachweis der Erschließungsleistungen deutscher Archive zu erreichen, damit der Forschung ein umfassendes Rechercheinstrument zu Archivgut zur Verfügung gestellt werden kann.[3] Über die Anbindung an die Deutsche Digitale Bibliothek (DDB)[4] ist überdies ein nachhaltiges Betriebskonzept für das Archivportal genauso gegeben wie konkrete Weiterentwicklungsperspektiven. Diese sind derzeit insbesondere bei der Etablierung themenbezogener Suchmöglichkeiten zu verorten. Dem bereits etablierten sachthematischen Zugang zu Quellen der Weimarer Republik[5] folgen analog Themenportale zu kurpfälzischen Urkunden[6] sowie zu dem Themenkomplex Wiedergutmachung von NS-Unrecht[7]. Im Rahmen dieser projektförmig realisierten Themenzugänge im Archivportal werden in Kooperation mit dem FIZ Karlsruhe bereits Anwendungsprototypen etwa zur (teil-) automatisierten Verschlagwortung oder Handschriftenerkennung entwickelt. Es erscheint naheliegend, bei der Erprobung und Entwicklung innovativer Verfahren zur Prozessierung von Archivdaten auf der Aggregationsebene des Archivportals anzusetzen, anstatt bei der heterogenen Landschaft der jeweiligen Archiv-Fachinformationssysteme (AFIS). Eine gezielte Weiterentwicklung des Archivportals-D sowie der assoziierten Dienste und Werkzeuge wird damit die Basisausstattung unseres *digitalen Werkzeugkastens*.

---

2 https://www.archivportal-d.de/ (aufgerufen am 05.10.2022). Allgemein zum Archivportal-D vgl. Daniel *Fähle* u.a.: Archivportal-D. Funktionalität, Entwicklungsperspektiven und Beteiligungsmöglichkeiten. In: Archivar 68, Heft 1 (2015) S. 10–19.

3 Im Unterschied zu den Ansätzen von „Schaufenster-Portalen", die in der Regel eine niedrigschwellige Präsentation von Highlight-Beständen bieten, aber für wissenschaftliche Recherchezwecke ungeeignet erscheinen.

4 https://www.deutsche-digitale-bibliothek.de/ (aufgerufen am 05.10.2022).

5 Aufbau einer Infrastruktur zur Implementierung sachthematischer Zugänge im Archivportal-D am Beispiel des Themenkomplexes *Weimarer Republik*, gefördert durch die DFG von 2017–2021, https://www.archivportal-d.de/themenportale/weimarer-republik (aufgerufen am 05.10.2022).

6 Vgl. Artikel *Kurpfälzisches Urkundenprojekt in vier Bundesländern gestartet* (2022) auf der Website des Landesarchivs Baden-Württemberg: https://www.landesarchiv-bw.de/de/aktuelles/nachrichten/74266 (aufgerufen am 05.10.2022).

7 Das Themenportal wurde vom Bundesministerium der Finanzen initiiert. Ziel ist es, einschlägige Aktenbestände des Bundes, der Länder und perspektivisch weiterer Stellen zusammenzuführen: https://www.archivportal-d.de/themenportale/wiedergutmachung (aufgerufen am 05.10.2022).

Auch beim Aufbau der Nationalen Forschungsdateninfrastruktur (NFDI), konkret dem bereits von Peter Haslinger in dieser Sektion vorgestellten Konsortium NFDI4Memory,[8] manifestiert sich die zentrale Bedeutung des Archivportals-D, indem es zu dessen Schlüsselangeboten (*Key Services)* zählt. Zu den Herausforderungen von 4Memory zählt die Schaffung eines gemeinsamen Datenraums (*Data Space*) – eines aus föderierten Infrastrukturen, Datenquellen und Diensten zu etablierenden digitalen Ökosystems der historisch arbeitenden Disziplinen: Wissenschaftlerinnen und Wissenschaftler sollen über Fach- und Repositoriengrenzen hinweg Quellen und assoziierte Daten finden können – Archivalien, wissenschaftliche Literatur aus Bibliotheken,

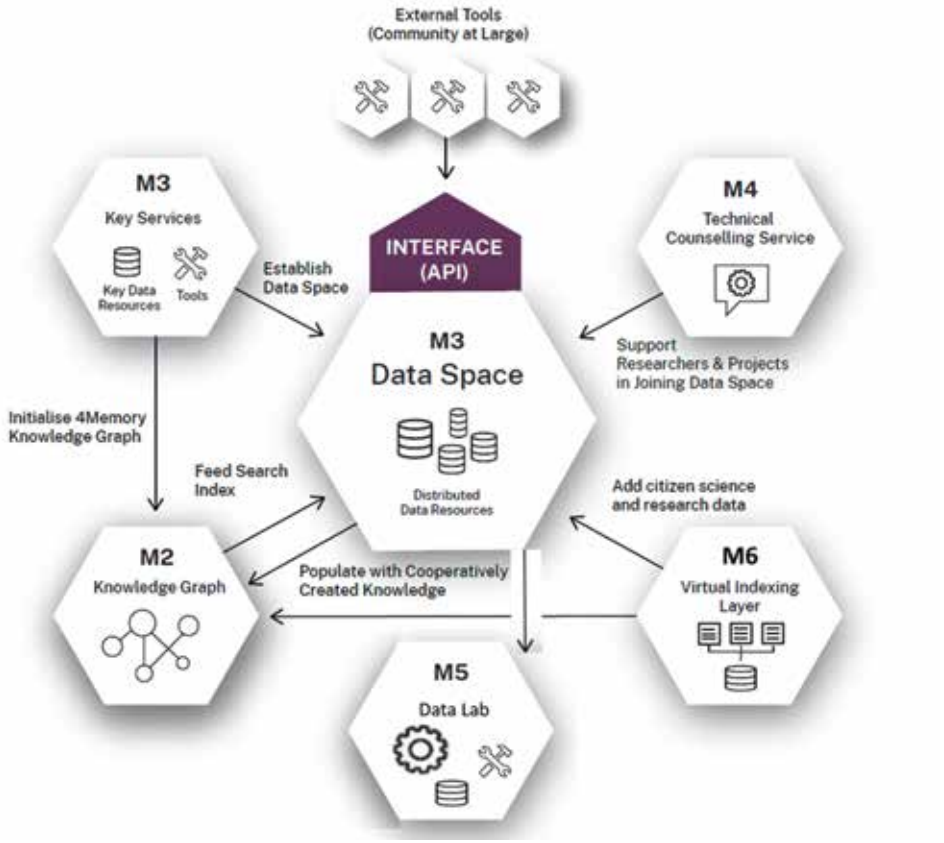

Abb. 1: Der *Data Space* von NFDI4Memory.

[8] Informationen zur NFDI-Konsortialinitiative der historisch arbeitenden Geisteswissenschaften: https://4memory.de/ (aufgerufen am 05.10.2022).

Informationen über Objekte aus Museen sowie weitere Forschungsdaten aus Projekten oder Spezialrepositorien. Ein überwölbender Wissensgraph (*Knowledge Graph*) wird neuartige Zugänge wie etwa eine explorative Recherche innerhalb des *Data Space* ermöglichen. Das ebenfalls im Rahmen von 4Memory zu realisierende *Data Lab* bietet Forschenden im Rückgriff auf den Datenraum eine Plattform für die Erforschung und Anwendung neuer Methoden des maschinellen Lernens, semantischer Annotation, fortgeschrittener OCR-Techniken etc.[9] Dem Netzwerk und Wirken des NFDI-Konsortiums 4Memory entlang des geplanten Arbeitsprogramms dürfte damit eine ganz wesentliche Rolle bei der Ausgestaltung des *digitalen Werkzeugkastens* zukommen.

Im Kontext von 4Memory, aber auch unabhängig davon, gewinnt ein Handlungsfeld für die Archive besondere Bedeutung: Es geht um die Frage, wie sich der Umfang der digital vorhandenen Erschließungsdaten substantiell und bedarfsgerecht erweitern lässt.[10] Die Auswertungsmöglichkeiten von Archivgut, häufig genug bereits schon die Auffindbarkeit, kranken an dem Grundproblem einer aus chronischem Personal- und Ressourcenmangel resultierenden flachen Erschließung. Nicht nur die digital Forschenden sind jedoch auf aussagekräftige Metadaten angewiesen, auch maschinelle Verfahren und Algorithmen benötigen als Ausgangspunkt eine umfangreichere Datenbasis. Um Abhilfe zu schaffen, muss der nächste große Ausbauschritt im Kontext einer forcierten Digitalisierung von Archivgut angegangen werden: die Generierung und Bereitstellung von Volltexten. Technologisch sind die einschlägigen Verfahren (OCR, HTR) zur optischen Zeichen- und Handschriftenerkennung inzwischen so weit vorangeschritten, dass deren Einsatz zunehmend alltagstauglich wird. Zweifellos dürfte die Erweiterung des Datenangebots um Volltexte, und seien es auch zunächst nur die Ergebnisse von *schmutziger OCR* für die Recherche, geradezu ein Quantensprung sein. Die Volltexte werden eine ideale Basis für qualifizierte Mining- und Auswertungsverfahren bieten, die die digitale Forschung dringend benötigt.

Datenqualität und Schnittstellen

Die angesprochene quantitative Ausweitung des Datenangebots von Archiven ist ein wichtiger Aspekt. Im Kontext unseres Themas erscheint jedoch die Art und Weise, wie die Daten bereitgestellt werden (müssen) mindestens als Frage von gleichem Gewicht. Welche Anforderungen ergeben sich an die Qualität der Primär- und der Erschließungsdaten? In welcher Form und Struktur müssen Daten vorliegen? Wie sollen Daten bereitgestellt werden? Eine wichtige Orien-

---

[9] Vgl. Fabian *Cremer* u. a.: Data meets history: A research data management strategy for the historically oriented humanities. In: Cultural Sovereignty beyond the Modern State. Hg. von Gregor *Feindt*, Bernhard *Gissibl* und Johannes *Paulmann* (Jahrbuch für Europäische Geschichte 21). Berlin/Boston 2021. S. 155–178.

[10] Daniel *Fähle*, Gerald *Maier* und Andreas *Neuburger*: Bereitstellung, Aufbereitung, Langzeitsicherung: Funktionen der Archive in der Forschungsdateninfrastruktur. In: Archivar 73 (2020) S. 13–18.

tierung bieten hier zunächst die FAIR-Prinzipien,[11] die als allgemeine Leitplanken zu den angesprochenen Fragen gelten können. Gute Datenqualität zeichnet sich demnach insbesondere durch standardisierte und maschinenlesbare (Meta-)Daten, die stringente Verwendung von Normdaten und kontrollierten Vokabularen, den Einsatz von gängigen Lizenzmodellen zur rechtssicheren Nachnutzung, persistente Identifikation als Basis von Zitierfähigkeit sowie die Implementierung von dokumentierten bzw. standardisierten Schnittstellen (Application Programming Interfaces – API) aus.

Werfen wir einen Blick auf den Status quo, so ist festzuhalten, dass einige Voraussetzungen durchaus schon erfüllt werden. Gerade die Entwicklungen rund um die übergreifenden Archivportale und die DDB haben eine tiefgreifende standardisierende Wirkung entfaltet. Mit der Etablierung der einschlägigen Lieferformate EAD(DDB) und METS/MODS, dem Einzug von UUIDs in archivische Fachinformationssysteme, einem bereits recht weitgehenden Rückgriff auf die verbreiteten Creative-Commons-Lizenzen[12] sowie einer inzwischen bemerkenswert hohen Dynamik bei der Verwendung und auch Produktion von Normdaten ist die Archivsparte auf einem guten Weg.

Defizite lassen sich vor allem bei der Bereitstellung von Daten ausmachen, wenn z. B. Images nur in unzureichender Qualität angeboten werden oder über den Einzeldownload hinaus keine Schnittstellen für automatisierte Datenabrufe existieren. Dies behindert die digitale Forschung mit Archivdaten erheblich. Denn nur über geeignete Wege des Datenabrufs lassen sich anwenderseitig individuell relevante Datensets bzw. Korpora für spezifische Forschungsfragen zusammenstellen und für Verfahren zur Extraktion von Informationen aus Images (Computer Vision, Text- und Mustererkennung) – wie sie im zweiten Teil dieses Beitrags von Harald Sack noch eingehend vorgestellt werden – heranziehen. Mit Blick auf die Datenbereitstellung lässt sich zumindest festhalten, dass über die DDB-API[13] die im Portal aggregierten Erschließungsinformationen (Metadaten) in verschiedenen Formaten sowie Derivate der zugehörigen Mediendateien (*Binaries*) maschinell abgerufen werden können. Doch die Bedarfe der digitalen Forschung sind inzwischen sehr viel weitgehender. Auf die Frage, in welcher Form genau die Daten benötigt werden, wird immer häufiger geantwortet: Zum einen am besten über Exportmöglichkeiten als semantisch codierte Metadaten, z. B. in RDF, zum anderen indem v. a. die Images über IIIF-Schnittstellen angeboten werden.[14] Beide Anforderungen sind durchaus voraussetzungsreich. Betrachten wir zunächst den Bedarf an semantischen Daten, dann erscheint die Übersetzung bzw. Transformation von beispielsweise EAD-Daten in RDF-XML als äußerst mühevolles Unterfangen. Erst wenn die Erschließungsinformationen bereits in den Erfassungssystemen semantisch modelliert vorliegen, sind hier Fortschritte zu erwarten. Die Entwicklung des neuen internationalen Erschließungsstandards Records in Contexts (RiC)[15] weist jedenfalls klar in diese Richtung. Archiv-

---

11 FAIR als Akronym für *findable*, *accessible*, *interoperable* und *reusable*.

12 https://creativecommons.org/licenses/ (aufgerufen am 05. 10. 2022).

13 https://labs.deutsche-digitale-bibliothek.de/app/ddbapi/ (aufgerufen am 05. 10. 2022).

14 https://iiif.io/ (aufgerufen am 05. 10. 2022).

15 Vgl. https://www.ica.org/en/records-in-contexts-conceptual-model (aufgerufen am 05. 10. 2022).

daten sollen demnach nicht mehr ausschließlich der hierarchischen Struktur ihrer Provenienz-Zugehörigkeit entsprechen, sondern können diverse Relationen in unterschiedlichen Kontexten abbilden. Dies erfordert ein semantisches Datenmodell, wie es die RiC-Ontologie in der derzeit diskutierten Fassung widerspiegelt. Einer der ersten Versuche, den neuen Erschließungsstandard in einer Archivsoftware zu implementieren, wird im Rahmen des DFG-Projekts EEZU[16] unternommen. Dieses Vorhaben hat zum Ziel, eine einfache Erschließungssoftware für kleinere Archive zu realisieren. Hierdurch ist gewährleistet, dass nicht nur die großen Archiveinrichtungen mit den oben skizzierten Anforderungen Schritt halten können. EEZU wird auch IIIF-Schnittstellen bieten. Dieses API-Framework umfasst drei Schnittstellen – für Metadaten, Images und Präsentation – und ist auf gutem Wege, insbesondere bei der Bereitstellung von Digitalisaten zum internationalen Standard und mithin Garanten der vielgeforderten Interoperabilität zu werden. IIIF zeichnet sich durch eine sehr hohe Performanz aus, ermöglicht vor allem aber die anwenderseitige bzw. dezentrale Zusammenführung von Images aus unterschiedlichen Quellen und Ursprungssystemen in einem Viewer. Damit ergeben sich eine Vielzahl an innovativen Nutzungsmöglichkeiten in entsprechenden Applikationen, die über die bloße Anzeige weit hinausreichen und z. B. das Vergleichen, Annotieren, Transkribieren und Georeferenzieren ermöglichen.

Anwendungen und Kompetenzen

Über die skizzierten Handlungsfelder 1) Erweiterung der digitalen Datengrundlage für die Forschung und 2) Ausbau von Archiv-Fachinformationssystemen (AFIS) sowie Archivportalen zu zeitgemäßen Informationsinfrastrukturen werden die Voraussetzungen geschaffen, damit ein *digitaler Werkzeugkasten* überhaupt zum Einsatz kommen kann. Werfen wir nun einen genaueren Blick auf die Werkzeuge und Methoden, die zur Verfügung stehen, so lassen sich diese grobschematisch in verschiedene Anwendungsbereiche gruppieren:

- Analyse von Daten, die im Grunde jeder weiteren Verarbeitung vorausgehen muss
- Bereinigung und Strukturierung, um benötigte Datensets ggf. für die weitere Prozessierung aufzubereiten
- Extraktion, um notwendige Informationen aus den digitalen Daten zu gewinnen (Texterkennung, Bild- und Objekterkennung, Named Entity Recognition)
- Erweiterung und Anreicherung, um mittels geeigneter Verfahren zusätzliche Informationen zu gewinnen und Datenbasis wie grundlage zu verbessern (z. B. Entity Linking, Annotation, Georeferenzierung)
- Suche und Präsentation, um Daten und Informationen darzustellen, recherchierbar zu gestalten und auszuwerten (semantisches Information Retrieval, Visualisierung, Geoinformationssysteme)

[16] Vgl. https://www.fiz-karlsruhe.de/de/forschung/eezu (aufgerufen am 05. 10. 2022).

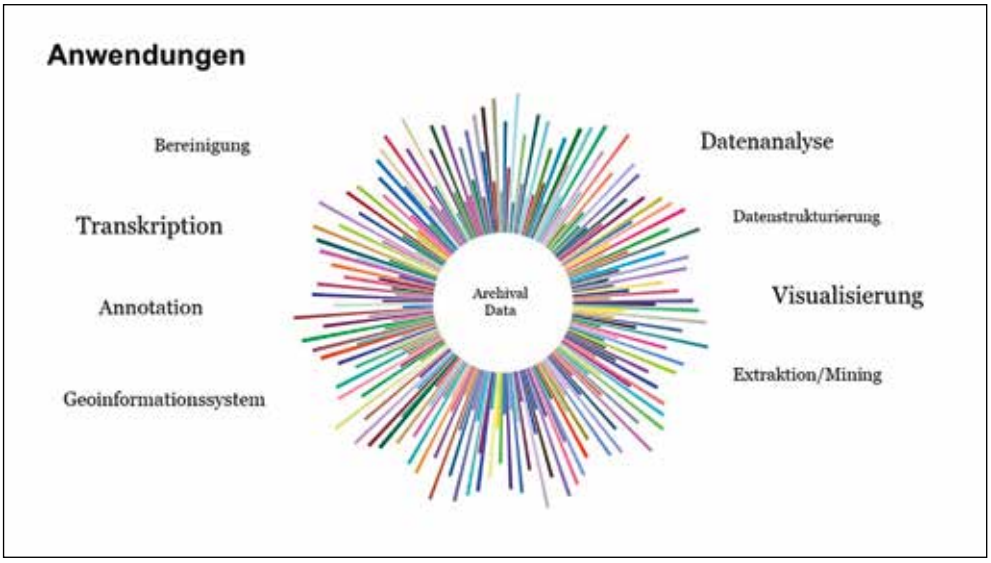

Abb. 2: Anwendungsbereiche.

Die relevanten Methoden und Tools sind dabei in der Regel nicht archivischer Provenienz, sondern entstammen ganz unterschiedlichen Kontexten und Disziplinen, z. B. der Informatik, Informationswissenschaft und Computerlinguistik. Nutzbar gemacht werden diese bisher vor allem von den Digital Humanities oder etwa im Rahmen der Digital History. Wie gelangen wir aber zu passgenau anwendbaren Werkzeugen für digitale Archivdaten? Indem die Bedarfe der digitalen Forschung zunächst ermittelt, dann geeignete Tools identifiziert und für den archivspezifischen Einsatz erprobt, ggf. angepasst und schließlich bereitgestellt werden. Die Archive allein, auch die leistungsfähigsten, werden das nicht in vollem Umfang leisten können, auch wenn es entsprechende Ansätze wie Forschungs- oder Datenlabore gibt.[17] Es ist zwingend nötig, sich mit Vertreterinnen und Vertretern der Forschung, der Digital Humanities und Datenwissenschaften bzw. der Informatik zu vernetzen. Auch mit Blick auf die Pflege, Weiterentwicklung und Interoperabilität des Werkzeugportfolios[18] bedarf es des Zusammenwirkens, sei es in kleineren Kooperationsprojekten oder in übergreifenden Konstellationen wie den NFDI-Konsortien.

Eine zentrale Voraussetzung für den Einsatz des *digitalen Werkzeugkastens* darf nicht unerwähnt bleiben: Neben dem unerlässlichen Infrastrukturaufbau und der damit verbundenen Aneignung von Kompetenzen auf Archivseite mit Blick auf die adäquate Bereitstellung von

[17] Das FDMLab des Landesarchivs Baden-Württemberg erprobt zum Beispiel Data-Science-Methoden und Techniken für den Einsatz im Archiv: https://fdmlab.ib2m.de/projekt/ (aufgerufen am 05. 10. 2022).

[18] Charakteristikum dieser Interoperabilität ist beispielsweise, dass die von einem Tool erzeugten Dateien auch in einem anderen Tool weiterverarbeitet werden können.

Archivdaten, werden auf der anderen Seite auch die Forscherinnen und Forscher die Kompetenz benötigen, mit den bereitgestellten Daten umzugehen und digitale Werkzeuge anwenden zu können. Hierzu wird es erforderlich sein, *Data Literacy*[19] als künftige Schlüsselkompetenz für Historikerinnen und Historiker zu stärken, vielleicht sogar als neue Hilfswissenschaft anzuerkennen. Archive können hier zwar einen Beitrag z. B. über Dokumentationen und Schulungsformate leisten, letztlich werden aber substanzielle Fortschritte nur durch entsprechende Lehrangebote an den Hochschulen zu erreichen sein.

Wissensextraktion von historischen Forschungsdaten

Automatisierte Verfahren zur Wissensextraktion aus historischen Forschungsdaten, z. B. Akten und Archivdokumenten, werden meist in Form einer Verarbeitungspipeline implementiert. Vorverarbeitung, Datenaufbereitung, unterschiedliche Analyseverfahren und gegebenenfalls Nachverarbeitung werden schrittweise und vielfach aufeinander aufbauend in konsekutive Einzelschritte unterteilt, um so eine effizientere Parallelisierung der Verarbeitung von *Massendaten* zu gewährleisten. Die folgenden für die Wissensextraktion aus historischen Forschungsdaten relevanten Verfahren und Werkzeuge sollen hier kurz vorgestellt werden:

- *Optical Character Recognition* (OCR)
- visuelle Analyseverfahren
- Sprachmodelle (Large Language Models, Foundation Models)
- Wissensrepräsentation von Archivdaten, und -prozessen
- Ontologien, Wissensgraphen und Normdaten

Optical Character Recognition

Am Beginn der eigentlichen Wissensextraktion aus Akten und Archivdokumenten steht zunächst die Digitalisierung, i. e. die Überführung der ursprünglich meist analogen Dokumente in eine digitale Form, z. B. durch Fotografie oder Scan. Dabei liegen die Dokumente zunächst in Form unstrukturierter Bilddaten vor, die im Folgeschritt verschiedenen visuellen Analysen unterzogen werden können. Textinhalte werden dabei über *Optical Character Recognition* (OCR) analysiert und transkribiert. Meist erfordern die untersuchten Akten und Archivdokumente aufgrund ihrer Ausprägung und inhaltlichen Darstellung spezielle Herangehensweisen, um handelsübliche OCR-Verfahren erfolgreich einsetzen zu können. Formulare mit unterschiedlichem Layout, teils mit Fließtexten, Tabellen oder strukturierten Datenerfassungsfeldern erfordern spezielle Verfah-

[19] Definition nach Chantel Ridsdale: […] *the ability to collect, manage, evaluate, and apply data in a critical manner*. Vgl. Chantel *Ridsdale* u. a.: Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report. Halifax 2015.

ren zur Layout-Analyse, um zusammenhängende Textregionen korrekt erfassen zu können. Unterschiedliche gedruckte Schriftarten und -typen auf den Dokumenten liegen oft zusammen mit maschinengeschriebenen Texten und handschriftlichen Passagen ergänzt vor. Dazu können noch Stempel, Unterschriften, Bemerkungen und Sichtvermerke auftreten.

Zuverlässige OCR erfordert speziell auf Archivbedürfnisse angepasste und trainierte Modelle. Dabei ist es allgemein unwahrscheinlich, dass ein einzelnes trainiertes Modell den Anforderungen der verschiedenen Inhaltsvarianten gerecht werden kann. Heterogene Archivdokumente lassen sich nach vorgegebenen inhaltlichen Kriterien automatisiert gruppieren (*clustern*) beziehungsweise in unterschiedliche Inhaltstypen (gedruckter Text, Handschriften, Stempel, etc.) separieren, die dann getrennt einer Weiterverarbeitung durch ein speziell angepasstes Modell zugeführt werden können, um so die Qualität der erzielten Transkriptionsergebnisse zu verbessern.[20]

Visuelle Analyseverfahren

Deep Learning Verfahren haben die visuelle Analyse von Bildinhalten in der vergangenen Dekade enorm verbessern können. Auf diese Weise lassen sich Objekte in historischen Bildern und Fotografien identifizieren oder bildbeschreibende Zusammenfassungen in Textform generieren. Eine potentielle Fehlerquelle liegt dabei in der den Bilderkennungsverfahren zugrunde liegenden Trainingsdaten, die vielfach aus zeitgenössischen, webbasierten Korpora gewonnen werden und damit ein diachrones Ungleichgewicht (Bias) aufweisen. Die mit zeitgenössischen Daten vortrainierten Modelle werden im Archivkontext auf historisches Bildmaterial angewandt und resultieren oft in diachronen Fehlklassifikationen, wie der fehlerhaften Identifikation von modernen elektronischen Geräten oder Sportgeräten in historischem Bildgut.[21]

Abb. 3: Diachrone Fehlidentifikation eines Skateboards in einer mittelalterlichen Illustration.[22]

[20] Vgl. Mahsa *Vafaie* u. a.: Handwritten and Printed Text Identification in Identification in Historical Archival Documents. In: Proceedings of Archiving Conference 19 (2022) S. 15–20, https://doi.org/10.2352/issn.2168-3204.2022.19.1.4 (aufgerufen am 07. 12. 2022).

[21] Vgl. Harald *Sack*: Ein Skateboard für den Papst oder Warum es maschinelles Lernen ohne Semantik so schwer hat. Netzwerk maschinelle Verfahren in der Erschließung, Deutsche Nationalbibliothek, Frankfurt, 11. Oktober 2019.

[22] Heinrich IV (1050–1106) bittet Markgräfin Mathilde von Tuszien und seinen Taufpaten Abt Hugo von Cluny um Vermittlung; Vita Mathildis des Donizio, um 1115. Vatikanstadt, Bibliotheca Apostolica Vaticana, Ms. Vat. lat. 4922, fol. 49 v.

Sprachmodelle (*Large Language Models*, *Foundation Models*)

Sprachmodelle ermöglichen die Repräsentation natürlicher Sprache in einem statistischen Modell. Statistische Kookkurrenz von Wörtern, Wortfolgen und ganzen Sätzen erlaubt implizite Rückschlüsse auf semantische Ähnlichkeit und Bezugnahme. Worte, Wortfolgen, Sätze, bis hin zu ganzen Dokumenten lassen sich als Vektor in einem niedrigdimensionalen Vektorraum darstellen, wobei semantische Ähnlichkeiten und Bezugnahme über einfache Vektoroperationen zugänglich gemacht werden.[23] Über sogenanntes *Transfer Learning* werden moderne Sprachmodelle mit Hilfe gigantischer, oft aus dem World Wide Web gewonnener, Textkorpora ohne direkte Supervision trainiert (*Self-supervised Pre-Training*) und lassen sich auf einfache Weise an spezielle Aufgabenstellungen, z. B. Textklassifikation, *Keyword Extraction*, Textzusammenfassung, Entitätenerkennung und Identifikation, oder die Beantwortung sachbezogener Fragen anpassen (*Finetuning*).[24] In den vergangenen fünf Jahren haben die verfügbaren Sprachmodelle zunehmend an Komplexität und Leistungsfähigkeit gewonnen. Jedoch bleibt zu berücksichtigen, dass – trotz beeindruckender Ergebnisse – eine explizite Kontrolle der damit erzielten Ergebnisse vollständig automatisiert aktuell nicht möglich ist und oft den Menschen als Kontrollinstanz erfordert.

Wissensrepräsentation von Archivdaten und -prozessen

Die im vorangegangenen Absatz behandelten Sprachmodelle repräsentieren das in ihnen vorhandene Wissen nur in impliziter, subsymbolischer Form. Diese vom *Deep Learning* geprägte Form der Wissensrepräsentation wird in der künstlichen Intelligenz durch symbolische Wissensrepräsentationstechniken, z. B. symbolische Logiken, Inferenz- und Schlussfolgerungstechniken und Ontologien ergänzt. Während subsymbolische KI-Techniken im Bereich des (maschinellen) Lernens brillieren, sind ihre Fähigkeiten zur Abstraktion sehr beschränkt. Umgekehrt verhält es sich mit symbolischen KI-Techniken. Ontologien als explizite Wissensrepräsentationen basieren auf unterschiedlich ausdrucksstarken mathematischen Logiken und ermöglichen das Schlussfolgern von implizit verborgenem Wissen und damit die Generierung neuen Wissens. Um diese Möglichkeiten z. B. zum Zweck der Datenintegration im Archivbereich nutzen zu können, ist es erforderlich, sowohl die im Archivgut enthaltenen Informationen als auch die Archivorganisation selbst sowie die darauf bezogenen Archivprozesse mit Hilfe von Ontologien und Wissensgraphen zu

---

[23] Vgl. Jacob *Devlin* u. a.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (2019), Human Language Technologies, Band 1 (Long and Short Papers), Minneapolis, Minnesota, Association for Computational Linguistics. S. 4171 – 4186.

[24] Vgl. Tom B. *Brown* u. a.: Language models are few-shot learners. In: Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS'20). Curran Associates Inc., Red Hook, New York 2020. Artikel 159. S. 1877 – 1901.

repräsentieren.[25] Die bereits erwähnte *Records in Context Ontologie* (RiC-O) oder die zur Repräsentation von Provenienzinformationen verwendete PROV Ontologie (PROV-O)[26] bieten schon erste Ansätze zur Wissensrepräsentation von Archivdaten. Dynamische Prozesse und Vorgänge im Archiv, wie Veränderungen und Anpassungen der Systematik, können über die *Archive Dynamics Ontology* (ArDO)[27] repräsentiert werden.

Ontologien, Wissensgraphen und Normdaten

Um Archivdokumente explizit mit semantischen Informationen zu annotieren, müssen semantische Entitäten, z. B. Personen, Orte und Ereignisse, sowie zugehörige Informationen in diesen Dokumenten erkannt und identifiziert werden. Über die Verknüpfung dieser Informationen mit fachbezogenen Ontologien und Normdaten entstehen Wissensgraphen, die die im Archivgut enthaltenenInformationen semantisch repräsentieren und damit für moderne Such- und Explorationstechniken zugänglich machen (siehe auch Abschnitt *Angewandte Wissensgraphen*). Ontologien formalisieren die Bedeutung der durch sie repräsentierten Konzepte und Relationen. Wissensgraphen verknüpfen Bedeutungsinhalte auf der Basis dieser Ontologien. Normdaten bilden dabei eine traditionelle Referenzbasis in Form kontrollierter Vokabulare und strukturierter Daten. Im Zusammenspiel miteinander verknüpft bilden Ontologien, Wissensgraphen und Normdaten die Basis zur effizienten Umsetzung der FAIR-Prinzipien für Archivdaten und historische Forschungsdaten.

Angewandte Wissensgraphen

In den folgenden Abschnitten werden beispielhaft Anwendungsszenarien für historische Forschungsdaten basierend auf dem Einsatz von Wissensgraphen vorgestellt.

---

25 Vgl. Mahsa *Vafaie* u. a.: Modelling Archival Hierarchies in Practice: Key Aspects and Lessons Learned. In Proceedings of the 6th International Workshop on Computational History (Histoinformatics2021), co-located with JCDL2021. CEUR workshop proceedings, Band 2981 (2021), http://ceur-ws.org/Vol-2981/paper6.pdf (aufgerufen am 07. 12. 2022). – Vgl. Oleksandra *Bruns* u. a.: Towards a Representation of Temporal Data in Archival Records: Use Cases and Requirements. In: Proceedings of the International Workshop on Archives and Linked Data (LinkedArchives), co-located with the 25th International Conference on Theory and Practice of Digital Libraries (TPDL 2021), CEUR workshop proceedings 3019. S. 128 – 134, http://ceur-ws.org/Vol-3019/ (aufgerufen am 07. 12. 2022).

26 PROV-O, the PROV Ontology, https://www.w3.org/TR/prov-o/ (aufgerufen am 07. 12. 2022).

27 Vgl. Oleksandra *Vsesviatska* u. a.: ArDO: An Ontology to Describe the Dynamics of Multimedia Archival Records. In: Proceedings of the 36th ACM/SIGAPP Symposium On Applied Computing (ACM SAC 2021). S. 1855 – 1863. DOI: https://doi.org/10.1145/3412841.3442057 (aufgerufen am 07. 12. 2022).

Zusammenspiel symbolischer und subsymbolischer KI

Wie bereits erläutert werden die oft beeindruckenden Ergebnisse moderner *Deep Learning*-Verfahren geschmälert, bedingt durch ihre mangelnde Zuverlässigkeit, die ihren Einsatz im wissenschaftlichen Kontext oft fragwürdig erscheinen lassen. Obwohl *Deep Learning* Modelle innerhalb eng fokussierter, spezieller Anwendungsbereiche, wie der zuverlässigen Identifikation von Krebszellen auf der Basis visueller Analyse, den Menschen in seiner Urteilsfähigkeit übertreffen,[28] scheitern diese oft an allgemeineren oder weiter gefassten Aufgabenstellungen, z. B. der korrekten Beschreibung von Bildinhalten in historischen Gemälden. Um inhaltlichen Fehlern oder Fehlklassifikationen vorzubeugen, können diese subsymbolischen KI-Technologien durch Wissensrepräsentationsverfahren der klassischen symbolischen KI ergänzt werden. Mit Hilfe von Wissensrepräsentationen in Form von Ontologien oder Regeln können inhaltliche logische Inkonsistenzen aufgedeckt und die betreffenden Ergebnisse in Zweifel gezogen werden. Ein Beispiel dazu liefert die im folgenden Abschnitt vorgestellte Überprüfung von Objektidentifikationsergebnissen aus der visuellen Analyse. Neben dem Aufdecken logischer Inkonsistenzen in Klassifikationsergebnissen können von *Deep Learning*-Verfahren generierte Texte auch mit bekannten Fakten aus Wissensgraphen verglichen werden. Widerspricht ein generierter Text den in einem Wissensgraphen repräsentieren Fakten, heißt das nicht unbedingt, dass diese Fakten zwangsläufig falsch sind. Diese Beurteilung könnte in einem Folgeschritt durch menschliche Experten und Expertinnen erfolgen, die gegebenenfalls den zugrunde liegenden Wissensgraphen inhaltlich ergänzen könnten.

Korrektur visueller Analyseverfahren

Ausgehend von der in Abb. 3 dargestellten Fehlidentifikation eines Skateboards in einer mittelalterlichen Handschrift, können zur Überprüfung der erzielten Objektergebnisse die zum Bild zugehörigen Metadaten in Kombination mit einer öffentlich zugänglichen Wissensbasis eingesetzt werden. Aus den Metadaten des betreffenden Bildes konnte das Entstehungsdatum mit dem Jahr 1115 ermittelt werden. Die erkannten Objekte im Bild, z. B. „Person“ und „Skateboard“, werden über *Entity Linking* mit den korrespondierenden Entitäten aus der Wissensbasis Wikidata verknüpft. Aus den darin vorhandenen Fakten zu den identifizierten Entitäten kann ermittelt werden, dass Skateboards seit den 1950er Jahren bekannt sind (vgl. Abb. 4). Eine entsprechende einfache logische Regel zur Überprüfung der in einem historischen Bild identifizierten Objekte sollte daher prüfen, ob die mit den identifizierten Objekten in einer externen Wissensbasis referenzierten Entstehungs-, Entdeckungs- oder Lebenszeiten mit der Entstehungszeit des zugrunde

[28] Vgl. Diego *Ardila* u. a.: End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography. In: Nature Medicine 25 (2019). S. 954–961, https://doi.org/10.1038/s41591-019-0447-x (aufgerufen am 07.12.2022).

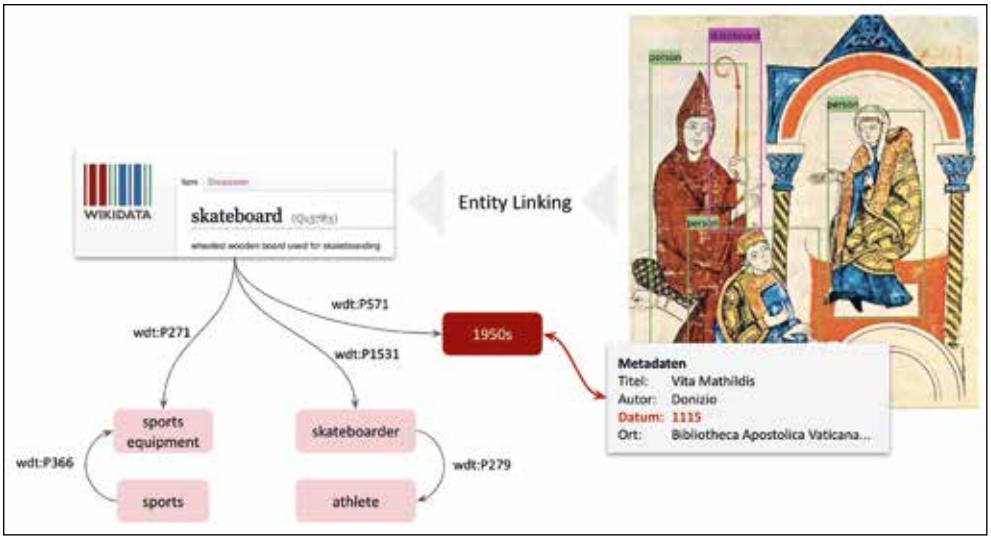

Abb. 4: Einsatz symbolischer Wissensrepräsentation zur Identifikation von logischen Inkonsistenzen in den Ergebnissen einer auf *Deep Learning* basierten visuellen Bildanalyse.

liegenden Bildes im Einklang steht; d. h. dass kein erkanntes Bildobjekt erst nach der durch die Metadaten belegten Entstehungszeit des Bildes existieren kann. In formalisierter Form lässt sich diese Regel durch einen sogenannten *Reasoner*, eine Software zum allgemeinen Ziehen logischer Schlussfolgerungen, überprüfen[29].

Semantische Suche

Moderne Informationssysteme bieten meist eine textbasierte Suche über den in ihnen repräsentierten Dokumentenbestand an. In den meisten Fällen basiert diese Suche auf einem syntaktischen Vergleich der in den Dokumentenmetadaten oder -inhalten vorhandenen Texte mit einer vom Benutzenden eingegebenen textuellen Suchphrase. Semantische Ähnlichkeiten oder Beziehungen zwischen den in den Texten und Suchphrasen repräsentierten Entitäten sowie Ambiguitäten, Synonyme oder Umschreibungen werden in traditionellen Informationssystemen nicht berücksichtigt. In der semantischen Suche wird zusätzlich zum traditionellen Textindex auch ein semantischer Index erzeugt, der die im Text enthaltenen Entitäten inklusive semantischer Relationen, wie hierarchische Ordnung oder Klassenzugehörigkeit, bereitstellt. Dieser semantische Index basiert meist auf einem Wissensgraphen, der die im Retrievalprozess erzielten Ergebnisse noch

[29] Vgl. Harald *Sack*, wie Anm. 21.

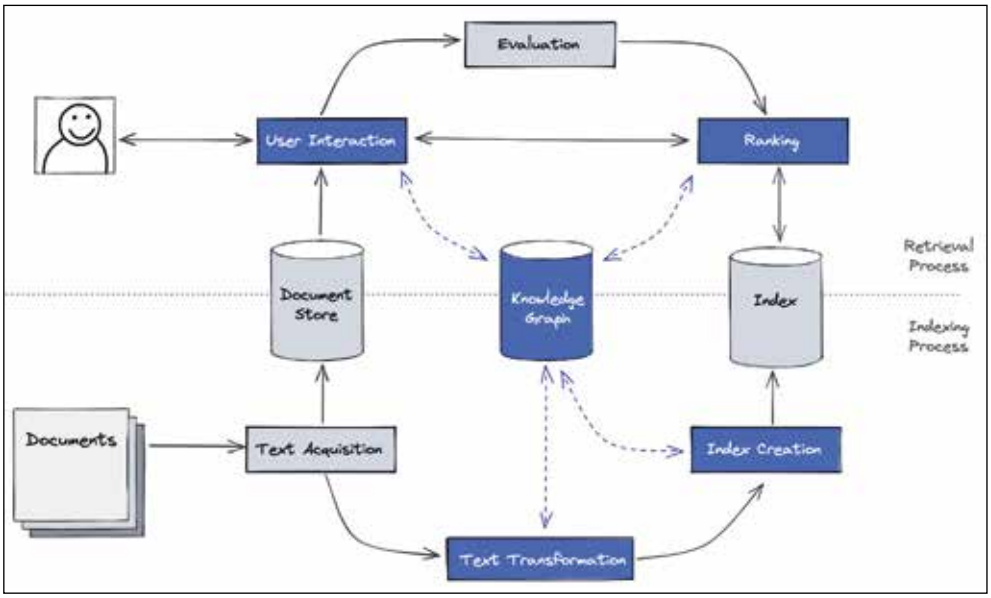

Abb. 5: Semantisches *Information Retrieval* als Grundlage der semantischen Suche auf der Basis von Wissensgraphen.

durch weitere suchbezogene Fakten ergänzen kann bzw. direkt auch für die Beantwortung von Sachfragen (*Question Answering*) verwendet werden kann (vgl. Abb. 5).

Des Weiteren wird das Ranking der erzielten Suchergebnisse über den zugrunde liegenden Wissensgraphen gesteuert, indem semantische Ähnlichkeiten und Beziehungen zwischen Suchphrasen und Dokumenteninhalten miteinander in Bezug gesetzt werden können.[30] Zusätzlich können inhaltsbasierte Suchfacetten generiert werden, die über eine entsprechende Filterung im Rahmen der Benutzendenschnittstelle eine detaillierte Exploration der Suchergebnisse gestatten.[31]

Im Mittelpunkt einer semantischen Suche stehen semantische Entitäten (*Things not Strings.*), wie Personen, Objekte, Ereignisse, Orte, etc. Die entitätenzentrierte semantische Suche verbindet eine zielgenaue Suche in Dokumenten, die die betreffende Entität enthalten oder beschreiben mit

[30] Vgl. Jörg *Waitelonis*, Claudia *Exeler* und Harald *Sack*: Linked Data Enabled Generalized Vector Space Model To Improve Document Retrieval. In: Proceedings of 3rd Int. Workshop on NLP&DBpedia 2015, co-located with ISWC 2015, CEUR workshop proceedings 1581 (2015) S. 33–44, https://nlpdbpedia2015.files.wordpress.com/2015/08/nlpdbpedia_2015_submission_7.pdf (aufgerufen am 07. 12. 2022).

[31] Vgl. Claudia *Exeler*, Jörg *Waitelonis* und Harald *Sack*: Linked Data Annotated Document Retrieval, Poster & Demo Track. In: Proceedings of 14th Int. Semantic Web Conference 2015 (ISWC 2015), CEUR Workshop proceedings 1486 (2015), http://ceur-ws.org/Vol-1486/paper_109.pdf (aufgerufen am 07. 12. 2022).

einer ähnlichkeitsbasierten Suche. Dabei steht die semantische Ähnlichkeit einer im Suchfokus stehenden Entität mit z. B. über- oder untergeordneten Konzepten, nahe verwandten Konzepten oder aber auch mit Konzepten, die im engen Bezug dazu stehen, im Vordergrund und ergänzt die direkten (exakten) Suchergebnisse mit diesen naheliegenden Ergebnissen.[32]

Explorative Suche

Auf der Grundlage der bereits beschriebenen semantischen Suche lässt sich eine zielgerichtete explorative Suche des einem Informationssystem zugrunde liegenden Dokumentenbestands realisieren. Explorative Suche ist, im Gegensatz zur traditionellen Websuche, oft dadurch gekennzeichnet, dass das finale Suchziel im Sinne eines bestimmten, eventuell bereits bekannten Dokuments, zu Beginn einer Suche nicht vorliegt.[33] Oft müssen sich Suchende zunächst einen Einblick in eine bislang noch unbekannte Domäne verschaffen, um eine bestimmte Suchintention ausdrücken zu können, z. B. weil das benötigte Fachvokabular fehlt. In diesem Kontext spielen weiterführende gezielte Suchempfehlungen sowie Visualisierungen des Dokumentenbestands und eventueller Sachzusammenhänge eine entscheidende Rolle.[34]

Ein Beispiel einer Implementierung eines Werkzeugs zur explorativen Suche in Wordpress Content Management Systemen ist *refer.cx*,[35] das als frei verfügbares Wordpress Plug-in realisiert wurde und die semantische Annotation und interaktive Visualisierung von Entitäten sowie deren Zusammenhänge in Blogbeiträgen ermöglicht.[36] Abb. 6 zeigt die Visualisierung möglicher Zusammenhänge zwischen Gottfried Wilhelm Leibniz und René Descartes auf der Grundlage der *DBpedia*-Wissensbasis und der in der Beispielanwendung verwalteten Blogbeiträge zur Wissenschaftsgeschichte.[37]

---

32 Vgl. Harald *Sack*: The Journey is the Reward – Towards New Paradigms in Web Search, invited keynote at 18th Int. Conf. on Business Information Systems 2015 (BIS 2015). In: Lecture Notes in Business Information Processing 228. Hg. von Witold *Abramowicz*. Cham u. a. 2015. S. 15–26.

33 Vgl. Gary *Marchionini*: Exploratory search: from finding to understanding. Communications of the ACM 49, 4 (2006) S. 41–46, https://doi.org/10.1145/1121949.1121979 (aufgerufen am 07. 12. 2022).

34 Vgl. Jörg *Waitelonis* und Harald *Sack*: Towards exploratory video search using linked data,. In: Multimedia Tools and Applications 59,2 (2012) S. 645–672, DOI: 10.1007/s11042-011-0733-1.

35 refer.cx Web page, https://refer.cx/ (aufgerufen am 07. 12. 2022).

36 Vgl. Tabea *Tietz* u. a: Semantic Annotation and Information Visualization for Blogposts with *refer*. In: Proceedings of 2nd. Int. Workshop on Visualization and Interaction for Ontologies and Linked Data 2016, co-located with ISWC 2016, Band 1704 (2016) S. 28–40, http://ceur-ws.org/Vol-1704/ (aufgerufen am 07. 12. 2022).

37 Vgl. Harald *Sack*: Let us Calculate – the last Universal Academic Gottfried Wilhelm Leibniz, SciHi Blog – daily blog on science, technology, and art in history, 2018, http://scihi.org/universal-academic-gottfried-wilhelm-leibniz/ (aufgerufen am 07. 12. 2022).

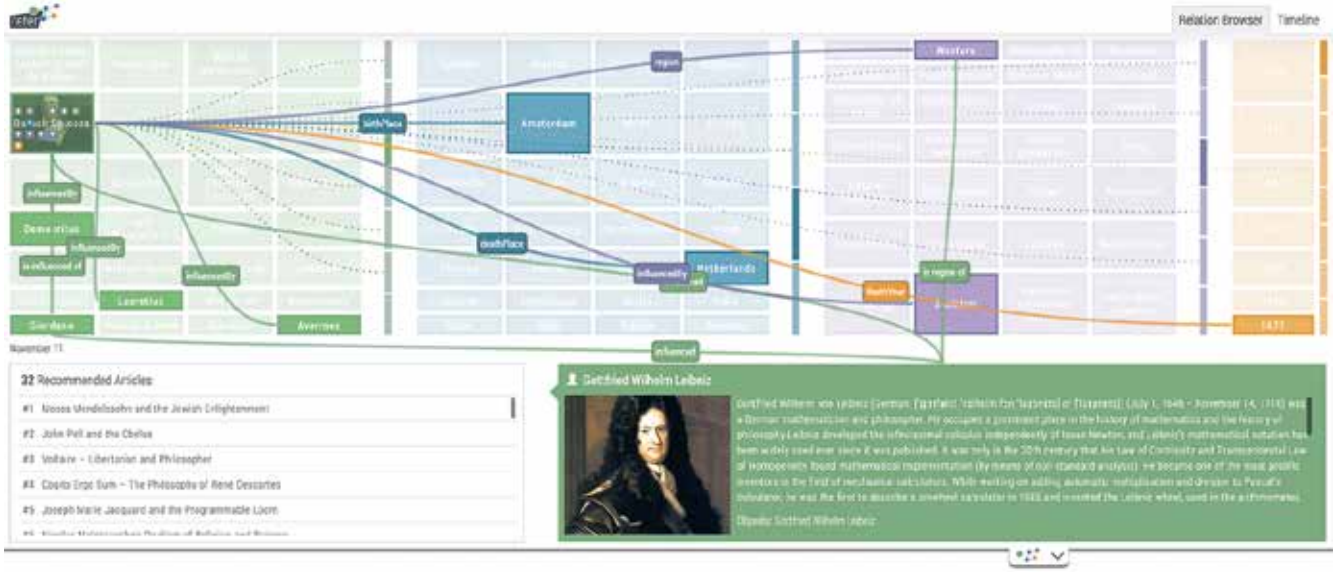

Abb. 6: Explorative Suche am Beispiel des refer.cx Wordpress PlugIns im SciHi Blog – daily blog on science, technology, and art in history.

Zusammenfassend lässt sich festhalten, dass die zukünftige digitale Forschung im Archivbereich maßgeblich von der Bereitstellung digitalisierter und originär digitaler Quellen sowie der Entwicklung innovativer Werkzeuge und Methoden für Analyse und Anreicherung dieser Daten abhängt. Dabei stellen die Erweiterung des Datenangebots, die Verbesserung der Datenqualität und die Optimierung der Bereitstellungskanäle entscheidende Handlungsfelder dar.