Fachbeitrag

Elisabeth Klindworth und Benjamin Rosemann

Das FDMLab@LABW

The FDMLab@LABW – Using Data Science Methods and Techniques in Archives

Data-Science-Methoden und -Techniken für den Einsatz im Archiv

https://doi.org/10.1515/abitech-2022-0029

Zusammenfassung: Das FDMLab des Landesarchivs Baden-Württemberg evaluiert geeignete Werkzeuge und Methoden aus dem Data Science- und KI-Bereich, passt diese für den Einsatz im Archiv an und führt erste Praxistests durch. Ziel ist es, Archivgut noch besser auffindbar und nachnutzbar zu machen und insbesondere die Bedarfe der digital forschenden, historisch orientierten Geisteswissenschaften zu berücksichtigen. Auch archivinterne Arbeitsprozesse der Erschließung sollen technisch besser unterstützt werden. Der Beitrag vermittelt einen Überblick über die bisherigen Projekte des FDMLabs in den Bereichen automatische Volltexterkennung, Werkzeuge zur Datenanalyse und -anreicherung sowie Interoperabilität und Schnittstellen.

Schlüsselwörter: Erschließung, Data Science, FAIR

Abstract: The FDMLab of the Baden-Wuerttemberg State Archives evaluates tools and methods in the field of data science and AI. Suitable tools are adapted for use in archives and initial practical tests are carried out. The aim is to make it easier to find and reuse archival material. The needs of the digital humanities are particularly taken into account. The aim is also to provide better technical support for internal archival indexing processes. The article provides an overview of the FDMLab's projects in the areas of automatic full-text recognition, tools for data analysis and enrichment, as well as interoperability and application programming interfaces.

Keywords: archival description, data science, FAIR

1 Einleitung und Ausgangslage

Die Forschungspraxis der historisch arbeitenden Geisteswissenschaften ist zunehmend digital und datengetrieben. Dadurch sehen sich die Archive gestiegenen Erwartungen der Nutzerinnen und Nutzer gegenüber: Forschungsdaten sollen digital vorliegen, maschinell durchsuchbar, standardisiert und mit anderen Daten kombinierbar, in individuelle Forschungsumgebungen integrierbar, gut dokumentiert und damit verständlich sowie nachnutzbar sein. Sowohl das Archivgut selbst in analoger, digitalisierter oder originär digitaler Form (Primärdaten) als auch die (Erschließungs-)Metadaten in Archiven sind Gegenstand der Forschung und daher Forschungsdaten.[1] Das Landesarchiv Baden-Württemberg verfügt mit seiner reichhaltigen Überlieferung vom Mittelalter bis in die Gegenwart über sehr umfangreiche Forschungsdaten, die für ein breites wissenschaftliches Themenspektrum von großem Interesse sind.

Um die Bedürfnisse seiner Nutzerinnen und Nutzer optimal zu unterstützen und die Zugänglichkeit und Benutzbarkeit seines Archivguts zu steigern, baut das Landesarchiv Baden-Württemberg eine eigene Basisinfrastruktur im Bereich E-Science und Forschungsdatenmanagement auf. Das Projekt „Forschungsdatenmanagementlabor am Landesarchiv Baden-Württemberg“, abgekürzt FDMLab@LABW, wird durch das Ministerium für Wissenschaft, Forschung und Kunst im Rahmen der Zukunftsoffensive III

1 Zur archivfachlichen Definition des Forschungsdatenbegriffs vgl. Konferenz der Leiterinnen und Leiter der Archivverwaltungen des Bundes und der Länder (KLA). „Archive als Informationsdienstleister und Infrastruktureinrichtungen“. 2018, 1, online: https://www.bundesarchiv.de/DE/Content/Downloads/KLA/positionspapier-forschungsdateninfrastruktur.pdf?__blob=publicationFile (09.05.2022). Außerdem: Maier, Gerald, Daniel Fähle, Andreas Neuburger. „Bereitstellung, Aufbereitung, Langzeitsicherung. Funktionen der Archive in der Forschungsdateninfrastruktur.“ *Archivar* Nr. 73,1 (2020): 13–14.

gefördert. Während der zweijährigen Laufzeit identifiziert und evaluiert das FDMLab seit August 2020 geeignete Werkzeuge und Methoden aus dem Data Science und KI-Bereich, passt diese ggf. für den Einsatz im Archiv an und führt erste Praxistests durch.

Dieser Beitrag umreißt zunächst die Zukunftsperspektiven, die sich durch den Einsatz von KI bzw. Maschinellem Lernen für die Bereitstellung und Nutzung archivischer Forschungsdaten eröffnen. Daran anschließend wird ein Überblick über die Aufgabenfelder und bisherigen Projekte des FDMLabs gegeben. Es wird ein Einblick vermittelt, welche Werkzeuge und Methoden im FDMLab bereits untersucht wurden, was sich bewährt hat und was nicht – und es wird dargestellt, wo aus Sicht des FDMLabs weiteres Entwicklungspotenzial für die Archive liegt.

2 Zukunftsperspektiven für die Nutzung archivischer Forschungsdaten – Warum braucht das Landesarchiv Baden-Württemberg ein FDMLab?

Um effizient digital forschen zu können, benötigen Wissenschaftlerinnen und Wissenschaftler von Seiten der Archive aggregierte Datensammlungen, auf die zeit- und ortsunabhängig über zentrale Sucheinstiege zugegriffen werden kann.[2] Das Landesarchiv Baden-Württemberg stellt seine Erschließungsinformationen und Digitalisate von Archivalien über das hauseigene Online-Findmittelsystem (OLF)[3] im Internet bereit und gibt seine Daten darüber hinaus unter anderem an das Archivportal-D[4] weiter, das als zentrales, institutionenübergreifendes Rechercheportal für Archivgut deutscher Archive dient. Die Digitalisierung von Archivgut wird seit Jahren mit einem erheblichen Ressourceneinsatz vorangetrieben. Seit im Jahr 2003 die ersten Digitalisate online gestellt wurden, wächst die Zahl der digitalisierten Archivalien stetig, sodass heute bereits 16,8 Millionen Digitalisate online recherchiert, angesehen und heruntergeladen werden können (Stand Mai 2022).

Während auf der einen Seite die Digitalisierung weiter vorangetrieben werden muss, ist die Bereitstellung von Scans auf der anderen Seite lediglich ein erster Schritt, um eine zukunftsweisende wissenschaftliche Nutzung von Archivgut zu unterstützen. Um textbasiertes Archivgut besser auffindbar und auch maschinell durchsuchbar zu machen, ist die zusätzliche Bereitstellung von Volltexten erforderlich. Volltexte können außerdem als Datengrundlage für weitere Services dienen. Beispielsweise können durch Datenextraktion und Mining-Verfahren (semi-) automatisiert Erschließungsinformationen aus Volltexten generiert werden. Dies wäre ein möglicher Ansatz, trotz des bestehenden Personal- und Ressourcenmangels in der Erschließung eine tiefere und damit qualitativ hochwertigere Erschließung zu erreichen. Darüber hinaus sind Anwendungen zur Datenanalyse, Bereinigung, Annotation, Verknüpfung mit Daten aus anderen Datenquellen (beispielsweise Normdaten) oder auch zur Datenvisualisierung denkbar. Diese Anwendungen können die Archivarinnen und Archivare bei der Erschließung technisch unterstützen, damit diese sich stärker auf die intellektuellen Aspekte ihrer Tätigkeit konzentrieren können und weniger Zeit mit rein „mechanischen“ Aufgaben verbringen müssen. Für Nutzerinnen und Nutzer ergäben sich neue Analysemöglichkeiten, um aus historischen Daten Wissen zu generieren.

Das FDMLab arbeitet daran, einen geeigneten Werkzeugkasten zusammenzustellen, um solche Funktionalitäten umsetzen zu können. Das Projekt steht dabei in engem Zusammenhang mit dem Engagement des Landesarchivs Baden-Württemberg im Konsortium NFDI4Memory,[5] das den Aufbau einer Nationalen Forschungsdateninfrastruktur (NFDI) für die historisch arbeitenden Wissenschaften zum Ziel hat. Auch innerhalb des Landesarchivs besitzt das Projekt eine besondere Relevanz: Da das Landesarchiv Baden-Württemberg bis 2025 seine zentralen IT-Systeme für die Erschließung und die Online-Präsentation durch eine Neuentwicklung ablösen wird, gilt es gerade jetzt, die richtigen Weichen für eine zukunftsweisende Datenbereitstellung für die digitalen Geisteswissenschaften und andere Nutzergruppen zu stellen. Das FDMLab baut hierzu Know-how auf, das in den Entwicklungsprozess des neuen archivischen Fachinformationssystems (AFIS) einfließen wird. Darüber hinaus unterstützt das FDMLab auch weitere Projekte des Landesarchivs, in denen innovative Methoden wie z. B. automatische Texterkennung zum Einsatz kommen (s. u.).

2 Wettlaufer, Jörg. „Welche Services braucht die digitale Geschichtswissenschaft von Bibliotheken, Archiven, Museen und Datenzentren?“ Online: https://www.historikertag.de/Muenchen2021/sektionen/gedaechtnisinstitutionen-in-der-digitalen-welt-bibliotheken-museen-archive-und-die-geschichtswissenschaft/ (5.10.2021).

3 Online-Findmittelsystem des Landesarchivs Baden-Württemberg unter https://www2.landesarchiv-bw.de/ofs21/ (13.05.2022).

4 Archivportal-D unter https://www.archivportal-d.de/ (13.05.2022).

5 Webseite des Konsortiums unter https://4memory.de/ (13.05.2022).

3 Aufgabenfelder des FDMLabs

Das Kernteam des Projekts wird von einer Archivarin und einem Data Scientist (Autorin und Autor dieses Beitrags) gebildet. Diese Kombination zeigt bereits den Ansatz des Projekts, die archivfachlichen Anforderungen mit der Expertise aus dem Data Science-Bereich zusammenzubringen – ein wesentlicher Erfolgsfaktor für das Projekt FDMLab. Die Aufgabenfelder des FDMLabs gliedern sich in die drei Arbeitsbereiche Recherchierbarkeit von Volltexten, Werkzeuge zur Datenanalyse und -anreicherung sowie Interoperabilität und Schnittstellen. Zwischen den drei Bereichen bestehen zahlreiche inhaltliche Querverbindungen, weshalb parallel an den verschiedenen Aufgaben gearbeitet wird.

Das Projekt ist – da mit der Anwendung maschinellen Lernens im Archiv Neuland beschritten wird – experimentell angelegt und bietet den notwendigen Raum, verschiedene Methoden und Technologien auszuprobieren und erste Erfahrungen damit zu sammeln. Dass nicht alles, was getestet wird, unmittelbar im Produktivbetrieb eingesetzt werden kann, gehört zum Charakter eines Forschungsprojekts. Bei der Beschäftigung mit den genannten Arbeitsbereichen bewegen wir uns im FDMLab auf verschiedenen Ebenen der inhaltlichen Auseinandersetzung: Die oberste Ebene ist das Sammeln von Informationen und das Ermitteln von Optionen. Die nächst tiefere Ebene ist das praktische Ausprobieren und Evaluieren von Werkzeugen und Methoden in einem Testumfeld. Auch Erfahrungs- und Fehlerberichte, Bugfixes oder Codebeiträge bei Open-Source-Projekten gehören im FDMLab zur Evaluationsphase. Daran anschließend ist die Produktivsetzung von Werkzeugen und die Integration einer neuen Methode in den Arbeitsalltag die tiefste Ebene der Auseinandersetzung. Die Unterscheidung dieser Ebenen bietet im Projektmanagement einen Orientierungsrahmen, um Ziele für das FDMLab zu definieren und den Standpunkt auf einem Themenfeld zu bestimmen.

3.1 Recherchierbarkeit von Volltexten

Durch die Volltexterkennung von digitalisiertem Archivgut wird für die Forschung eine gegenüber der ausschließlichen Bereitstellung von Bilddateien deutlich breitere Datengrundlage geschaffen. Sie ermöglicht eine Volltextsuche über den gesamten Text von Dokumenten, ähnlich wie Nutzerinnen und Nutzer es von gängigen Internetsuchmaschinen gewohnt sind. Volltexte bilden darüber hinaus die Ausgangsbasis für die Bereitstellung einer semantischen Suche, die sich an der tatsächlichen inhaltlichen Bedeutung einer Suchanfrage orientiert und daher perspektivisch präzisere Ergebnisse liefern kann als eine Suche, die lediglich auf dem Auffinden von Schlüsselwörtern im Text basiert. Aus diesem Grund bildet die Volltexterkennung einen Schwerpunktbereich in der bisherigen Arbeit des FDMLabs.

3.1.1 Evaluation von OCR- und HTR-Tools

Begrifflich wird zwischen der Erkennung von gedrucktem Text, der Optical Character Recognition (OCR), und der Erkennung von Handschrift (Handwritten Text Recognition – HTR), unterschieden. Im FDMLab wurden für die Erfassung von Volltexten aus Digitalisaten mittels OCR und HTR zunächst verschiedene einschlägige Softwarelösungen evaluiert. Als Testdaten dienten hierbei die digitalisierten Kriegs- und Friedensstammrollen aus der Abteilung Hauptstaatsarchiv Stuttgart (Bestand LABW HStAS M 430/1 – M 631) und die ebenfalls digitalisierten Kriegsgräberlisten aus der Abteilung Staatsarchiv Ludwigsburg (Bestand LABW StAL EL 20/1 VI). Aufgrund ihrer tabellenartigen Struktur und der Mischung aus unterschiedlichsten Maschinen-, Hand- und Druckschriften in einem Dokument weisen diese Bestände Merkmale auf, die auch für andere projektrelevante Teile des jüngeren Archivguts im Landesarchiv typisch sind.

Mit diesen Beständen konnte wertvolles Wissen für die Auswahl geeigneter Strategien zur Volltexterkennung von Archivgut gewonnen werden. Um historische Handschriften zu erkennen, hat sich die Software Transkribus[6] als besonders geeignet erwiesen. Für die Bearbeitung von gedruckten Archivalien des Landesarchivs können vor allem mit den im Projekt OCR-D[7] entwickelten Werkzeugen sowie dem Werkzeug OCR4All[8] geeignete Workflows aufgebaut werden.

Durch die Evaluation der OCR- und HTR-Tools hat sich die Vermutung bestätigt, dass es keine „One fits all"-Lösung für alle Bestände des Landesarchivs geben kann. Da das Archivgut im Schriftbild und vor allem auch im Layout sehr heterogen ist, müssen für unterschiedliche Bestände jeweils eigene Workflows definiert und die Parameter der Tools neu konfiguriert werden. Um komplexe Layouts wie z. B. Tabellen und Formulare erkennen zu können, ist es außerdem notwendig, jeweils eigene Erkennungsmodelle für ein spezifisches Layout zu trainieren.

6 Read-Coop. „Transkribus." Online: https://transkribus.eu/ (13.05.2022).
7 OCR-D. Online: https://ocr-d.de/ (13.05.2022).
8 OCR4all. Online: http://www.ocr4all.org/ (13.05.2022).

Dazu müssen zunächst Ground Truth Daten in größerem Umfang manuell generiert werden, mit deren Hilfe das Modell die korrekte Erfassung eines Layouts „erlernt". Der Personal- und Zeitaufwand für ein solches Modelltraining ist dabei nicht zu unterschätzen. Im Anschluss können mit dem Modell Volltexte automatisch generiert werden. Manuelle Arbeitsschritte zur Verbesserung der automatisch generierten Ergebnisse fallen dennoch an, um eine hohe Qualität der Volltexte zu erreichen. Die Aufgabe, aus den textuellen Digitalisaten des Landesarchivs Volltexte in größerem Umfang zu generieren, wird daher auch nach Abschluss der zweijährigen Projektlaufzeit des FDMLabs eine Aufgabe für die kommenden Jahre bleiben.

Die Qualitätskontrolle von OCR- und HTR-Ergebnissen ist derzeit eine technisch noch nicht vollständig gelöste Aufgabe. Dies gilt insbesondere für eine objektive Kontrolle der Ergebnisse der Layoutsegmentierung, für die noch keine zuverlässig einsetzbaren Tools zur Verfügung stehen.[9] Um künftig auch Qualitätskontrollen anhand objektiver Kennzahlen durchführen zu können, wurde das Werkzeug „Dinglehopper" aus dem Qurator-Projekt[10] zur Analyse von OCR-Ergebnissen evaluiert und an die Bedürfnisse des Landesarchivs angepasst. Dazu gehört die praktische Implementierung der Metrik „Flexible character accuracy"[11] zur Bewertung von OCR-Ergebnissen aus Dokumenten mit komplexer Layoutgestaltung.

3.1.2 OCR-Technologie

Generell stellten wir in der Projektlaufzeit fest, dass im Open-Source-Bereich der Zugang zu OCR-Technologie mit Projekten wie eScriptorium,[12] OCR4All und OCR-D deutlich vereinfacht wurde. Die Werkzeuge können Endanwenderinnen und Endanwendern über eine Weboberfläche zur Verfügung gestellt werden, so dass als technologische Voraussetzung lediglich ein aktueller Webbrowser mit Internetzugang notwendig ist. Die Integration der verschiedenen Betriebsmodi und Parameter für die OCR-Prozesse in die Bedienung der Weboberfläche bedeutet für die Entwicklerinnen und Entwickler jedoch einen separaten Arbeitsschritt. Das hat zur Folge, dass in den Weboberflächen der einzelnen Werkzeuge nicht der volle Funktionsumfang von manuell konfigurierten OCR-Prozessen zur Verfügung gestellt wird. So bleibt zum Beispiel die Erkennung von Material mit komplexen Layoutanforderungen nach wie vor ein Problem, das aktuell mit vielen manuellen Korrekturen oder umfangreichem technischem Fachwissen gelöst wird.

Die Erfahrungen des FDMLabs mit der Volltexterkennung unterstreichen, dass zu Beginn eines jeden OCR- oder HTR-Projektes im Archiv grundsätzlich überlegt werden sollte, welcher Qualitätsanspruch an das Ergebnis zu stellen ist. Dieser hängt stark vom Anwendungsfall ab. Soll der Volltext als Basis für eine Suche nach bestimmten Schlüsselbegriffen in einem Dokument genutzt werden (*Keyword Spotting*)? Soll der Volltext gemeinsam mit dem ursprünglichen Digitalisat in einem Viewer angezeigt werden, um Nutzerinnen und Nutzern als Lesehilfe für historische Schriften zu dienen, oder soll der Volltext gar Basis einer wissenschaftlichen Quellenedition sein? Für das Keyword Spotting ist mitunter bereits eine Erkennungsgenauigkeit von 70–80 % akzeptabel, für das Lesen eines Textes ist eine Genauigkeit von 90 % oder mehr anzustreben.[13] Im Fall einer Vergabe der Volltexterkennung an einen externen Dienstleister empfiehlt die DFG in ihren Leitlinien zur Digitalisierung, eine Genauigkeit von möglichst nicht unter 95 % zu vereinbaren.[14] Nach den bisherigen Erfahrungen des FDMLabs dürften die Erkennungsraten, die sich mit öffentlich nachnutzbaren Modellen erzielen lassen, bei vielen handschriftlichen Archivbeständen eher unter 90 % liegen (Stand Mai 2022). Um die Qualität der Texterkennung für Archivgut zu erhöhen, braucht es das Engagement der Archive in der OCR/HTR-Community: Erkennungsmodelle, die in einem Archiv trainiert wurden, können für andere Archive bereitgestellt werden. Dies gilt ebenfalls für mit hohem Ressourcenaufwand erstellte Ground-Truth-Daten, die für das Training weiterer Modelle nachgenutzt werden können.

3.1.3 Scanstifte

Neben den bereits erwähnten Softwarelösungen gibt es eine weitere interessante Variante, um automatische

9 Hinweise auf praktisch einsetzbare Tools zur Qualitätskontrolle der Layoutsegmentierung nehmen Autor und Autorin dieses Beitrags gern entgegen.

10 Qurator. Online: https://qurator.ai (13.05.2022).

11 Vgl. Clausner, C., S. Pletschacher, A. Antonacopoulos. „Flexible character accuracy measure for reading-order-independent evaluation." *Pattern Recognition Letters*, Nr. 131 (2020): 390–397.

12 eScriptorium. Online: https://gitlab.com/scripta/escriptorium (13.05.2022).

13 Vgl. Empfehlungen für Transkribus-Texterkennungsmodelle der Read-Coop. „How To Train and Apply Handwritten Text Recognition Models in Transkribus." Online: https://readcoop.eu/transkribus/howto/how-to-train-a-handwritten-text-recognition-model-in-transkribus/ (12.05.2022).

14 Deutsche Forschungsgemeinschaft (DFG). „DFG-Praxisregeln Digitalisierung". S. 35. Online: https://www.dfg.de/formulare/12_151/12_151_de.pdf (12.05.2022).

Texterkennung in archivische Arbeitsabläufe einzubetten: Einige Hersteller vertreiben spezielle Scanstifte, die – ähnlich wie ein herkömmlicher Textmarker – über ein Blatt Papier geführt werden können. Dabei wird ein Text Zeile für Zeile erfasst und direkt in eine Textdatei am eigenen PC übertragen. Bei einigen Scanstiftmodellen kann der gescannte Text auch direkt in ein Formular, beispielsweise die Maske einer archivischen Erschließungssoftware, übertragen werden.

Im FDMLab haben wir mit zwei verschiedenen Scanstiftmodellen exemplarisch getestet, wie sich damit Informationen aus analogen Findmitteln digitalisieren lassen. Unsere bisherigen Tests ergaben, dass die Scanstifte bei modernen gedruckten Texten sehr gute Ergebnisse erzielen, bei der Erkennung älterer maschinenschriftlicher Findbücher jedoch recht fehleranfällig sind. Fraktur oder historische Handschriften wurden von den getesteten Modellen nicht unterstützt, da die von diesen verwendete OCR-Software auf moderne Dokumente abgestimmt ist. Auch ein Training eigener spezifischer Erkennungsmodelle für bestimmte Schrifttypen ist für die Scanstifte nicht vorgesehen.

Demgegenüber steht der Vorteil, dass die Scanstifte auch ohne tieferes technisches Verständnis sofort eingesetzt werden können. Die Handhabung der Scanstifte erfordert aber etwas Übung: Sowohl der Winkel, in dem der Stift gehalten wird, als auch Lichtverhältnisse und Scangeschwindigkeit haben Einfluss auf das Ergebnis. Nach ein bis zwei Stunden Einarbeitungszeit beherrschten wir die Arbeitsweise jedoch schon gut. Der Einsatz der Scanstifte soll im Landesarchiv Baden-Württemberg weiter praktisch erprobt werden. Als potenziellen Anwendungsfall sehen wir etwa die Digitalisierung von gedruckten Findbüchern, bei denen statt einer kompletten OCR-Erkennung die Informationen direkt in die dafür vorgesehenen Felder der Erschließungsmaske gescannt werden sollen.

3.2 Werkzeuge zur Datenanalyse und -anreicherung

Nahezu alle Archive kennen die Problematik, dass für die fachgerechte Erschließung des Archivguts nur begrenzte Personal- und Zeitressourcen zur Verfügung stehen. Um die Menge der neu übernommenen Unterlagen zu bewältigen und gleichzeitig auch noch Erschließungsrückstände abzubauen, müssen bei der Erschließungstiefe oftmals Abstriche hingenommen werden. Dies erschwert den Nutzerinnen und Nutzern in vielen Fällen die Recherche. Der Einsatz von Verfahren des Data-Minings könnte dieses Problem in Zukunft abmildern, da sich mit deren Hilfe Erschließungsmetadaten (semi-)automatisch generieren lassen. So können beispielsweise bestimmte Entitäten wie Personen und Körperschaften oder auch sachthematische Referenzen in vorhandenen Textkorpora identifiziert, maschinell extrahiert und als Erschließungsinformationen zur Verfügung gestellt werden. Die Qualität und Quantität der Erschließungsinformationen ließe sich so trotz knapper Personalressourcen erhöhen.

An dieser Stelle soll noch einmal deutlich gesagt werden, dass es im FDMLab nicht darum geht, vorhandenes Personal durch Automatisierung einzusparen. Vielmehr ist das Ziel, Archivarinnen und Archivare mit neuem Handwerkszeug auszurüsten, das ihnen die Arbeit mit schwach strukturierten, uneinheitlich formatierten Daten erleichtert. Denn sonst verwenden Archivarinnen und Archivare zu viel ihrer wertvollen Arbeitszeit darauf, Schritte zur Datenbereinigung manuell auszuführen, die eine Datenspezialistin oder ein Datenspezialist in wenigen Minuten erledigt hätte. Dabei kann nicht erwartet werden, dass alle Archivarinnen und Archivare nun selbst Programmieren lernen. Deswegen legen wir im FDMLab ein besonderes Augenmerk auf die Frage, wie wir das neue „Handwerkszeug“ in einer auch für Personen ohne tiefere technische Kenntnisse nutzbaren Form zur Verfügung stellen können.

3.2.1 Anreicherung mit Normdaten

Während das Data Mining darauf abzielt, mit algorithmusbasierten Analyseverfahren Informationen aus einer schwach strukturierten Datenbasis zu extrahieren, birgt auch der umgekehrte Weg, Erschließungsdaten mit Daten aus anderen, externen Datenquellen anzureichern, großes Potenzial: Normdaten, beispielsweise aus der Gemeinsamen Normdatei (GND), ermöglichen es, Personen, Körperschaften, Geografika und Sachbegriffe eindeutig zu referenzieren und institutionenübergreifend zu verschlagworten. Die Anreicherung der Erschließungsdaten mit Normdaten hilft so, die gezielte Recherche zu verbessern sowie Archivgut zu kontextualisieren und mit anderen Informationsquellen in Beziehung zu setzen.

Das Landesarchiv Baden-Württemberg setzt bereits seit mehr als zehn Jahren auf den Einsatz von Normdaten der GND in der archivischen Erschließung, sodass heute bereits knapp zwei Millionen Normdatenverknüpfungen bestehen (Stand Mai 2022). Es ist geplant, die Erschließung mit Normdaten in den nächsten Jahren noch deutlich auszuweiten. Im Zuge der anstehenden AFIS-Neuentwicklung sollen derzeit manuell durchgeführte Prozesse toolgestützt automatisiert und vereinheitlicht werden. So

ist für das AFIS eine eigene Indizierungskomponente vorgesehen, die eine (teil-)automatisierte Anreicherung normierter Indexansetzungen ermöglichen soll. Dadurch soll das Einbinden von Normdaten noch einfacher gestaltet werden. Hiermit eng im Zusammenhang steht das Projekt GND4C, in dem das Landesarchiv gemeinsam mit der Deutschen Nationalbibliothek und weiteren Partnern an der Öffnung der bisher stark bibliothekarisch geprägten GND für archivische Anwendungskontexte arbeitet.[15] Das FDMLab steht in engem Austausch mit dem GND4C-Projekt und gibt seine Erfahrungen mit Tools und Workflows zur Normdatenanreicherung weiter.

3.2.2 Data Mining

Eine für das Data Mining verwendete Technik ist die so genannte Named Entity Recognition. Hierfür werden Klassifikatoren verwendet, die basierend auf Stichworten und/oder Satzkonstruktionen Entitäten wie Personen, Körperschaften, Geografika und Sachbegriffe in einem Text identifizieren. Diese Entitäten können in einem Folgeschritt zum Beispiel mit Normdaten verknüpft werden (Entity Linking). Ein bereits erwähnter Anwendungsfall im Archiv ist die automatische Extraktion von Schlagwörtern aus Archivmaterial und die anschließende Anreicherung mit Normdaten, ein weiterer Anwendungsfall die strukturierte Aufbereitung von gedruckten Findbüchern für archivische Fachinformationssysteme. Im FDMLab testeten wir einige frei verfügbare vortrainierte Modelle für Named Entity Recognition auf digitalisiertem Archivgut, das Formulare und Tabellen beinhaltet, sowie auf gedruckten Findbüchern. Die Anzahl der korrekt und vollständig erkannten Entitäten war jedoch sehr gering. Gründe hierfür sind zum Beispiel das für die vortrainierten Modelle unbekannte Vokabular und die sehr speziellen Textkonstruktionen in Formularen oder Findbucheinträgen.

Beide Ansätze – Data Mining und Anreicherung mit Normdaten – bergen ein großes Potenzial für die signifikante Verbesserung der Datengrundlagen für die historische Forschung und werden daher im FDMLab in verschiedenen praktischen Projekten vorangetrieben, die im Folgenden kurz vorgestellt werden sollen.

3.2.3 OpenRefine

Um Archivarinnen und Archivare im Landesarchiv bei der Verschlagwortung von Archivmaterial und der Digitalisierung von Findbüchern zu unterstützen, entwickelten wir als Alternative zur automatischen Named Entity Recognition Regeln und Rezepte für OpenRefine, ein Werkzeug mit Fokus auf die Datenaufbereitung.[16] Mit den in OpenRefine entwickelten Verfahren können zum Beispiel Schlagwörter automatisch extrahiert und anschließend gesammelt, sortiert, fachlich geprüft, korrigiert, und mit Normdaten verknüpft werden.

In einem ersten praktischen Projekt wurden im FDMLab analog vorhandene Erschließungsdaten des Landesarchivs digitalisiert, analysiert und aufbereitet. Insbesondere wurden Metadaten zu Beständen des Landesarchivs durch Verfahren der automatischen Informationsextraktion in ein strukturiertes Datenformat umgewandelt und erfolgreich in die Erschließungsdatenbank scopeArchiv importiert. Die für Archivnutzende bisher nur über eine gedruckte Version des Findmittels recherchierbaren Bestände stehen nun über das Online-Findmittelsystem (OLF) zur Verfügung. Bearbeitet wurden ein Findbuch der Abteilung Staatsarchiv Sigmaringen zum Bestand LABW StAS Dep 30/1 T 1 Grafschaft Friedberg-Scheer, Urkunden,[17] sowie neun Findbücher zur umfangreichen Aktenüberlieferung des frühneuzeitlichen Reichskammergerichts aus mehreren Archivabteilungen.[18]

Weiterhin werden im FDMLab derzeit bestimmte Orts- und Personenindizes der Findbücher der Reichskammergerichtsakten mit dem Programm OpenRefine aufbereitet. Dabei wird ein Workflow zur Deduplizierung und zum Abgleich der Indexdaten mit den Einträgen der Gemeinsamen Normdatei (GND) entwickelt und erprobt.

Weitere Erfahrung mit dem massenhaften Einsatz von Normdaten gewann das FDMLab bei der Aufbereitung des Bestands LABW StAL EL 68 IX Landesvermessungsamt Baden-Württemberg: Landesbefliegung Baden-Württemberg

15 Projektwebseite: https://www.dnb.de/DE/Professionell/ProjekteKooperationen/Projekte/GND4C/gnd4c.html (13.05.2022).

16 OpenRefine. Online: https://openrefine.org/ (13.05.2022).

17 Findbuch zum Bestand des Landesarchivs Baden-Württemberg, Staatsarchiv Sigmaringen, Dep. 30/1 T 1, Grafschaft Friedberg-Scheer, Urkunden. Online: http://www.landesarchiv-bw.de/plink/?f=6-2857&a=fb (13.05.2022).

18 *Akten des Reichskammergerichts im Hauptstaatsarchiv Stuttgart: Inventar des Bestands C*. Bearb. von A. Brunotte, R. J. Weber. Veröffentlichungen der Staatlichen Archivverwaltung Baden-Württemberg, Band 46, Nr. 1–8, Stuttgart 1993–2008. Außerdem: *Akten des Reichskammergerichts im Staatsarchiv Sigmaringen: Inventar des Bestands R 7*. Bearb. von R. J. Weber, Veröffentlichungen der Staatlichen Archivverwaltung Baden-Württemberg; Band 57, Stuttgart 2004.

1968 – Luftbilder und digitales Orthophoto,[19] aus dem Staatsarchiv Ludwigsburg. Das FDMLab begleitete hier die automatisierte Verknüpfung der Verzeichnungseinheiten mit insgesamt 195 988 (zum Teil neuen) normierten Ortsdeskriptoren in scopeArchiv.

Die Fähigkeit, Datensätze mit OpenRefine für eine Datenbank aufzubereiten, mit Normdaten zu verknüpfen und dadurch mit weiteren Daten anzureichern, nutzte das FDMLab in der Unterstützung weiterer Datenprojekte. Dazu gehört zum Beispiel die visuelle Aufbereitung der Schicksalswege von deportierten Jüdinnen und Juden aus Baden, der Pfalz und des Saarlandes im Projekt „Gurs 1940“ durch die Anreicherung mit Geodaten.[20]

Die Erfahrungen des FDMLabs bei der Arbeit mit OpenRefine für Datenprojekte im Archiv wurden in Form von Workshops dokumentiert und über Schulungen an interessierte Kolleginnen und Kollegen im Landesarchiv Baden-Württemberg und an weitere Kooperationspartner weitergegeben.[21]

Der aktuelle Stand der Technik erlaubt zwar eine erhöhte Automatisierung von Prozessen zur Digitalisierung von Archivmaterial, Erstellung von Volltexten und Verschlagwortung. Jedoch ist die Qualität der Bearbeitung historischer Daten (noch) nicht vergleichbar mit der Leistung von modernen Systemen auf aktuellen Daten. Umso wichtiger ist die Einbindung unserer Benutzerinnen und Benutzer, die häufig als Expertinnen und Experten auf ihren Fachgebieten Material aus dem Archiv nutzen. Dabei entstehen z. B. qualitativ hochwertige Transkriptionen und Metadaten, die jedoch selten veröffentlicht werden. Im FDMLab suchten wir daher auch nach Programmen und Projekten aus dem Bereich User Generated Content, die weitere Benutzerinnen und Benutzer darin unterstützen könnten, mit unserem Archivmaterial zu arbeiten, und dazu geeignet sind, in das Angebot des Landesarchivs eingebunden zu werden. Diese Erkenntnisse fließen direkt in die Planung des neuen AFIS am Landesarchiv Baden-Württemberg ein.

3.3 Interoperabilität und Schnittstellen

Das neue AFIS des Landesarchivs Baden-Württemberg soll nicht nur alle internen archivischen Arbeitsprozesse der Erschließung unterstützen, sondern auch eine Präsentationskomponente für die Recherche und Darstellung von Archivgut im Internet umfassen. Um die Entwicklung der Präsentationskomponente konzeptionell vorzubereiten, hat das FDMLab in einem ersten Schritt evaluiert, inwieweit die vorhandenen Funktionalitäten des derzeitigen Online-Findmittelsystems (OLF) des Landesarchivs die Anforderungen an eine zeitgemäße Datenbereitstellung bereits erfüllen. Zentral war dabei auch die Frage, ob Forschungsfragestellungen auf der Basis massenhafter digitaler Daten unterstützt werden. Als Richtschnur für die Evaluation dienten die im Forschungsdatenmanagement inzwischen kanonischen FAIR-Prinzipien für die Datenbereitstellung.[22] Das Akronym FAIR steht für auffindbare (*findable*), zugängliche (*accessible*), interoperable (*interoperable*) und nachnutzbare (*reusable*) Daten. Das Ergebnis der Betrachtung der insgesamt 15 FAIR-Kriterien zeigt, dass unser aktuelles Online-Findmittelsystem die Anforderungen an „FAIRe“ Daten in einigen Punkten bereits berücksichtigt, sie jedoch noch nicht vollumfänglich erfüllt.

Die Evaluation diente als Ausgangspunkt, um gezielt Anforderungen an das künftige Online-Präsentationssystem des Landesarchivs zu formulieren, das die FAIR-Prinzipien vollständig umsetzen soll. Das neue Präsentationssystem soll auf aktuellen technologischen Entwicklungen aufbauen, zudem soll es rechtliche Nutzungsbedingungen verlässlich und allgemein verständlich kommunizieren. Es gilt, gleichzeitig eine möglichst offene Bereitstellung von Kulturdaten und die Wahrung von Datenschutz, Urheberrecht und anderen Schutzrechten zu gewährleisten. Die Software soll keine künstlichen technischen Zugangshürden schaffen, wo aus rechtlicher Sicht keine erforderlich sind. Aus dieser Zielsetzung lassen sich unter anderem die folgenden Handlungsbedarfe ableiten:

– Um die Auffindbarkeit aller Daten und Metadaten des Landesarchivs dauerhaft und weltweit zu gewährleisten und die Verknüpfung von Metadaten und Primärdaten sicherzustellen, soll ein global eindeutiger Persistent Identifier (PID) eingeführt werden. Dieser Identifier ermöglicht außerdem eine sichere Zitierweise der Objekte. Persistent Identifiers sollen darüber hinaus auch genutzt werden, um auf die Dokumentation verwendeter Normdatenquellen und Er-

19 Online-Findbuch und Digitalisate zum Bestand Landesarchiv Baden-Württemberg, Staatsarchiv Ludwigsburg, EL 68 IX. Online: http://www.landesarchiv-bw.de/plink/?f=2-5684305 (13.05.2022).

20 Meldung „Geschichte, die nicht vergeht: Datenbank zu Gurs 1940 online“ unter https://www.landesarchiv-bw.de/de/aktuelles/nachrichten/73495 (10.05.2022).

21 Die Workshops und das zugehörige Datenmaterial sind über den Projektblog des FDMLabs online zugänglich unter https://fdmlab.landesarchiv-bw.de/workshops/.

22 GoFAIR. „FAIR Principles“. Online: https://www.go-fair.org/fair-principles/ (13.05.2022).

schließungsstandards zu verweisen und dadurch die Erschließung transparenter zu gestalten.
- Die Erschließungsmetadaten werden im derzeitigen Archivinformationssystem strukturiert erfasst und über EAD-XML als Austauschformat an die Deutsche Digitale Bibliothek (DDB) und das Archivportal-D übergeben. Anschließend sind die Erschließungsmetadaten über die API der DDB abrufbar. Das neue Archivinformationssystem soll ebenfalls eine frei zugängliche Schnittstelle besitzen, über die Nutzerinnen und Nutzer Metadaten in einem geeigneten maschinenlesbaren Format herunterladen können. Dabei soll sowohl der Download ausgewählter Datensätze als auch der Download einer größeren Anzahl von Datensätzen möglich sein. Selbstverständlich gilt dies ebenso für Digitalisate.
- Darüber hinaus hat das FDMLab konzeptionelle Vorüberlegungen zur Implementierung eines IIIF-kompatiblen Viewers geleistet, mit dem Digitalisate und zughörige Volltexte für Nutzerinnen und Nutzer im Internet angezeigt werden können. Unterschiedliche Umsetzungsvarianten wurden mit einem Prototyp praktisch getestet.
- Das Landesarchiv Baden-Württemberg stellt sämtliche Erschließungsinformationen unter der Creative Commons-Lizenz CC0 zu Verfügung. Digitalisate werden, soweit sie nicht ohnehin gemeinfrei sind und dies rechtlich möglich ist, unter der Creative Commons-Lizenz CC-BY veröffentlicht. Die Information zu diesen rechtlichen Bedingungen der Nachnutzung sollen künftig nicht nur an zentraler Stelle auf der Homepage des Landesarchivs, sondern auch in der Detailansicht eines jeden Objekts angezeigt werden, damit sie noch leichter auffindbar sind.
- Die Erschließungsinformationen sollten nicht nur maschinenlesbar sein, sondern auch als semantisches Datenmodell bzw. Knowledge-Graph-Database umgesetzt werden. Im neuen AFIS wird daher der Einsatz des neuen internationalen Erschließungsstandards Records in Contexts (RiC)[23] angestrebt.

Diese ersten Vorschläge für eine möglichst nutzerfreundliche Gestaltung des künftigen Online-Präsentationssystems werden im Landesarchiv Baden-Württemberg noch weiterentwickelt und verfeinert werden. Das FDMLab wird den Entwicklungsprozess des Online-Präsentationssystems weiterhin fachlich begleiten und konzeptionelle Überlegungen dazu anstellen.

4 Schlussüberlegungen

Die Methoden und Technologien aus dem Data-Science- und KI-Bereich, denen sich das Landesarchiv Baden-Württemberg mit dem FDMLab-Projekt widmet, haben für die Archive ein großes Potenzial, der Forschung eine qualitativ und quantitativ verbesserte Datengrundlage zur Verfügung zu stellen. Auch die Recherchemöglichkeiten für Nutzerinnen und Nutzer können perspektivisch deutlich komfortabler und effizienter gestaltet werden.

Die in diesem Beitrag vorgestellten Verfahren der automatischen Texterkennung und Werkzeuge zur Datenanalyse und -anreicherung werden im FDMLab und Projekten anderer GLAM-Institutionen praktisch erprobt. Erste Erfahrungen liegen bereits vor, insgesamt betrachtet handelt es sich jedoch nach wie vor eher um technisches „Neuland“, das es noch weiter zu erkunden gilt. Für Archive bedeutet dies auch, dass die Werkzeuge, die in der Regel außerhalb der Archivwelt konzipiert wurden, auf die eigenen Bedürfnisse angepasst werden müssen. Die historischen Daten der Archive mit ihrer oft sehr heterogenen Struktur und spezifischen, zeitbedingten Semantik erfordern doch immer wieder andere Regeln und Verfahrensweisen als moderne Daten. Die Mitarbeit bei der Entwicklung von Open-Source-Tools ist für Archive wichtig, um die archivspezifische Sichtweise einzubringen und bekannt zu machen.

Bisherige Erfahrungen des FDMLabs deuten darauf hin, dass sich in den meisten Fällen keine „One fits all“-Pipelines für die Bearbeitung der Daten des Landesarchivs aufbauen lassen, da die Daten zu unterschiedlich sind. Projekte erfordern daher trotz der Automatisierung von Teilschritten individuelle Anpassungen und intellektuelle Arbeit. Daten in größerem Umfang zu generieren, bleibt damit eine fortlaufende Aufgabe der kommenden Jahre.

Um hier voranzukommen, sind für das FDMLab Kooperation und Austausch von zentraler Bedeutung. Innerhalb des Landesarchivs entstanden im Gespräch mit Kolleginnen und Kollegen der verschiedenen Archivabteilungen immer wieder neue Ideen, wie das FDMLab sich bei der Lösung konkreter archivischer Alltagsprobleme einbringen kann. Insbesondere die insgesamt sechs Schulungen zur Datenaufbereitung und Normdatenanreicherung mit OpenRefine stießen auf große Resonanz. Über das Landesarchiv hinaus war das FDMLab bei verschiedenen Workshops, Coffee Lectures und Konferenzen ver-

23 Ein Onlinefassung des Standardentwurfs ist über die ICA-Webseite verfügbar: https://www.ica.org/en/records-in-contexts-conceptual-model (13.05.2022).

treten und berichtet auf dem projekteigenen Blog[24] regelmäßig über die eigene Arbeit.

Der Kompetenzaufbau zum Einsatz von KI in Archiven benötigt zusätzliche Ressourcen. Ihr Einsatz dürfte sich aber mehr als auszahlen, wenn es die Archive schaffen, sich dadurch als moderne Institutionen der deutschen Wissenschaftsinfrastruktur zu präsentieren. Im Kontext des Aufbaus der Nationalen Forschungsdateninfrastruktur (NFDI) geht es auch darum, hier nicht den Anschluss zu verlieren. Für das Landesarchiv Baden-Württemberg hat das FDMLab einen Anfang gemacht. Eine Fortführung der Arbeit des FDMLabs über die aktuelle Projektlaufzeit hinaus wird angestrebt.

Autoreninformationen

Benjamin Rosemann
Landesarchiv Baden-Württemberg
Zentrale Dienste
Projekt FDMLab@LABW
Eugenstraße 7
70182 Stuttgart
benjamin.rosemann@la-bw.de
orcid.org/0000-0002-0780-3979

Elisabeth Klindworth
Landesarchiv Baden-Württemberg
Archivischer Grundsatz
Projekt FDMLab@LABW
Eugenstraße 7
70182 Stuttgart
elisabeth.klindworth@la-bw.de
orcid.org/0000-0003-1848-5870

24 Projektblog des FDMLabs unter https://fdmlab.landesarchiv-bw.de/ (13.05.2022).