Gerrit Heim

Named Entity Recognition in Digitalen Sammlungen – Ein Werkstattbericht aus der Badischen Landesbibliothek

Named Entity Recognition in Digital Collections – A Workshop Report from the Baden State Library

https://doi.org/10.1515/bd-2023-0039

Zusammenfassung: Die Badische Landesbibliothek hat im Rahmen eines Pilotprojekts die Named Entity Recognition (NER) in den Digitalen Sammlungen für ausgewählte Zeitungsbestände realisiert. Grundlage ist eine technische Neuentwicklung in Visual Library, die auf Google Cloud Natural Language basiert. Diese ermöglicht die Erkennung von Normdaten und deren Verknüpfung in den mittels OCR generierten Volltexten. Diese Datenanreicherung schafft neue Rechercheeinstiege für die Nutzerinnen und Nutzer und ermöglicht die Anwendung neuer Recherchemethoden für die Wissenschaft.

Schlüsselwörter: Normdaten, AI, Digital Humanities, Kulturgutdigitalisierung, Erfahrungsbericht

Abstract: As part of a pilot project, the Baden State Library has implemented Named Entity Recognition (NER) for selected newspaper holdings in their Digital Collections. Using a technical innovation in Visual Library based on Google Entities, this enables the recognition and linking of authority data in the full texts generated by OCR. The data enhancement has created new entry points for researchers and users, providing for new research methods in the sciences and humanities.

Keywords: Authority data, AI, digital humanities, digitalization of cultural assets, field report

Gerrit Heim: heim@blb-karlsruhe.de

1 Einleitung

Die Badische Landesbibliothek betreibt seit über 10 Jahren erfolgreich ein umfangreiches Kulturgutdigitalisierungsprogramm mit einer leistungsfähigen hausinternen Digitalisierungswerkstatt.[1] Die Digitalisierung soll eine breite Verfügbarkeit der herausragenden Bestände der Badischen Landesbibliothek für die interessierte Öffentlichkeit und die wissenschaftliche Forschung ermöglichen.

Der gesamte Digitalisierungsworkflow und die Präsentation erfolgen auf Basis von Visual Library von semantics.[2] Diese Wahl wurde 2010 vor dem Hintergrund der damals verfügbaren Produkte und unter Berücksichtigung der Leistungsfähigkeit der IT-Abteilung getroffen. Diese Entscheidung hat sich im Nachhinein als sehr positiv herausgestellt, da die Badische Landesbibliothek hier in Zusammenarbeit mit semantics die Digitalisierung stetig weiterentwickeln und neue innovative Funktionen etablieren konnte.

Digitalisierung bedeutet längst nicht mehr nur die Produktion und Bereitstellung von Digitalisaten.[3] Dies ist zwar immer noch das unerlässliche Kerngeschäft eines Digitalisierungsprogramms, doch verschiebt sich der Fokus sukzessive auf die Erzeugung hochwertiger Forschungsdaten. Hierzu stellt die BLB qualitativ hochwertige Volltexte mittels OCR für gedruckte Materialien in Antiqua und Fraktur bereit und optimiert diese Prozesse stetig. Die weitgehend automatisierte Volltexterkennung macht aber auch vor Handschriften nicht mehr halt. Digitalisierte Handschriften werden mittels Transkribus aufbereitet, was inzwischen ebenso wichtig ist, wie die ständige Erprobung neuer Werkzeuge in diesem Bereich.

Die wissenschaftlichen Bedarfe der Digital Humanities und die Anforderungen einer breiten öffentlichen Nutzergruppe stehen dabei nicht im Widerspruch. Dies lässt sich sehr gut am Beispiel der Normdaten verdeutlichen. Die Verknüpfung von qualitativ hochwertigen Volltexten mit Normdaten ermöglicht es mit neuen Fragestellungen der Digital Humanities, an diese Texte heranzugehen, aber gleichzeitig auch innovative Rechercheeinstiege für alle Nutzerinnen und Nutzer der Digitalen Sammlungen. Die Entitätenerkennung in Volltexten mittels des NER-Verfahrens soll diese spannenden Perspektiven verdeutlichen.

---

1 Syré, Ludger: Aufbruch in eine neue Zeit: Die Anfänge der Digitalisierungswerkstatt und der Digitalen Sammlungen an der Badischen Landesbibliothek. In: Siebert, Irmgard (Hg.): Digitalisierung in Regionalbibliotheken. Frankfurt am Main 2012, S. 173–194. Syré, Ludger: Die Digitalisierung feiert Geburtstag – zehn Jahre Digitale Sammlungen der Badischen Landesbibliothek. In: BIT online – Bibliothek, Information, Technologie 24.1 (2021), S. 72–82.

2 https://digital.blb-karlsruhe.de [Zugriff: 23.03.2023].

3 Schütte, Jana Madlen: Die Zukunft der Kulturgutdigitalisierung an Landesbibliotheken am Beispiel der Badischen Landesbibliothek (BLB). In: Bibliotheksdienst 56.2 (2022), S. 103–114, https://doi.org/10.1515/bd-2022-0021.

2 Digitalisierungskonzept

Das in der Planungsphase erarbeitete Digitalisierungskonzept der Badischen Landesbibliothek hat sich im Kern bewährt und wird bis heute verfolgt und dabei kontinuierlich weiterentwickelt. Im Mittelpunkt stehen die unikalen Bestände der Badischen Landesbibliothek. Diese Schwerpunktsetzung resultiert nicht zuletzt aus den Erfahrungen der Vergangenheit. Die vollständige Zerstörung in der Nacht vom 2. auf den 3. September 1942 betraf nicht nur das Gebäude, sondern auch alle nicht ausgelagerten Bestände und schuf ein bis heute anhaltendes Bewusstsein für die Bedeutung der langfristigen Sicherung und öffentlichen Zugänglichkeit der unikalen Bestände und über Jahrzehnte mühsam wieder aufgebauten regionalen Sammlung.

Die Badische Landesbibliothek hat daher frühzeitig mit der Digitalisierung mittelalterlicher und frühneuzeitlicher Handschriften aus ihren Beständen begonnen. So waren die Digitalisierungsprojekte in diesem Bereich bereits weit fortgeschritten, als die DFG 2016 eine nationale Strategie zur Handschriftendigitalisierung auf den Weg brachte. Wichtige Bestände sind heute bereits digitalisiert und ein erfolgreicher Abschluss der Handschriftendigitalisierung zeichnet sich ab.

Einen weiteren Schwerpunkt der Digitalisierung bilden die Musikalien. Ausgehend von den markgräflichen Beständen der Hofbibliothek ist hier durch Neuerwerbungen der letzten Jahrzehnte eine herausragende Sammlung entstanden. Die Badische Landesbibliothek bewahrt die Sammlungen säkularisierter Klöster und die Hofmusiken aus den Residenzen Karlsruhe, Rastatt und Baden-Baden sowie zahlreiche bedeutende Nachlässe von Komponisten und Musikern der badischen Musikgeschichte.[4] Von besonderer Bedeutung sind die Musikhandschriften und Musikdrucke der Fürstlich Fürstenbergischen Hofbibliothek Donaueschingen. Die Digitalisierung der herausragenden Donaueschinger Musikalien ist bereits weit fortgeschritten und wird voraussichtlich 2025 abgeschlossen werden.

In ihrer Eigenschaft als eine von zwei Landesbibliotheken des Bundeslandes Baden-Württemberg ist die Badische Landesbibliothek nicht nur ihren Sondersammlungen verpflichtet, sondern sammelt darüber hinaus landeskundlich relevante Informationen und stellt diese einer breiten Öffentlichkeit zur Verfügung. Es war daher von Beginn an klar, dass die Digitalisierungsstrategie als dritte Säule neben Handschriften und Musikalien regional bedeutsame Werke und Quellen beinhalten muss. Es sind vor allem diese Bestände, die das Digitalisierungspro-

4 Geyer, Brigitte; Knödler-Kagoshima, Brigitte; Krumeich, Kirsten u. a.: Musiknoten digital. Zum Stand der Musikaliendigitalisierung in Deutschland. In: Zeitschrift für Bibliothekswesen und Bibliographie 69.4 (2022), S. 196–209, https://doi.org/10.3196/186429502069422.

gramm einer Landesbibliothek von vergleichbaren Digitalisierungsprojekten anderer wissenschaftlicher Bibliotheken unterscheiden.

Eine besondere Kategorie innerhalb der Regionalia bilden die badischen Zeitungen. Die Badische Landesbibliothek hat 2014 ihre umfangreichen Bestände an historischen Zeitungen in ihre Digitalisierungsstrategie aufgenommen.[5] Dank verschiedener Förderprogramme konnten hier zügig umfangreiche Bestände digitalisiert werden und seit 2019 ist die Digitalisierung der historischen Tageszeitungen aus der Region Karlsruhe weitgehend abgeschlossen.[6] Die digitalisierten Periodika erfreuen sich einer besonders hohen Nachfrage in der geschichtswissenschaftlichen Forschung und bei der interessierten Öffentlichkeit. Mehr noch als bei anderen Beständen handelt es sich bei der Zeitungsdigitalisierung um ein Massengeschäft. So machen Zeitungen mit über 2,1 Millionen Images etwas weniger als die Hälfte der insgesamt 4.6 Millionen Images in den Digitalen Sammlungen der Badischen Landesbibliothek aus. Ab 2019 verlagerte sich der Fokus von der Tagespresse auf thematisch spezifischere Zeitungen aus ausgewählten Themenbereichen wie Verkehrswesen in Baden oder dem Bereich der Wirtschaft. Dadurch konnten zahlreiche weitere Presseerzeugnisse digitalisiert und für spezifische Forschungsfragen zur Verfügung gestellt werden. Diese besondere Bedeutung der digitalisierten Zeitungen prädestiniert diesen Bestand für exemplarische Datenanreicherungsprojekte.

3 Einstiege in die Recherche

Recherchen in diesen ständig wachsenden digitalen Beständen stellen eine Herausforderung für die Nutzenden dar. Eine wichtige Rolle spielt dabei natürlich eine qualitativ hochwertige OCR,[7] die eine direkte Recherche in den Beständen über Schnell- und Detailsuche ermöglicht. Neben der OCR mit Abbyy kommen hier seit einiger Zeit Tesseract-Modelle aus dem Projekt OCR-BW[8] zum Einsatz.

Darüber hinaus entwickelt die Badische Landesbibliothek seit geraumer Zeit spezifische Recherchemöglichkeiten für ihre Digitalen Sammlungen, um die

5 https://digital.blb-karlsruhe.de/zeitungen/topic/view/2965491 [Zugriff: 23.03.2023].

6 Im Sinne einer kooperativen Digitalisierung haben auch die Universitätsbibliotheken Heidelberg und Freiburg, das Marchivum in Mannheim sowie das Kreisarchiv Calw Zeitungbestände digitalisiert und tragen gemeinsam dazu bei, die umfangreichen historischen badischen Zeitungsbestände zumindest in Teilen digital verfügbar zu machen.

7 Hertling, Anke; Klaes, Sebastian: Volltexte für die Forschung: OCR partizipativ, iterativ und on Demand. In: o-bib. Das offene Bibliotheksjournal (2022), S. 1–11; https://doi.org/10.5282/O-BIB/5832.

8 https://ocr-bw.bib.uni-mannheim.de [Zugriff: 23.03.2023].

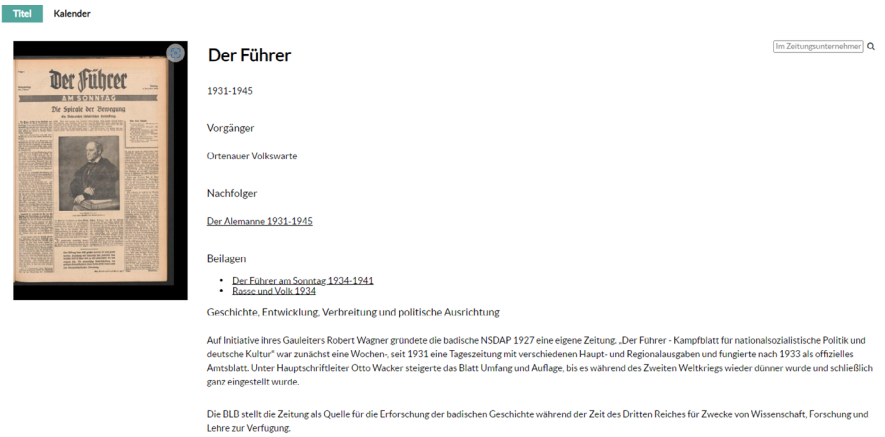

Abb. 1: Das Zeitungsunternehmen „Der Führer" mit Kontextinformationen.

umfangreichen Bestände leicht zugänglich zu machen. Diese orientieren sich an den spezifischen Anforderungen der digitalisierten Bestände und sollen einen präzisen Zugang zu den Digitalen Sammlungen ermöglichen. In diese Kategorie fällt beispielsweise das Geographica-Tool, das die Literatur zum Oberrhein georeferenziert darstellt[9] und eine gezielte Recherche in der Rheinliteratur ermöglicht. Ein anderes Beispiel findet sich bei den Landtagsprotokollen. Für diesen prominenten Bestand der digitalisierten Regionalia steht zur zielgerichteten Suche eine Personen- und Redensuche zur Verfügung, um den umfangreichen Bestand besser zugänglich zu machen.[10]

Speziell für die Zeitungen bieten die Digitalen Sammlungen den Nutzerinnen und Nutzern zwei verschiedene Zugänge. Die derzeit über 126 Zeitungen und ihre Beilagen sind in so bezeichneten Zeitungsunternehmen organisiert.[11] Diese zeigen übersichtlich die Vorgänger und Nachfolger, listen die unterschiedlichen Beilagen auf und liefern Basisinformationen zur jeweiligen Zeitung. Nutzende können sich so schnell orientieren. Über den Kalendereinstieg können die jeweiligen Ausgaben tagesgenau abgerufen werden.

9 Schütte, Jana Madlen: Geographica digital – Ansichten und Denkmäler aus dem Oberrheingebiet in den Digitalen Sammlungen der Badischen Landesbibliothek. In: Bibliotheksdienst 54.7–8 (2020), S. 565–576, https://doi.org/10.1515/bd-2020-0071.

10 https://digital.blb-karlsruhe.de/topic/view/792873 [Zugriff: 23.03.2023].

11 Beispielhaft: Das Karlsruher Tagblatt.

Abb. 2: Zielgerichtete Suche in Zeitungen über den Kalender.

4 Named Entity Recognition in Digitalen Sammlungen

Die Normdatenverknüpfung in den deskriptiven Daten ist ein zentraler Bestandteil der Digitalen Sammlungen der Badischen Landesbibliothek. Getreu dem Motto *„Kein Digitalisat ohne Katalogisat“* hat der bibliothekarische Bereich der Katalogisierung und Verknüpfung mit Normdaten seit jeher eine zentrale Bedeutung für den Digitalisierungsprozess. Eine Verknüpfung des Verfassers mit GND und Wikipedia ist daher ebenso selbstverständlich wie die Bereitstellung der Digitalisate über Wikisource und andere Plattformen. Erst die Verknüpfung mit Normdaten ermöglicht die Auffindbarkeit für Mensch und Maschine und eine weitere automatisierte Datenanreicherung im Rahmen sich stetig weiter entwickelnder Forschungsmethoden.

Die Erkennung von benannten Entitäten in den Volltexten digitaler Sammlungen wird seit einigen Jahren als eine Möglichkeit diskutiert, um die Digitalisierung weiterzuentwickeln. Das zugrundeliegende NER-Verfahren gilt als sehr zuverlässig und die automatische Extraktion von Entitäten aus Volltexten als ein in der Linguistik gelöstes Problem.[12] Lediglich die Umsetzung in den verbreiteten Digitalisierungssoftwarelösungen steht noch aus. Die Badische Landesbibliothek hat daher 2022 als Vorreiter an einem Projekt zur Named Entity Recognition (NER) im Rahmen von Visual Library teilgenommen und daran mitgearbeitet, die Normdatenverknüpfung noch tiefer in die Digitalen Sammlungen zu integrieren.

12 Schumacher, Mareike: Named Entity Recognition (NER), forText. Literatur digital erforschen (2018), https://fortext.net/routinen/methoden/named-entity-recognition-ner [Zugriff: 23.02.2023]; More, Jacqueline: Theorie und Anwendung von Named Entity Recognition in den Digital Humanities mit Fokus auf historische Texte des 17. Jahrhunderts, Masterthesis. Graz 2021, https://resolver.obvsg.at/urn:nbn:at:at-ubg:1-166482 [Zugriff: 23.03.2023].

Diese Eigennamenerkennung oder Erkennung von benannten Entitäten ist ein Teil der maschinellen Aufbereitung von Texten. Mit diesem Verfahren werden Eigennamen, Organisationen und Themen in Texten identifiziert, kategorisiert und aufbereitet. Die technische Basis bildet in der Umsetzung innerhalb von Visual Library die Technologie „Google Entities“ und die zugehörige „Natural Language API (NLP)“, die in natürlicher Sprache nach bekannten Entitäten sucht.[13] Die Google API basiert auf fortlaufend trainierten Modellen zur Erkennung von Entitäten, wodurch das NER-Verfahren einen praktischen Anwendungsfall von Machine learning-Verfahren im Bereich der Digitalisierung darstellt. Voraussetzung ist eine qualitativ hochwertige OCR als Basis für das Verfahren. Die Google API ermittelt anschließend in den Volltexten die benannten Entitäten, verlinkt diese mit Wikidata und liefert diese Informationen als JSON zurück. Hierauf aufbauend identifiziert die Lösung von semantics verlinkte GND-Informationen aus Wikidata, um die Verbindung von benannter Entität und GND herzustellen. Diese Daten kombiniert die Visual Library mit der generierten OCR (hOCR) für die Darstellung auf der Oberfläche und speist daraus die verschiedenen Indices für Personen, Themen und Orte in der Visual Library.

Für das aus Sondermitteln des Ministeriums für Wissenschaft, Forschung und Kunst 2021 finanzierte Pilotprojekt wurden mehrere Margen zusammengestellt und nacheinander mit NER angereichert. Das Projekt begann mit zwei eher kleinen Zeitungen, deren Erscheinungszeitraum nach 1945 lag. Der Umfang wurde dann sukzessive über vier Margen hinweg gesteigert. Das Ziel war es, eine möglichst große Auswahl an Zeitungen zusammenzustellen, die eine ausreichend große Testmenge bildeten und gleichzeitig eine gewisse thematische Bandbreite gewährleisteten. Die Named Entity-Erkennung erfolgte daher über ausgewählte Zeitungen aus dem Themenbereich Wirtschaft, typische kleinere regionale Zeitungen sowie große Tageszeitungen nach 1945.

Zeitung	Jahrgänge	VL-ID
Der Südkurier	1945–1952	https://digital.blb-karlsruhe.de/6993401
Der neue Tag / Unser Tag	1947–1950	https://digital.blb-karlsruhe.de/6617602
Badisches Volksecho	1946–1950	https://digital.blb-karlsruhe.de/6992162
Das Volk	1946–	https://digital.blb-karlsruhe.de/6597585
Südwestdeutsche Volkszeitung	1946–1949	https://digital.blb-karlsruhe.de/6992164
Bruchsaler Post	1950–1952	https://digital.blb-karlsruhe.de/6073670
Bauländer Bote und Boxberger Anzeiger	1914–1918	https://digital.blb-karlsruhe.de/6585155
Badische Warte	1917–1920	https://digital.blb-karlsruhe.de/6120600
Offenburger Wochenblatt	1796–	https://digital.blb-karlsruhe.de/6594362
D’r Alt Offeburger	1899–1933	https://digital.blb-karlsruhe.de/6602613

13 https://cloud.google.com/natural-language/docs/analyzing-entities?hl=de; https://cloud.google.com/natural-language?hl=de [Zugriffe: 17.02.2023].

Zeitung	Jahrgänge	VL-ID
Bericht der Badischen Industrie- und Handelskammer, Karlsruhe, über die Wirtschaftslage in Baden	1934	https://digital.blb-karlsruhe.de/6759378
Bericht der badischen Industrie- und Handelskammern über die Wirtschaftslage in Baden	1935–1936	https://digital.blb-karlsruhe.de/6488881
Bericht der im Badischen Industrie- und Handelstag vereinigten Handelskammern (Freiburg, Heidelberg, Karlsruhe, Konstanz, Lahr, Mannheim, Pforzheim, Schopfheim, Villingen) über die Wirtschaftslage in Baden: im ... Vierteljahr	1930–1932	https://digital.blb-karlsruhe.de/6488880
Berichte des Badischen Gewerbeaufsichtsamtes: erstattet an das Ministerium der Finanzen und Wirtschaft	1931–1936	https://digital.blb-karlsruhe.de/6488879
Jahres-Bericht der Großherzoglich Badischen Fabrik-Inspektion / erstattet an Großherzogliches Ministerium des Innern	1888–1910	https://digital.blb-karlsruhe.de/6488877
Jahres-Bericht des Großherzoglich-Badischen Fabrik-Inspektors / veröffentlicht auf Anordnung des Großherzoglich-Baden'schen Ministeriums für Handel	1879	https://digital.blb-karlsruhe.de/6488876
Jahresbericht des Bad. Gewerbeaufsichtsamtes und des Bad. Bergamtes: erstattet an das Ministerium des Innern	1911–1930	https://digital.blb-karlsruhe.de/6488878
Badische Gewerbe- und Handwerkerzeitung	1910–1020	https://digital.blb-karlsruhe.de/6485128
Heimat und Handwerk	1910–1925	https://digital.blb-karlsruhe.de/6485130
Badische Gewerbezeitung	1867–1909	https://digital.blb-karlsruhe.de/6485127
Das Badische Handwerk	1921–1931	https://digital.blb-karlsruhe.de/6485129
Badische Wirtschaftszeitung	1922–1942	https://digital.blb-karlsruhe.de/6485132
Lebendiges Handwerk	1942	https://digital.blb-karlsruhe.de/ 6485131
Monatsheft der Technik	1935–1940	https://digital.blb-karlsruhe.de/ 6488875
Oberrheinisches Wirtschaftsblatt	1943–1944	https://digital.blb-karlsruhe.de/ 6485133
Das Wirtschaftsjahr in Baden	1937	https://digital.blb-karlsruhe.de/ 6488882

Insgesamt konnten auf diese Weise 18.464 Orte und 35.612 Personen und Körperschaften identifiziert werden. Dafür verarbeitete das System 127.486 Seiten.

Die Erkennungsqualität war dabei sehr hoch. Dies ist umso wichtiger, als ein effizienter Personaleinsatz, ähnlich wie bei der OCR, eine intensive Bereinigung solch automatisiert generierter Daten nicht zulässt. Allerdings lassen sich bei Bedarf einige Normdatengruppen ausschließen, die mit hoher Wahrscheinlichkeit nicht in den Digitalisaten vorkommen und der Prozess sich somit steuert. Eine stichprobenartige Qualitätssicherung anhand der Zeitung „Das Volk" ergab null bis maximal sechs

Abb. 3: Zusätzliche Rechercheeinstiege nach NER im Zeitungsunternehmen.

Fehler pro Seite. Dabei konnten einige neuralgische Punkte ermittelt werden. Bindestriche sind ebenso ein Problem wie Abkürzungen und Umlaute. Beispielsweise erkannte NER bei dem Wort „über“ eine Verknüpfung zum kalifornischen Start-up „Uber“. In der Abkürzung Freiburg i. Br., die in badischen Zeitungsbeständen natürlich häufig vorkommt, ermittelte die Software den Bayerischen Rundfunk (BR).

Zeitungsunternehmen, die mit NER aufbereitet wurden, erhalten in Visual Library neue Einstiege für die erkannten Entitäten.

Dadurch stehen den Nutzerinnen und Nutzern nun Reiter für Personen/Körperschaften, Orte und Themen zur Verfügung, um gezielt an die entsprechenden Stellen in den Zeitungen springen zu können.

Semantics hat die Darstellung in den letzten Monaten auf Basis der Erfahrungen der Badischen Landesbibliothek und anderer Erstanwender kontinuierlich weiterentwickelt und konnte den Nutzen in den letzten Monaten weiter erhöhen. Personennamen oder Orte, die in der Schnellsuche eingegeben werden, zeigen nun in der Trefferliste und in der Vollanzeige sofort den entsprechenden Normdatensatz und die zugehörigen Fundstellen an.

Im Hintergrund bietet das VL-TextLab den Mitarbeiterinnen und Mitarbeitern der Digitalisierung die Möglichkeit, durch NER ermittelte Normdaten zu überarbeiten und einzelne Entitäten zu korrigieren.

Bei der Vorbereitung weiterer Erkennungsprojekte kann die Digitalisierungsabteilung über dieses Werkzeug die notwendigen NER-Einheiten für einen definierten Bestand prognostizieren. Dabei handelt es sich um eine akkurate Schätzung auf Basis der erkannten Zeichen. Diese Prognose der NER-Einheiten ist für die Abwicklung in Form der bei semantics üblichen Volumenpakete notwendig.

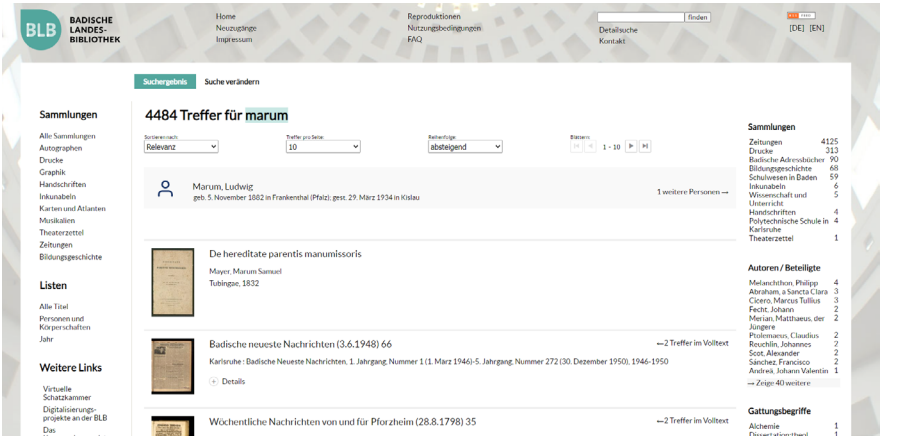

Abb. 4: Suche nach „marum“ zeigt Normdatensatz zu Ludwig Marum in der Trefferliste.

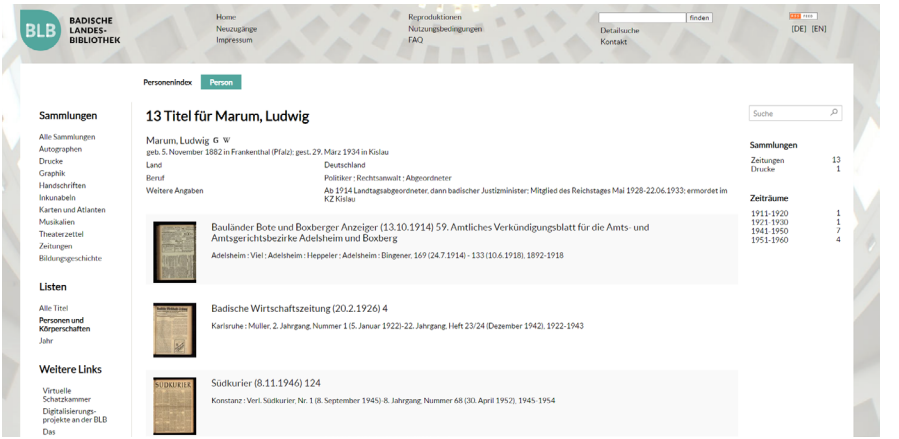

Abb. 5: Biographische Informationen und Verweise auf GND und Wikipedia bei Treffern zu Ludwig Marum.

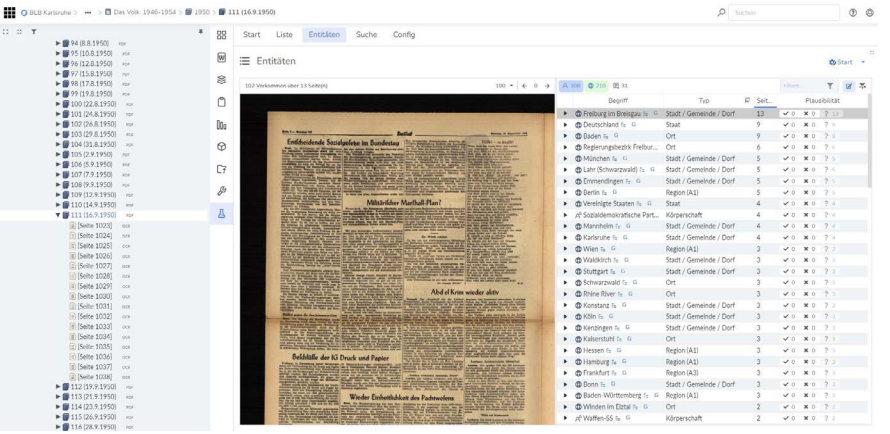

Abb. 6: Administrationsmöglichkeiten für erkannte Normdaten im VL-TextLab.

5 Herausforderungen

Eine Herausforderung für die strategische Planung einer solchen technologischen Erweiterung stellt der große Bestand an bereits digitalisierten Zeitungen dar. Ähnlich wie bei der nachträglichen OCR-Erkennung, die in der Badischen Landesbibliothek nahezu abgeschlossen ist, bleibt für die Nutzerinnen und Nutzer schwierig zu erkennen, welche Bestände bereits mit NER bearbeitet sind und welche nicht. Dies ist nicht an allen Stellen ein Problem. In den sogenannten Zeitungsunternehmen werden nach dem NER-Verfahren bei einigen Zeitungen zusätzliche Reiter angeboten. Die Trefferlisten in diesen Reitern sind aber vollständig für die jeweilige Zeitung, da hier bereits eine zielgerichtete Suche erfolgt und die Reiter sich nur auf die jeweilige Zeitung beziehen. Bei der Suche in den Personen- und Ortslisten ist für die Benutzerinnen und Benutzer jedoch nicht sofort ersichtlich, dass über die entsprechenden Trefferlisten nur ein Ausschnitt der Digitalen Sammlungen präsentiert wird. Gleiches gilt für die Normdatensätze in den Ergebnissen der Schnellsuche, die die Benutzerinnen und Benutzer fast unmerklich zu den mit NER versehenen Beständen führen.

Die Lizenzstruktur von Visual Library zur Abgeltung der Erschließungskosten für NER bedarf einer sorgsamen Kalkulation. Dies betrifft sowohl den bereits vorhandenen Bestand an Digitalisaten als auch zukünftige Digitalisierungsprojekte. Für eine weitgehende Implementierung von Normdaten in die Präsentation der digitalen Sammlung müsste ein retrospektiver Ausbau der NER-Erschließung erfolgen, vergleichbar mit der langjährigen retrospektiven OCR. Ebenso müsste bei zukünftigen Projekten die NER in die pauschalisierten Produktionskosten pro

Volltext-Statistik

	Gesamtzahl	NER-Ergebnis
Seiten	6.615	6.615
Blöcke	207.762	207.762
Absätze	780.645	780.645
Zeilen	3.083.275	3.083.275
Wörter	21.928.600	21.928.600
Zeichen	121.432.475	121.432.475
NER-Einheiten	124.742	124.742

Abb. 7: Kalkulation der benötigten NER-Einheiten auf Basis der mittels OCR erkannten Zeichen.

Image einbezogen werden. Die Badische Landesbibliothek hat durch ihre mehr als zehnjährige Kulturgutdigitalisierung eine so große digitale Sammlung aufgebaut, dass ein solches Projekt nur langfristig realisierbar wäre.

Dabei sind nicht alle Bestandsgruppen gleichermaßen erfolgversprechend. Grundvoraussetzung ist eine hinreichende Anzahl an Normdaten in den Volltexten. Die bisherigen Tests erfolgten ausschließlich mit Zeitungsbeständen. Auch wenn es verlockend ist, diesen erfolgreichen Weg weiter zu beschreiten, sind für die Zukunft auch andere Regionalia denkbar. Ein mögliches Entwicklungsprojekt wäre ein weiterer stark nachgefragter Bestand aus der Säule Regionalia: die Landtagsprotokolle.[14]

Zusammenfassend bleibt die Anreicherung der volltexterkannten Zeitungsbestände mit NER trotz einiger unvermeidbarer Defizite bei solch automatisierten Verfahren ein großer Erfolg und ermöglicht eine präzisere Suche in umfangreichen Beständen sowie gezielte Einstiege über Orte, Themen und Personen. Die Anwendung von Machine-learning-Verfahren, um Digitalisate mit Normdaten zu verknüpfen, schafft Mehrwerte über die reine digitale Verfügbarkeit der historischen Bestände hinaus und bietet ganz neue Möglichkeiten für die Forschung.

Gerrit Heim
Leiter der Abteilung Regionalia
Fachreferent für Geschichte
Badische Landesbibliothek
Erbprinzenstraße 15
76133 Karlsruhe
Deutschland
E-Mail: **heim@blb-karlsruhe.de**

14 Syré, Ludger: Die Protokolle des Badischen Landtags in digitaler Form: der Beitrag der Badischen Landesbibliothek zum Landesjubiläum. In: Badische Heimat 93.2 (2013), S. 607 KB, https://doi.org/10.57962/REGIONALIA-450.