Sina Menzel, Hannes Schnaitter, Josefine Zinck, Vivien Petras,
Clemens Neudecker, Kai Labusch, Elena Leitner, Georg Rehm

Named Entity Linking mit Wikidata und GND – Das Potenzial handkuratierter und strukturierter Datenquellen für die semantische Anreicherung von Volltexten

1 Einführung

Named Entities[1] (benannte Entitäten) – wie Personen, Organisationen, Orte, Ereignisse und Werke – sind wichtige inhaltstragende Komponenten eines Dokuments und sind daher maßgeblich für eine gute inhaltliche Erschließung. Die Erkennung von Named Entities, deren Auszeichnung (Annotation) und Verfügbarmachung für die Suche sind wichtige Instrumente, um Anwendungen wie z. B. die inhaltliche oder semantische Suche in Texten, dokumentübergreifende Kontextualisierung oder das automatische Textzusammenfassen zu verbessern. Inhaltlich präzise und nachhaltig erschlossen werden die erkannten Named Entities eines Dokuments allerdings erst, wenn sie mit einer oder mehreren Quellen verknüpft werden (Grundprinzip von Linked Data, Berners-Lee 2006), die die Entität eindeutig identifizieren und gegenüber gleichlautenden Entitäten disambiguieren (vergleiche z. B. *Berlin* als Hauptstadt Deutschlands mit dem Komponisten *Irving Berlin*). Dazu wird die im Dokument erkannte Entität mit dem Entitätseintrag einer Normdatei oder einer anderen zuvor festgelegten Wissensbasis (z. B. Gazetteer für geografische Entitäten) verknüpft, gewöhnlich über den persistenten Identifikator der jeweiligen Wissensbasis oder Normdatei. Durch die Verknüpfung mit einer Normdatei erfolgt nicht nur die Disambiguierung und Identifikation der Entität, sondern es wird dadurch auch Interoperabilität zu anderen Systemen hergestellt, in denen die gleiche Normdatei benutzt wird, z. B. die Suche nach der Hauptstadt Berlin in verschiedenen Datenbanken bzw. Portalen. Die Entitätenverknüpfung (*Named Entity Linking*, NEL) hat zudem den Vorteil, dass die Normdateien oftmals Relationen zwischen Entitäten enthalten, sodass Dokumente, in denen Named Entities erkannt wurden, zusätzlich auch im Kontext einer größeren Netzwerkstruktur von Entitäten verortet und suchbar gemacht werden können (z. B. die Ausweitung einer Suche

1 Wir verwenden in diesem Beitrag nicht den deutsch- (*Entitys*), sondern den englischsprachigen Plural (*Entities*).

https://doi.org/10.1515/9783110691597-012

von der Hauptstadt Berlin auf andere Städte in Deutschland über die Hierarchiebeziehung *Berlin → Deutschland*). Damit ist das Named Entity Linking eine Methode zur Inhaltserschließung von Volltexten und ermöglicht neben der semantisch strukturierten Volltextsuche nach bestimmten Entitäten oder Entitätentypen (z. B. nur Personen) in einer Sammlung auch die Verknüpfung von Volltextinhalten mit anderen Systemen mit identischen Normdatensätzen für eine sammlungsübergreifende, netzwerk- oder graphbasierte Suche.

Eine erfolgreiche und damit qualitativ hochwertige Entitätenverknüpfung hängt im Wesentlichen davon ab, wie viele der erkannten Entitäten korrespondierenden Normdatensätzen zugewiesen werden können (Abdeckung oder *recall*) und ob – dies ist besonders bei automatischen, computerlinguistischen Verfahren ohne intellektuelle Kontrolle entscheidend – lexikalische Ambiguitäten überwunden werden können, um die Disambiguierung und eine exakte und korrekte Identifikation der Entität zu gewährleisten (Genauigkeit oder *precision*). Dies hängt einerseits von der zur Verknüpfung benutzten Normdatei und andererseits von den eingesetzten maschinellen Disambiguierungsverfahren ab.

In diesem Beitrag präsentieren wir eine Studie zur Entitätenerkennung und -verknüpfung in historischen, retrodigitalisierten Zeitungstexten, die die Abdeckung und Verknüpfungsqualität zweier verschiedener Normdateien für die inhaltliche Suche miteinander vergleicht: die im deutschsprachigen Raum verwendete bibliothekarische Normdatei Gemeinsame Normdatei (GND) und die globale Wissensdatenbank Wikidata. Die stark qualitätskontrollierte GND ist insbesondere auf Entitäten fokussiert, die publizierend tätig waren, auch wenn die Normdatei mittlerweile stark erweitert wird (Balzer et al. 2019). Wikidata enthält dagegen eine sehr viel größere Anzahl von *Objekten* (so benannt in Wikidata), unterliegt aber bislang einer weniger ausgeprägten Qualitätskontrolle.[2] Es ist unklar, welche Normdatei besser für eine Verknüpfung geeignet ist. In experimentellen Ansätzen der Sprachverarbeitung wird Wikidata, ähnlich wie z. B. auch DBpedia, oft wegen seiner großen Abdeckungsrate der Vorzug gegeben, in der bibliothekarischen Erschließung bevorzugt man die GND, insbesondere aufgrund ihrer qualitätsgeprüften Angaben und Relationen (Hochstein 2011; Piscopo und Simperl 2019; Voß et al. 2014). Die vorliegende Studie fokussiert insbesondere auf eine Fehleranalyse, um die Herausforderungen der Entitätenverknüpfung mit den beiden Quellen aufzuzeigen.

Der Beitrag ist wie folgt strukturiert: In Abschnitt 2 wird der aktuelle Forschungsstand zu Verfahren der Entitätenverknüpfung sowie aktuelle Evaluati-

2 Vgl. dazu den Beitrag *Wissensbasen für die automatische Erschließung und ihre Qualität am Beispiel von Wikidata* in dem vorliegenden Band.

onsansätze zur Messung der Ergebnisqualität, einschließlich typischer Fehlerkategorien beschrieben. Abschnitt 3 gibt einen Überblick über die GND und Wikidata, wobei auch die verwendeten Entitätenkategorien (Typen) sowie die Bereinigungsprozesse zur Verarbeitung der Wissensbasen für die Entitätenverknüpfung dargestellt werden. Abschnitt 4 beschreibt das Projekt *SoNAR*, in dem die Studie durchgeführt wurde, und stellt die eingesetzten Textkorpora und Algorithmen für die Entitätenerkennung und -verknüpfung vor. Der Evaluationsabschnitt 5 beschreibt das Forschungsdesign der Studie zum Vergleich beider Quellen und legt die intellektuell evaluierten Resultate vor, wobei besonders Fehlerkategorien beschrieben werden. Abschnitt 6 zieht ein Fazit und diskutiert das Potenzial beider Quellen für die Inhaltserschließung von Entitäten in historischen Zeitungstexten.

2 Named Entity Linking – Stand der Forschung

2.1 Verfahren des Named Entity Linking

Named Entity Linking (NEL) beschreibt den Prozess der Verknüpfung von Entitäten mit dem korrespondierenden Datensatz einer Normdatei oder, allgemeiner, einer Wissensbasis (*Knowledge Base*, Balog 2018: 148 f.). Die Entitäten müssen zuvor in einem gegebenen Text identifiziert werden (*Named Entity Recognition*, NER). Die einzelnen Normdatensätze sind dabei in der Regel mit persistenten Identifikatoren referenzierbar. Der Verknüpfungsprozess führt über die Disambiguierung hin zur Identifizierung des korrespondierenden Normdatensatzes unter den wahrscheinlichsten Kandidaten und schließlich zur Verknüpfung der Zeichenkette der Entität, wobei jedoch nicht bei allen NEL-Ansätzen eine Disambiguierung stattfindet.

Konkrete NEL-Ansätze unterscheiden sich bzgl. der angewandten Techniken, Textgenres/-arten, Kategorien der ausgezeichneten Entitäten und verwendeten Wissensbasen (Rao et al. 2013; Shen 2015; Sevgili 2020). Nach derzeitigem Stand erzielen Ansätze, die auf neuronalen Netzen aufbauen (*Deep Learning*), bessere Ergebnisse als Ansätze, die auf klassischen maschinellen Lernverfahren basieren. Neuronale Netze werden dabei für alle Verarbeitungsschritte verwendet, d. h. Erkennung sowie Disambiguierung, einschließlich Kandidatenauswahl und Bestimmung der wahrscheinlichsten Rangfolge der Kandidaten. *Word Embeddings* und *Entity Embeddings* (Ganea 2017) spielen eine wichtige Rolle für die erzielte Erkennungsleistung. Kolitsas et al. (2018) beschreiben ein neurona-

les Netzwerk-Modell, das sowohl lokale als auch globale Kontextinformation in Form von *Word*, *Entity* und *Mention Embeddings* benutzt und Entitätenerkennung mit Entitätenverknüpfung kombiniert. Für zeitgenössische Texte legen Yamada et al. (2020) mit einem BERT-Modell (Devlin et al. 2019) führende Ergebnisse vor, das kontextualisierte *Embeddings* der Wörter und Entitäten lernt.

Historische Texte, die im Projekt *SoNAR* im Vordergrund stehen (vgl. Abschnitt 4), enthalten im Vergleich zu zeitgenössischen Texten viele Herausforderungen für NEL. Sie liegen ursprünglich in gedruckter Form vor, wobei Bibliotheken seit einiger Zeit große Mengen dieser Texte mithilfe von *Optical Character Recognition* (OCR) sowie weiteren NLP-Verfahren, speziell für die Erschließung, digitalisieren. Diese Vorverarbeitungsschritte arbeiten nicht perfekt und erzeugen somit Fehler, ferner sind in historischen Texten Schreibweisen weniger standardisiert. Zudem unterliegen beispielsweise die Bezeichnungen von Orten und Organisationen einem stetigen Wandel. Um neue und insbesondere neuronale NER- und NEL-Verfahren für historische Zeitungstexte zu testen und die Ergebnisse systematisch zu evaluieren, wurde der Wettbewerb *Identifying Historical People, Places and other Entities* (HIPE)[3] im Rahmen der *Conference and Labs of the Evaluation Forum* (CLEF) 2020 organisiert. Hierbei wurden zwei unterschiedliche NEL-Problemstellungen betrachtet, einerseits *End-to-End*-NEL, d. h. inklusive NER, oder NEL-*only* wobei die zu verknüpfenden Entitäten bereits im Text markiert waren (Ehrmann et al. 2020).

Die besten Ergebnisse im Rahmen des HIPE-Wettbewerbs erzielte das L3i-System (Boros et al. 2020). Es besteht aus folgenden Komponenten:

1. Erzeugung von Wissensbasen für die englische, französische und deutsche Sprache
2. Erkennung der Entitäten mit einem verbesserten BERT-Modell
3. Erstellung von *Entity Embeddings* basierend auf dem Ansatz von Ganea (2017)
4. Disambiguierung der Entitäten mit einem *End-to-End*-Modell von Kolitsas (2018)
5. Kandidatenfilterung

Zusätzlich gibt es ein Modul für die Vorverarbeitung, das einen negativen Einfluss der Eingabetexte minimieren soll, denn je mehr OCR-bedingte Fehler existieren, desto schlechter sind die NER- und NEL-Ergebnisse (Ehrmann et al. 2020).

Insgesamt wurden drei Systeme entwickelt, die NEL in deutschsprachigen historischen Zeitungstexten realisieren: L3i (Boros et al. 2020), SBB (Labusch 2020) und UVA.ILPS (Provatorova et al. 2020), wobei letzteres *End-to-End*-NEL

3 https://impresso.github.io/CLEF-HIPE-2020/

arbeitet; L3i und SBB können sowohl für *End-to-End*- als auch für NEL-*only*-Problemstellungen verwendet werden. Der beste F1-Wert[4] für *End-to-End*-NEL liegt bei 0,534 mit einem strikten Maß[5] und bei 0,557 mit einem unscharfen Maß. Die besten F1-Werte für NEL-*only* sind mit 0,582 (strikt) sowie 0,602 (unscharf) etwas höher. Diese Ergebnisse stammen von L3i. Die zweitbesten F1-Werte konnte das Team der SBB erzielen. Für *End-to-End*-NEL erreicht das System F1-Werte von 0,389 und 0,403 und für NEL-*only* 0,445 und 0,461. Im Vergleich zum Französischen und Englischen sind die Ergebnisse für das Deutsche am niedrigsten. Dieser Umstand lässt sich dadurch erklären, dass weniger Trainingsdaten für das Deutsche als für das Französische existieren. Das *Interannotator Agreement* für Wikidata-Identifikatoren (genannt QID) ist mit 0,69 hier niedriger als bei anderen Sprachen.

2.2 Evaluationsansätze und Fehlertypen in der Entitätenverlinkung

Um die Güte von automatisiert erzeugten NEL-Ergebnissen insgesamt zu messen, ist der Vergleich mit menschlichen Annotationen (Gold-Standards) notwendig.

Hachey et al. (2013) empfehlen für die NEL-Evaluation eine Unterteilung in Kandidatensuche und Disambiguierung. Sie untersuchen dazu drei wegweisende NEL Systeme (Bunescu und Pasca 2006, Cucerzan 2007, und Varma et al. 2009) anhand einer Methodik, die zwischen den Schritten *Eigennamenextraktion aus Text*, *Kandidatensuche* und *Disambiguierung* unterscheidet. Bei dieser differenzierteren Betrachtung stellen sie erhebliche Unterschiede zwischen der Performanz der drei Systeme auf den verschiedenen Ebenen fest und schlagen daher detaillierte Metriken für die jeweiligen Aufgaben vor. Dabei betonen sie insbesondere die große Bedeutung der Kandidatensuche für die nachfolgenden Schritte, während gängige Metriken zumeist nur die Disambiguierung und das Ranking der Kandidaten angemessen betrachten. Zuletzt weisen sie auch auf die Dynamik von Änderungen in Wikipedia (gleiches gilt für Wikidata) als Wissensbasen hin, die eine Vergleichbarkeit der Evaluationsergebnisse von NEL über einen längeren Zeitraum schwierig machen.

---

4 Das F1-Maß kombiniert die *precision* und den *recall* mithilfe des gewichteten harmonischen Mittels.

5 Bei dem strikten (genannt *strict*) F1-Maß existiert eine richtige Verknüpfung, die als korrekt erkannt bewertet wird. Bei dem unscharfen (genannt *fuzzy*) F1-Maß wird eine Menge von verwandten Verknüpfungen als korrekt erkannt bewertet (Ehrmann et al. 2020).

Für den CLEF-HIPE-2020 Wettbewerb (Ehrmann et al. 2020) wurde für die Evaluierung ein *Scorer* implementiert, der auf der NER-Evaluation von Batista 2014 beruht. Die Verknüpfung einer Entität wird als Label interpretiert. Um also als korrekt gezählt zu werden, benötigt die Systemantwort nur ein überlappendes Link-Label mit dem Goldstandard. Wörtliche und metonymische Verknüpfungen werden getrennt bewertet. Die strikte NEL-Metrik berücksichtigt nur die am höchsten gerankte Vorhersage des Systems, während die Fuzzy-Metrik die Systemantworten um eine Reihe von historisch verwandten Entitäts-QIDs erweitert.

Ling et al. (2015) merken an, dass für die Annotation von Entitäten-Verlinkungen – anders als für die Annotation von Entitäten selbst – kaum Richtlinien existieren. Sie unterscheiden für die Evaluierung der automatisierten Verlinkung von Entitäten sechs Fehlertypen (Ling et al. 2015: 325 f.):

- Typ 1 Metonymie: Bei metonymischem Gebrauch einer Entität kann es vorkommen, dass die Voraussage, auf der die automatisierte Verlinkung basiert, stattdessen eine bzw. die nicht-metonymische Entität am höchsten rankt.
  Beispiel:
  Falsch „Der [Kreml]LOC hat entschieden."
  → Verlinkung auf Festungsgebäude
  Richtig „Der [Kreml]ORG hat entschieden."
  → Verlinkung auf russ. Staatsmacht
- Typ 2 Falscher Entitätentyp: Die Ursache dieses Fehlers liegt nicht im Schritt der Verlinkung von Entitäten, sondern in der vorangehenden Auszeichnung von Entitäten. Wird hier der falsche Typ zugeordnet, werden falsche Verlinkungen wahrscheinlicher.
  Beispiel:
  Falsch „Wir wohnen an der [Ostsee]ORG"
  → Verlinkung auf Ostsee-Sparkasse Rostock
  Richtig „Wir wohnen an der [Ostsee]LOC"
  → Verlinkung auf Ostsee (als Gewässer)
- Typ 3 Koreferenz: Umfasst Fehler in der Disambiguierung von Koreferenzen auf dieselbe Entität.
  Beispiel:
  Falsch „[Davids]PER Vater heißt auch [David]PER."
  → Verlinkung auf dieselbe Entität
- Typ 4 Kontextfehler: Dieser Typ fasst Fehler zusammen, die durch Homonymie entstehen, die trotz bestehendem Kontext nicht aufgelöst wurde.
  Beispiel:

Falsch „[Winston Churchill]PER verließ die [Downing Street]LOC im Jahr 1955."
→ Verlinkung auf den US-amerikanischen Schriftsteller
Richtig „[Winston Churchill]PER verließ die [Downing Street]LOC im Jahr 1955."
→ Verlinkung auf den britischen Premierminister

- Typ 5 Spezifität: Da die Einträge in Wissensbasen hierarchische Strukturen (Taxonomien) besitzen, kann es passieren, dass eine Verlinkung auf eine zu tiefe Hierarchieebene und damit auf einen zu spezifischen Eintrag erfolgt.
  Beispiel:
  „[Paris]LOC ist im Herbst am schönsten."
  → Verlinkung auf „Arrondissement du Louvre" als Stadtteil von Paris.
- Typ 6 Sonstige: Falsche Verlinkungen, die keinem der anderen Typen zugeordnet werden können.

In dieser Studie wurden auch die Qualität der Wissensbasen sowie des vorliegenden Textkorpus als signifikante Fehlerquellen identifiziert (vgl. Abschnitt 5).

3 Verwendete Datenquellen

3.1 Gemeinsame Normdatei (GND)

Die Gemeinsame Normdatei (GND) hat über 8,6 Mio. Datensätze, die Personen, Körperschaften, Konferenzen, Geografika, Sachbegriffe und Werke beschreiben. Insgesamt werden 50 Entitätentypen (Satzarten) verwendet. Jede Entität besitzt einen eindeutigen Bezeichner (GND-ID), eine bevorzugte Namensform und Merkmale, die sich je nach Entitätentyp unterscheiden, wie z. B. Veranstaltungsort und -datum bei Konferenzen, Erscheinungsort und -datum bei Werken etc. Zusätzlich sind Entitäten untereinander verknüpft und diese Verknüpfungen werden mit Merkmalen detailliert beschrieben. Laut Online-Dienst lobid-gnd[6] existieren in der GND 1 152 856 Einträge, die auch über einen Wikidata-Identifizierer verfügen.

Der Linked-Data-Service der Deutschen Nationalbibliothek (DNB) stellt Datensätze frei unter einer CC0 1.0-Lizenz[7] in den Formaten RDF, Turtle und

6 https://lobid.org/gnd
7 https://creativecommons.org/publicdomain/zero/1.0/deed.de

JSON-LD zur Verfügung.[8] Auf einzelne Datensätze der GND kann man auch über den Katalog der DNB[9] zugreifen und die Datensätze in den Formaten MARC 21 XML und RDF herunterladen.

Abb. 1 stellt einen GND-Datensatz dar, der die Person *Douglas Adams* beschreibt. Der Entitätentyp wird durch *piz* kodiert. Die Entität besitzt einen Geburts- und einen Sterbeort, die ihrerseits über Links mit den korrespondierenden geografischen Entitäten verknüpft sind. Die Entität besitzt auch Merkmale wie Geschlecht, andere Namensformen, Beruf etc.

GND	
Link zu diesem Datensatz	http://d-nb.info/gnd/119033364
Person	Adams, Douglas
Geschlecht	männlich
Andere Namen	Adams, Douglas Noe¨l
Quelle	Bibliogr. Lex. der utop.-phantast. Lit.
Zeit	Lebensdaten: 1952-2001
Land	Großbritannien (XA-GB); USA (XD-US)
Geografischer Bezug	Geburtsort: Cambridge Sterbeort: Santa Barbara, Calif.
Beruf(e)	Science-fiction-autor
Systematik	12.2p Personen zu Literaturgeschichte (Schriftsteller)
Typ	Person (piz)
Autor von	157 Publikationen 1. *Per Anhalter ins All* *Adams, Douglas. - Leipzig : Deutsche Nationalbibliothek, 2020* 2. *[Doctor Who and the Krikkitmen]* *Doctor Who und die Krikkit-Krieger* *Adams, Douglas. - Köln : Bastei Entertainment, 2019, 1. Auflage 2019* 3. ...
Beteiligt an	39 Publikationen 1. *[The restaurant at the end of the universe]* *Das Restaurant am Ende des Universums* *Leipzig : Deutsche Nationalbibliothek, 2020* 2. *Dirk Gently's holistic detective agency* *Leipzig : Deutsche Nationalbibliothek, 2020* 3. ...

Abb. 1: Ein beispielhafter GND-Datensatz (http://d-nb.info/gnd/119033364

8 https://data.dnb.de/opendata/
9 https://portal.dnb.de/

Das Projekt *SoNAR* (s. Abschnitt 4) arbeitet mit einem GND-Datendump im Format MARC 21 XML, der für den Import zu einer Graphdatenbank transformiert wurde.

3.2 Wikidata

Wikidata ist eine sprachunabhängige Wissensbasis für strukturierte Daten, die in Wikipedia, Wikivoyage, Wiktionary, Wikisource usw. Verwendung finden. Die Daten sind nach Objekten (*Items*), Eigenschaften (*Properties*) und Aussagen (*Statements*) strukturiert. Im Jahr 2020 umfasst Wikidata 90 440 252 Objekte und 8 056 Eigenschaften. Wikidata wird in den Formaten JSON, RDF und XML zur Verfügung gestellt.[10] Die zentrale Einheit in Wikidata ist das *Item*. Ein Item-Identifizierer besteht aus dem Präfix Q und einer numerischen ID. *Q42* ist die ID des Items für *Douglas Adams* (siehe Abb. 2). Items umfassen ein oder mehrere sprachspezifische Einträge, in denen ein Name, eine Kurzbeschreibung und Alternativnamen zusammengefasst sind. Das Beispiel *Q42* wird zudem mit weiteren Attribut-Wert-Paaren beschrieben, z. B. besuchte Bildungseinrichtung. Um gleichrangige Aussagen näher zu beschreiben, werden Qualifikatoren eingesetzt (Endzeitpunkt, Hauptfach im Studium, akademischer Grad und Startzeitpunkt). Einige Items aus Wikidata werden über die Property P227 mit der GND verknüpft. Über eine SPARQL-Abfrage[11] kann man die Anzahl verknüpfter GND-IDs ermitteln: 1 182 217. Der Beitrag *Wissensbasen für die automatische Erschließung und ihre Qualität am Beispiel von Wikidata* in diesem Band liefert weitere Details zur globalen Wissensbasis Wikidata.

10 https://www.wikidata.org/wiki/Wikidata:Database_download
11 https://query.wikidata.org

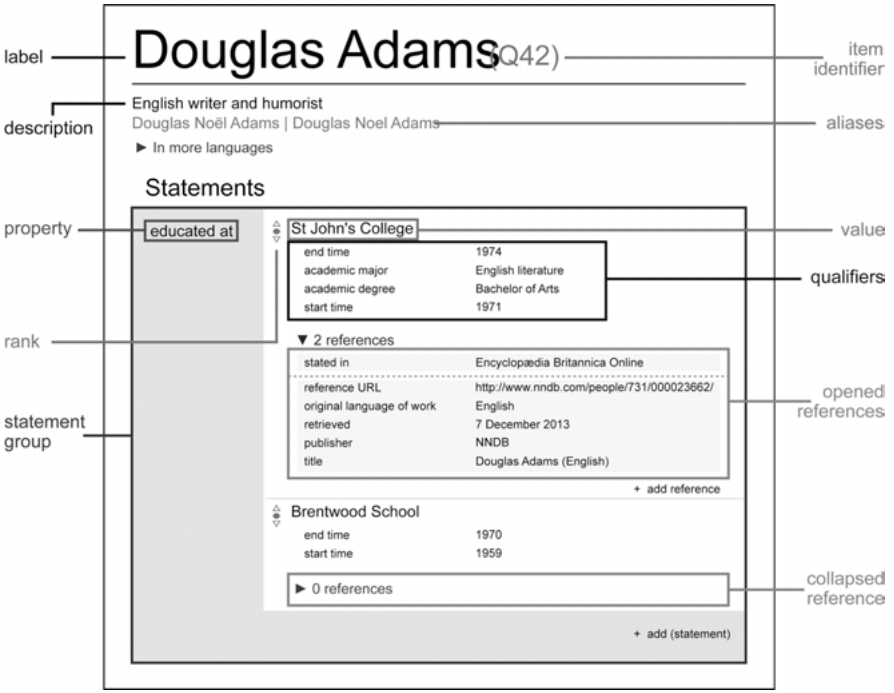

Abb. 2: Ein beispielhafter Wikidata-Datensatz (https://www.mediawiki.org/wiki/Wikibase/DataModel/Primer, 10.11.2020)

4 Named Entity Recognition und Named Entity Linking im Projekt *SoNAR*

Das Projekt *Interfaces to Data for Historical Social Network Analysis and Research* – kurz: *SoNAR* (IDH)[12] – beschäftigt sich damit, vorhandene Daten aus Gedächtnisinstitutionen wie Archiven und Bibliotheken speziell angepasst auf die Bedürfnisse der Historischen Netzwerkanalyse (HNA) aufzubereiten und auf dieser Basis das Konzept einer allgemeinen Forschungsinfrastruktur zu erarbeiten. Die in diesen Daten implizit und explizit enthaltenen Beziehungsinformationen werden dabei modelliert, analysiert und visualisiert (Bludau et al. 2020; Leitner et al. 2020).

12 https://sonar.fh-potsdam.de

Das Projekt nutzt eine Datenbasis von ca. 8 Mio. Normdatensätzen, ca. 22 Mio. Metadatensätzen und ca. 2 Mio. Volltextseiten, u. a. aus Datenbeständen des Kalliope-Verbundes, der Zeitschriftendatenbank und der GND. Die Metadaten beschreiben Ressourcen (Werke, Zeitungen, Akten u. ä.), die Normdaten beschreiben Entitäten sechs verschiedener Klassen (Personen, Geografika, Körperschaften, Konferenzen, Sachbegriffe und Werke). Die Volltexte stellen als direkte Repräsentanten der Ressourcen die Basis der semantischen Module im Projekt dar, zu denen das in diesem Artikel beschriebene Entity Linking gehört.

4.1 Korpora

Das Volltextkorpus besteht aus 2 078 127 digitalisierten Zeitungsseiten aus historischen Beständen der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz (SBB). Die Publikationsorte konzentrieren sich stark auf das Königreich (und später den Freistaat) Preußen, alle Titel sind in Deutschland publiziert worden. Die Digitalisate sind zum überwiegenden Teil öffentlich über das Zeitungsinformationssystem (ZEFYS)[13] zugänglich. Die Zeitungen umfassen Jahrgänge des späten 19. und frühen 20. Jahrhunderts (vgl. Tab. 1). Die Sprache ist – bis auf wenige Ausnahmen auf Artikelebene – fast ausschließlich Deutsch und weist normierte Orthografie auf.[14] Schriften und Seitenanordnung variieren dennoch stark, sowohl zwischen den einzelnen Zeitungstiteln, als auch innerhalb der Titel über die Jahrzehnte. Sowohl Antiqua- als auch Frakturdrucke sind vorhanden.

Tab. 1: Korpus mit Zeitungsvolltexten im Projekt *SoNAR* (IDH)

Titel	Zeitspanne	Anzahl der Dokumente	Anteil in %
Berliner Börsenzeitung	1872–1931	642 480	30,92
Berliner Tageblatt	1877–1939	489 983	23,58
Berliner Volkszeitung	1890–1930	142 403	6,85
Deutsches Nachrichtenbüro	1936–1940	7 429	0,36
Neueste Mittheilungen	1882–1894	1 322	0,06
Norddeutsche Allgemeine Zeitung	1878–1918	120 362	5,79
Provinzial-Correspondenz	1863–1884	1 087	0,05
Teltower Kreisblatt	1856–1896	25 819	1,24
Vossische Zeitung	1857–1917	647 242	31,15

13 http://zefys.staatsbibliothek-berlin.de

14 Typisch für deutschsprachige Periodika nach 1850, vgl. Labusch et al. 2019: 3.

Aus dem Volltextkorpus wurde eine repräsentative Stichprobe zur manuellen Erstellung von Goldstandards gezogen, die im Zuge des Projektes fortlaufend annotiert wird. Der hier untersuchte Datensatz bezieht sich auf diese Stichprobe und wird in Abschnitt 5.2 beschrieben.

4.2 Verarbeitungsschritte

Das Volltextkorpus wurde an der SBB mit einem aktuell im Rahmen des DFG-Projekts OCR-D[15] sowie des BMBF-Projekts *QURATOR*[16] in Entwicklung befindlichen Workflows erstellt. Dieser setzt sich aus zahlreichen einzelnen Verarbeitungsschritten zusammen und soll perspektivisch für alle in der SBB digitalisierten Dokumente zur Anwendung kommen.[17]

4.2.1 Bildvorverarbeitung

Die der Erstellung von Volltexten zugrundeliegenden Bilddigitalisate von historischen Zeitungen wurden aus Kosten- und Effizienzgründen von Mikroverfilmungen der Papieroriginale erstellt. Dies ermöglicht eine hochgradig automatisierte Digitalisierung, bringt aber auch zusätzliche Anforderungen an die Weiterverarbeitung mit sich. So werden z. B. im Zuge der Mikroverfilmung Aufnahmen von Doppelseiten erstellt, die in einem ersten Schritt in Einzelseiten separiert werden müssen (*Cropping*). In zahlreichen Fällen sind die Mikrofilmaufnahmen nicht perfekt ausgerichtet, d. h. die Aufnahmen der Zeitungsseiten sind um einige Grad rotiert. Dies beeinträchtigt die Erkennung und Trennung von Spalten ungemein und kann zu dem unerwünschten Effekt führen, dass mehrere Spalten (und ggf. Artikel) im Zuge der Texterkennung fälschlicherweise zusammengefügt werden. Derlei Defekte sind so zu korrigieren, dass sämtliche Textzeilen nicht von der Horizontalen abweichen (*De-skewing*). Zu guter Letzt ergibt sich durch die Mikroverfilmung häufig ein sehr geringer Kontrast zwischen dem unbedruckten Hintergrund (insbesondere. bei Zeitungspapier) und den gedruckten Inhalten, was die Analyse der Seite und die Texterkennung erheblich negativ beeinflusst. Daher wird noch ein weiterer Schritt ausgeführt, in

15 https://ocr-d.de

16 https://qurator.ai, siehe Rehm et al.

17 Im Rahmen von *QURATOR* werden zudem Konzepte erarbeitet, derartige Verarbeitungsketten mittels eines Workflow-Managers in flexibler Form zu orchestrieren (Moreno-Schneider et al. 2020).

dem sämtliche Pixel in entweder weiß (Hintergrund) oder schwarz (Vordergrund) umgewandelt werden, um den Kontrast deutlich zu erhöhen (*Binarisierung*). Entsprechende Verfahren wurden von der SBB im Projekt *QURATOR* entwickelt und basieren auf neuronalen Netzwerken, die in Keras[18] implementiert und mit denen entsprechende Modelle[19] trainiert wurden.

4.2.2 Segmentierung/Textzeilenerkennung

Im nächsten Schritt werden die vorverarbeiteten Digitalisate einer Segmentierung unterzogen. Hierbei werden diejenigen Bereiche im Bild, die Text enthalten, von denjenigen unterschieden, die z. B. Abbildungen, Tabellen oder Strukturelemente (Separatoren, Verzierungen usw.) darstellen. Anschließend werden die bedruckten Bereiche in einzelne Zeilen aufgetrennt. Erneut kommen Methoden des Deep Learning zum Einsatz[20], die zudem mit Heuristiken ergänzt werden.

4.2.3 Texterkennung

Die einzelnen Zeilen stellen schließlich die benötigte Eingabe für die Erkennung der Zeichen und Texte (OCR) dar. Hierzu kommt die Open-Source-OCR-Software Tesseract[21] zum Einsatz, die seit Version 4 ebenfalls auf Deep Learning beruht. In diesem Fall wurde auf der Grundlage des Datensatzes GT4HistOCR[22] ein Modell[23] (Springmann et al. 2020) trainiert. Als Teil der Texterkennung werden die in einer Textzeile enthaltenen Wörter aufgetrennt, was bei einer späteren Online-Präsentation und Volltextsuche ermöglicht, die entsprechenden Treffer im Digitalisat farblich hervorzuheben.

4.2.4 Transformation von PAGE-XML nach TSV

Für die anschließenden Bearbeitungsschritte muss die im Zuge der Texterkennung erstellte PAGE-XML-Datei[24] in ein *tab-separated-values*-Format (TSV-

18 https://keras.io
19 https://github.com/qurator-spk/sbb_binarization
20 https://github.com/qurator-spk/sbb_textline_detection
21 https://github.com/tesseract-ocr/tesseract
22 https://zenodo.org/record/1344132
23 https://github.com/tesseract-ocr/tesstrain/wiki/GT4HistOCR
24 https://github.com/PRImA-Research-Lab/PAGE-XML

Format) überführt werden, das die Basis sowohl für die automatische Anreicherung als auch für die manuelle Annotation mit Named Entities darstellt.[25] Das verwendete TSV-Format[26] baut auf dem Format des GermEval 2014 *Named Entity Recognition Shared Task*[27] auf, ergänzt dies aber in einigen Bereichen wie z. B. für überlappende Entitäten oder zur Integration von Metadaten für die Annotation.

4.2.5 Tokenisierung

Bevor die Texte mit einer automatisierten NER verarbeitet werden können, müssen sie tokenisiert werden (Textzerlegung in Wörter). Hierfür wird SoMaJo[28] eingesetzt.

4.2.6 Named Entity Recognition (NER)

Für die NER wird eine an der SBB im Projekt *QURATOR* entwickelte Software verwendet, die ein BERT-Modell (Devlin et al. 2019) verwendet. Ziel der Entwicklung war es, ein NER-System zu implementieren, das mit OCR-Fehlern behaftete historische Volltexte in den häufigsten Sprachen der digitalisierten Sammlungen verarbeiten kann. Hierzu haben wir ein multilinguales, von Google auf den 100 umfangreichsten Wikipedien (basierend auf natürlichen Sprachen) vortrainiertes BERT-Modell an die OCR-Volltexte der digitalisierten Sammlungen der SBB, die neben historischen Sprachvarianten auch zahlreiche OCR-Fehler enthalten, adaptiert. Dabei verwendeten wir einen Datensatz bestehend aus rund 2,3 Mio. Seiten[29] der digitalisierten Sammlungen der SBB für ein unüberwachtes Training. Die finale Optimierung für die NER-Aufgabe erfolgte mittels überwachtem Lernen unter Verwendung eines Korpus von 47 000 Token manuell für NER annotierter historischer Daten (Neudecker 2016) sowie unter Verwendung verschiedener zeitgenössischer NER-*Groundtruth*-Korpora (CONLL, GermEval). Weitere Details finden sich bei Labusch et al. (2019) sowie im Readme des veröffentlichten Quellcodes.[30]

25 https://github.com/qurator-spk/page2tsv
26 https://github.com/qurator-spk/neat/blob/master/README.md#22-data-format
27 https://sites.google.com/site/germeval2014ner/data
28 https://github.com/tsproisl/SoMaJo
29 https://zenodo.org/record/3257041
30 https://github.com/qurator-spk/sbb_ner

4.2.7 Named Entity Disambiguation und Linking

Die automatisierte Disambiguierung und Verlinkung der im Zuge der NER erkannten Entitäten erfolgt mittels einer weiteren an der SBB in *QURATOR* entwickelten Software. Diese verwendet ebenfalls ein BERT-Modell. Die zugrundeliegende Wissensbasis besteht aus mehreren Tabellen, in denen alle Personen, Orte und Organisationen aufgelistet sind, die das System kennt. Diese Tabellen entstehen durch rekursives Traversieren geeigneter Kategorien der deutschsprachigen Wikipedia. So ergibt sich die Tabelle der Personen z. B. durch alle Wikipedia-Artikel die den Kategorien *Frau* oder *Mann* angehören. Die Wissensbasis enthält nur diejenigen Wikipedia-Artikel, die auch einer Wikidata-ID zugeordnet werden können. Die Verknüpfung einer erkannten Entität zu einer oder mehreren Wikidata-IDs erfolgt in mehreren Schritten:

I. Zunächst werden mögliche Kandidaten durch eine *nächste-Nachbarn*-Suche (*k-Nearest-Neighbour, kNN*) in einem Embedding-Raum bestimmt, der durch BERT-Embeddings definiert ist. Hierzu wurde ein BERT-Modell auf Satzpaaren der Wikipedia trainiert, so dass es entscheiden kann, ob zwei gegebene Sätze die gleiche Entität enthalten. Dies wird durch die Verlinkungen ermöglicht, die durch die Wikipedia-Editoren erstellt werden.
II. Im nächsten Schritt werden für jeden Kandidaten Satzpaare erstellt, die jeweils aus einem Satz des Zieltextes und einem Satz der Wikipedia bestehen, in dem der jeweilige Kandidat referenziert wird. Das BERT-Modell bestimmt für alle Satzpaare jeweils die Wahrscheinlichkeit, dass in dem Satzpaar die gleiche Entität referenziert wird.
III. Im letzten Schritt wird durch ein Bewertungsmodell eine globale Rangfolge auf Basis aller Satzpaarwahrscheinlichkeiten bestimmt. Hierbei handelt es sich um ein klassisches Modell des maschinellen Lernens (*Random Forest*).

Das System erstellt somit eine Liste von Wikidata-IDs, die absteigend nach Trefferwahrscheinlichkeit sortiert sind. Falls kein Kandidat eine Trefferwahrscheinlichkeit oberhalb von 0.2 hat, ist die Liste leer, siehe Labusch et al. (2020) für weitere Details.

Im Rahmen von *SoNAR* haben wir die automatisiert erkannten Entitäten und ihre Verlinkung stichprobenartig intellektuell überprüft und erweitert. Diese intellektuelle Überprüfung dient als Vergleichsbasis für diese Untersuchung und wird im Folgenden beschrieben.

5 Evaluation der Verlinkung

5.1 Korpus-Stichprobe

Für die Studie wurde aus den in Abschnitt 4 digitalisierten Zeitungen eine zufällige Stichprobe von 10 Zeitungsseiten entnommen, um eine breite Abdeckung unterschiedlicher Zeitungen und Jahrgänge zu erhalten. Die Zeitungsseiten wurden im Zeitraum von 1868 bis 1936 veröffentlicht, wodurch beispielsweise mehrere Staatsformen und Hoheitsgebiete Deutschlands abgedeckt sind. Die Zeitungsseiten in der Stichprobe enthielten zwischen 1 109 und 5 924 Tokens (meist Wörter). Eine der Seiten wurde vollständig intellektuell auf Verlinkungen überprüft, bei den restlichen neun Seiten wurden jeweils die ersten 1 000 Tokens betrachtet. Von den 34 062 Tokens in den 10 Zeitungsseiten wurden 13 824 – also 40,6 % der Tokens – intellektuell überprüft.

Die durch das System automatisch erkannten Entitäten (NER) wurden auf Basis von im Projekt erstellten Annotationsrichtlinien für Entitätentypen[31] intellektuell überprüft und verbessert. Dieser Goldstandard von Zeitungsseiten mit korrekt annotierten Entitäten und ihren jeweiligen Typen wurde anschließend automatisch disambiguiert und mit Wikidata Datensätzen verlinkt. Eine Entität kann dabei aus einem oder mehreren Tokens bestehen. So kann die Entität Kaiser Wilhelm II (Q2677) als einzelne Token wie *Kaiser* oder *Kaisers* oder als die aufeinanderfolgenden *Kaiser* und *Wilhelm* im zugrundeliegenden Text auftauchen. Die Stichprobe enthält 791 Named Entities, deren Verlinkung überprüft wurde.

5.2 Forschungsdesign

Die automatisch verlinkten Entitäten wurden ebenfalls einer manuellen Annotation unterzogen, um einen Goldstandard nicht nur für die Entitätentypen, sondern auch für deren Verlinkung zu erstellen. Hierbei wurde jede tokenisierte Zeitungsseite bzw. die in ihnen enthaltenen Entitäten mit ihren Verlinkungen in mehreren Schritten überprüft und angepasst. Die automatische Entitätenverlinkung resultierte zunächst in einer Datei, in der die schon korrekt erkannten Entitäten entweder verlinkt oder nicht verlinkt waren. Waren Entitäten verlinkt, dann hat das System mehrere Wikidata-Entitäten als Kandidaten in einer nach Relevanz gerankten Reihenfolge vorgeschlagen. Für die Studie wurden nur die

31 https://github.com/qurator-spk/neat/blob/master/Annotation_Guidelines.pdf

erstgerankten Entitätenverlinkungen berücksichtigt. Dies entspricht dem realistischen Szenario, dass im Zeitungskorpus erkannte Entitäten mit derjenigen Entität verlinkt sind, die vom System als die vielversprechendste ausgewählt wurde.

Da die automatische Entitätenverlinkung mit Wikidata als Wissensbasis arbeitete, konnte die Entitätenverlinkung mit der GND nur durch eine manuelle Überprüfung simuliert werden. Folgende Evaluationsschritte wurden durchgeführt:

1. *Überprüfung der Korrektheit der Verlinkung:* Die verlinkte Wikidata-Entität wurde auf ihre Korrektheit gemäß unserer Annotationsrichtlinien überprüft, d. h. es wurde intellektuell evaluiert, ob die verlinkte Entität in Wikidata der erwähnten Entität in der Stichprobe entspricht.
2. *Recherche der falsch oder nicht verlinkten Entitäten in Wikidata:* Für alle fehlerhaft oder gar nicht verlinkten Entitäten in der Stichprobe wurde recherchiert, ob die Entitäten in Wikidata vorhanden sind, d. h. es wurde überprüft, ob der Algorithmus eine korrekte Entität hätte identifizieren können.
3. *Recherche aller Entitäten in der GND:* Alle Entitäten in der Stichprobe wurden ebenfalls in der GND recherchiert, um zu verifizieren, wie viele Entitäten durch eine automatische Entitätenverlinkung identifiziert hätten werden können. War die Entität in Wikidata vorhanden, wurde zunächst überprüft, ob der Wikidata-Eintrag mit der GND verlinkt (Voß et al. 2014) und die im Wikidata-Eintrag vorhandene GND-ID ausgewählt war. War eine GND-ID nicht im Wikidata-Eintrag vorhanden oder die Entität selbst existierte nicht in Wikidata, wurde sie im OGND-Portal des Bibliotheksservicezentrums Baden-Württemberg[32] recherchiert.
4. *Kategorisierung der Verlinkungsfehler bzw. -probleme:* Auftretende Fehler und Probleme in der Verlinkung wurden dokumentiert und kategorisiert. Im Gegensatz zu den von Ling et al. (2015) aufgestellten sechs Fehlertypen (s. Abschnitt 2.2) wurden im Projekt weitere Fehlerkategorien als aussagekräftig identifiziert: (a) Qualität der Wissensbasis, (b) Qualität des Korpus und (c) Fehler in der Disambiguierung.

Tab. 2 beschreibt die potenziellen Fälle in der evaluierten Stichprobe. Sie unterscheidet zwischen der Existenz der Entität in der entsprechenden Wissensbasis (wenn diese grundsätzlich vorhanden ist, könnte diese durch einen Algorithmus gefunden werden) und der Korrektheit der vom Algorithmus identifizierten bzw. verlinkten Entität. Inkorrekt verlinkte Entitäten können im Gegensatz zu nicht verlinkten Entitäten die Qualität der Inhaltserschließung noch stärker

32 http://ognd.bsz-bw.de

schmälern, da sie zu nicht-relevanten Suchergebnissen führen könnten. Die Korrektheit der Verlinkung konnte nur mit Wikidata-Verlinkungen überprüft werden, da der Algorithmus nur auf Wikidata-Daten arbeitete. Fälle A und F zeigen die Menge der korrekt identifizierten Entitäten in Wikidata, Fälle E und J zeigen die Menge der korrekt nicht verlinkten Entitäten in Wikidata (weil die Entität selbst nicht vorhanden war). Dies sind die korrekt vorgenommenen Verlinkungen bzw. Entscheidungen des Verlinkungsalgorithmus. Alle anderen Fälle stellen Fehlerfälle dar. Die Unterscheidung für die GND erlaubt uns die Aussage, ob in der GND andere Entitäten verlinkbar wären als in Wikidata und erlaubt so indirekt eine Aussage über das Abdeckungspotenzial beider Wissensbasen für die Entitätenverlinkung des SoNAR-Zeitungskorpus.

Tab. 2: Fallmatrix zur Evaluation

Entität in	**Wikidata korrekt**	**Wikidata inkorrekt, aber vorhanden**	**Wikidata inkorrekt, nicht vorhanden**	**Wikidata nicht verlinkt, aber vorhanden**	**Wikidata nicht verlinkt, nicht vorhanden**
GND vorhanden	A	B	C	D	E
GND nicht vorhanden	F	G	H	I	J

5.3 Vollständigkeit der Wissensbasen

Insgesamt wurden in den 13 824 Token 791 Entitäten maschinell erkannt. Um diese vollständig und korrekt zu verlinken, müssen diese in der Wissensbasis eine Entsprechung besitzen und die maschinelle Erkennung muss diese identifizieren. Die Entsprechung wurde durch intellektuelle Recherche überprüft und wird nachfolgend genauer dargestellt. Die Korrektheit der Verknüpfungen wird in Abschnitt 5.4 beschrieben.

5.3.1 Abdeckung der Entitäten in GND und Wikidata

Um herauszufinden, wie viele der vom Text benötigten Datensätze in den Wissensbasen vorhanden sind, betrachten wir zuerst die 791 Entitäten, ohne Dopp-

lungen herauszufiltern (Tab. 3). Dazu zählen wir, wie oft eine Verlinkung in Wikidata bzw. zur GND möglich wäre und wie oft nicht.

Tab. 3: Abdeckungsrate der Wissensbasen für die 791 identifizierten Entitäten der Stichprobe

	GND	Wikidata
Datensatz ist nicht vorhanden	252	241
Datensatz ist vorhanden	539	550
Abdeckungsrate (setzbare Links / Entitäten im Text)	68,1 %	69,5 %

In der Stichprobe finden sich viele Dopplungen in den erkannten Entitäten, z. B. wird die Stadt Berlin mehr als einmal erwähnt. Daher wurde überprüft, wie viele eineindeutige Datensätze es bräuchte, um die Entitäten in der vorliegenden Stichprobe korrekt zu verlinken und wie viele davon jeweils in Wikidata und in der GND existieren. Dazu wurden die analysierten Entitäten intellektuell dedupliziert. Wenn mehrere Entitäten intellektuell gleichbehandelt wurden, also mit dem gleichen Link versehen wurden, dann zählen diese als die gleiche Entität. Entitäten, die nicht verlinkt werden konnten, weil kein Datensatz existierte, wurden ebenfalls intellektuell dedupliziert. Insgesamt befinden sich 434 individuelle Entitäten in der Stichprobe. Tab. 4 zeigt die Abdeckung der deduplizierten eindeutigen Entitäten in den Wissensbasen. In der GND wurden durch die intellektuelle Recherche Nachweise für 225 Entitäten gefunden. Hierunter sind 14 Entitäten, die nicht in Wikidata auffindbar sind. Wikidata zeigt mit 235 verlinkbaren Entitäten die etwas bessere Abdeckung der möglichen Nachweise in beiden Normdatenbanken, wobei hier 24 Entitäten verzeichnet sind, die nicht in der GND verfügbar sind.

Tab. 4: Abdeckungsrate der Wissensbasen für die 434 deduplizierten identifizierten Entitäten der Stichprobe

	GND	Wikidata
Datensatz ist nicht vorhanden	209	199
Datensatz ist vorhanden	225	235
Abdeckungsrate (vorhandene Datensätze/benötigte)	51,8 %	54,1 %

Beide Wissensbasen enthalten knapp die Hälfte der in der Stichprobe identifizierten Entitäten. Dies impliziert, dass historische Korpora für eine inhaltliche Erschließung herausfordernder sind, da die Wissensbasen trotz ihres Umfangs nicht alle relevanten Entitäten enthalten. Trotz des ungleich größeren Umfangs

von Wikidata (Abschnitt 5) weisen für den untersuchten Korpus beide Wissensbasen eine ähnliche Abdeckung auf, wobei Wikidata etwas mehr benötigte Nachweise enthält. Tab. 4 enthält eine Übersicht der abgedeckten Entitäten kategorisiert nach Entitätentyp.

Tab. 5: Abdeckungsrate der Wissensbasen für die 434 deduplizierten identifizierten Entitäten der Stichprobe kategorisiert nach Entitätentyp

	GND		Wikidata		nicht vorhanden	Gesamt
Ereignis	2	11,1 %	3	16,7 %	15	18
Lokation*	104	81,3 %	117	91,4 %	10	128
Organisation	60	46,5 %	53	41,1 %	63	129
Person	54	37,2 %	57	39,3 %	88	145
Werk	5	35,7 %	5	35,7 %	9	14

* Eine Lokation kann beispielsweise ein Ort wie *Hannover* oder ein Gebiet wie *Österreich-Ungarn* sein.

Beide Wissensbasen decken vor allem Lokationen gut ab (Wikidata: 91,4 %; GND: 81,3 %). Die GND enthält mehr Datensätze für benannte Organisationen in der Stichprobe. Für Ereignisse sind beide Wissensbasen nur schlecht geeignet. Das liegt aber auch an den erkannten Ereignis-Entitäten, unter denen beispielsweise auch Veranstaltungen wie *die am 15. d. M. stattgefundene Sitzung des Aufsichtsrathes der Aktien-Zuckerfabrik Bauerwitz* zu finden sind. Die nur mittelmäßige Abdeckung bei Organisationen und Personen lässt sich durch deren relative geschichtliche Unbedeutsamkeit erklären. Auch wenn zum Zeitpunkt des Erscheinens einer Zeitung die Organisation *Gebrüder Scheller* oder eine Person namens *Conradi* für das Zeitgeschehen wichtig waren, so sind sie dies (zum jetzigen Zeitpunkt) nicht mehr und würden nur in Wissensbasen mit historischem Fokus aufgenommen werden.

5.3.2 Überlappung zwischen GND und Wikidata

Wikidata und die GND decken für die betrachtete Stichprobe mehrheitlich die gleichen Entitäten ab, können also nicht als komplementäre Wissensbasen betrachtet werden. Die GND enthält 13 Organisationen und eine Lokation mehr, die nicht in der Wikidata zu finden sind. Wikidata hingegen enthält Nachweise für ein Ereignis, 14 Lokationen, 6 Personen und 3 Organisationen, die nicht in der GND verzeichnet sind (Tab. 6).

Tab. 6: Überlappung der Datensätze in beiden Wissensbasen sowie die jeweils nur in einer der Basen existierenden Datensätze nach Eventtyp

	in beiden Wissensbasen verzeichnet		nur GND	nur Wikidata
Ereignis	2	11,1 %	0	1
Lokation	103	80,5 %	1	14
Organisation	47	36,4 %	13	6
Person	54	37,2 %	0	3
Werk	5	35,7 %	0	0
Gesamt	211	48,6 %	14	24

Die Datensätze in Wikidata sind mehrheitlich mit den passenden Datensätzen in der GND verknüpft, welches aufgrund der entsprechenden Initiativen nicht überrascht (Ohlig 2018). So enthalten, basierend auf einer groben Auswertung, 80–90 % der maschinell korrekt verlinkten Wikidata-Einträge ihrerseits einen Verweis auf den entsprechenden GND-Eintrag.

5.4 Automatische Verlinkung zu Wikidata

Um die Qualität der automatischen Verlinkung zu bestimmen, wurden die erkannten Entitäten einzeln überprüft. Jedes Mal, wenn eine als Entität erkannte Zeichenfolge auftritt, wird maschinell entschieden, ob es einen passenden Eintrag in der Wissensbasis gibt und welcher der möglichen Einträge am besten passt. Dabei wird der textliche Kontext berücksichtigt, weswegen die gleiche Zeichenfolge an unterschiedlichen Positionen im Text unterschiedlich verlinkt werden könnte. Die maschinelle Verlinkung fand bis zu 46 Wikidata-Einträge pro Entität. Der Durchschnitt maschinell gefundener Verlinkungen pro Entität in der Stichprobe lag bei 3,72. Für 213 Entitäten (von 791) wurde genau ein Eintrag in der Wissensbasis gefunden. Intellektuell wurde nur jeweils die erste maschinell gefundene Verlinkung überprüft und gegebenenfalls verbessert.

In Tab. 7 sind die möglichen Kategorien aufgelistet, wie die automatische Verlinkung intellektuell bewertet wurde. Kategorie A beinhaltet all jene maschinell gefundenen Links, die intellektuell als korrekt bewertet wurden. Kategorie B beinhaltet jene, bei denen maschinell ein falscher Link gesetzt wurde, ein anderer Link jedoch korrekt gewesen wäre. Kategorie C wiederum beinhaltet die Fälle, in denen ein Link gesetzt wurde, obwohl es nach der intellektuellen Überprüfung keinen passenden Eintrag in der Wissensbasis gibt. In Kategorie D sind alle Vorkommen gesammelt, für die ein Eintrag in der Wissensbasis existieren

würde, der aber nicht automatisch gefunden wurde. Und in Kategorie E wurden diejenigen Entitäten aufgeführt, für die maschinell kein Link gefunden wurde und für die es auch tatsächlich keinen Eintrag in der Wissensbasis gibt.

Die *precision* (Korrektheit) ist der Anteil der maschinell korrekt verlinkten Entitäten aus allen, die maschinell verlinkt wurden. Der *recall* (Abdeckung) ist der Anteil der maschinell korrekt verlinkten Entitäten aus allen Entitäten, die basierend auf ihrem Vorkommen in Wikidata korrekt verlinkt sein könnten. Für die gesamte Stichprobe lag die *precision* bei 37,9 % und der *recall* bei 38,5 %. Dies bedeutet, dass 37,9 % der automatisch verlinkten Entitäten korrekt verlinkt wurden und dass 38,5 % der potenziell verlinkbaren Entitäten korrekt verlinkt wurden.

Die Kategorisierung nach Entitätentyp zeigt ein differenziertes Bild (Tab. 7). Da die automatische Erkennung Ereignisse und Werke nicht verlinkte, konnte für diese auch keine *precision* berechnet werden und der *recall* liegt bei 0. Auch hier zeigt sich die oben beschriebene besonders hohe Abdeckung der Wissensbasis für Orte, während Personen und Organisationen weniger effektiv behandelt werden. So liegen für den Entitätentyp *Ort* die *precision* mit 55,5 % und der *recall* mit 46,0 % weit über den Werten für alle anderen Entitäten. Wenn man die automatische Verlinkung ohne intellektuelle Evaluation für die inhaltliche Erschließung einsetzen wollte, ist insbesondere zu beachten, dass mehr automatische Links falsch gesetzt wurden (Kategorien B und C) als korrekt. Die über 60 % automatischen inkorrekten Verlinkungen werden daher die Erschließung ineffektiver machen. Die Evaluation konnte allerdings nicht untersuchen, ob in den vorgeschlagenen Verlinkungen (hier wurde nur der erstgerankte Vorschlag untersucht) auch die korrekte Entität verlinkt wurde, welches zu einem Mehrwert führen könnte, falls man z. B. einem Suchenden alle Möglichkeiten zur Auswahl vorlegen könnte.

Tab. 7: Anzahl der Entitäten nach Typ in jeder Fehlerkategorie sowie die daraus resultierenden precision-, recall- und F1-Werte

	A korrekt verlinkt	**B falscher Link**	**C fälschlicherweise Link**	**D fälschlicherweise kein Link**	**E richtigerweise kein link**	**Summe der Fälle A-E**	**precision A/(A+B+C)**	**recall A/(A+B+D)**	**F1**
gesamt	212	202	146	136	95	791	37,9 %	38,5 %	0,382
Ereignis	0	0	0	4	19	23	-	0	-
Lokation	157	120	6	64	6	353	55,5 %	46,0 %	0,503

	A korrekt verlinkt	B falscher Link	C fälschlicherweise Link	D fälschlicherweise kein Link	E richtigerweise kein link	Summe der Fälle A-E	precision A/(A+B+C)	recall A/(A+B+D)	F1
Organisation	22	44	82	33	18	199	14,9 %	22,2 %	0,178
Person	33	38	55	28	40	194	26,2 %	33,3 %	0,293
Werk	0	0	0	7	11	18	-	0	-

5.5 Herausforderungen und Fehlerquellen

Die manuelle Evaluation der Verlinkung ermöglichte ebenfalls eine Kategorisierung von unterschiedlichen Fehlertypen, die bei der automatischen Entitätenverlinkung auftraten. Dabei wurden neben den Herausforderungen in der Disambiguierung von Entitäten, die die Mehrheit der Fehlertypen bei Ling et al. (2015) ausmachen und die auf den Verlinkungsalgorithmus zurückzuführen sind, auch die Qualität der Wissensbasis und des zu verlinkenden Textkorpus als signifikante Fehlerquellen identifiziert. Sonstige Fehler tauchen in zu vernachlässigenden Mengen auf, so dass sie hier nicht weiter beschrieben werden.

5.5.1 Qualität der Wissensbasis

Wenn Entitäten in einer Wissensbasis inkorrekt oder unvollständig angelegt oder beschrieben sind, entstehen auch Fehler in der Entitätenverlinkung, da die Entitäten schwerer zu identifizieren sind. Dies ist nicht einer Schwachstelle im Algorithmus zuzuschreiben, sondern eher den Qualitätsproblemen, die in der Wissensbasis an sich auftreten. Beispiele dafür sind:

- Entitäten sind in der Wissensbasis nicht enthalten (z. B. Fokus auf Autor:innen in der GND, während Zeitungstexte oft Personen des öffentlichen Lebens referenzieren).
- Für dieselbe Entität existieren mehrere Einträge in der Wissensbasis (z. B. *Q60685764* und *Q311124* für *Lüneburger Heide* in Wikidata).
- Für fast äquivalente Entitäten existieren mehrere Datensätze in der Wissensbasis (z. B. existieren für *Reichstag* eine Vielzahl an Einträgen, s. *Q20007287* und *Q29053853*, wobei es zu Überschneidungen kommt).

- Datensätze sind unvollständig, enthalten aber Namen von Entitäten, die ein Algorithmus findet (z. B. *Q72391398*).
- In Einträgen fehlen Varianten von Entitätennamen (z. B. finden sich unterschiedliche Schreibweisen in Zeitungstexten nicht in der Wissensbasis wieder).

Ein Sonderfall für historische Texte entsteht durch das Problem, dass Entitäten, die über einen langen Zeitraum existieren, verschiedene Ausprägungen annehmen, die in den Wissensbasen nicht vollständig nachvollzogen bzw. korrekt voneinander getrennt werden. Länder ändern bspw. ihre Grenzen, Behörden, Gremien und Organisationen ihre Namen und Befugnisse. Eine Herausforderung für die Entitätenverlinkung ist die Frage, ob die aktuelle Version eines Orts oder einer Organisation verlinkt werden sollte, insbesondere wenn verschiedene historische Ausprägungen der gleichen Entität in der Wissensbasis existieren. Für den Reichstag enthält Wikidata z. B. Datensätze für die verschiedenen historischen Ausprägungen: *Q321246* für 1867–1870, *Q160208* für 1871–1918, *Q637829* für 1919–1945 und *Q878525* für 1933–1945. Für Preußen existieren dagegen überlappende Datensätze. So umschließt z. B. *Preußen* (*Q38872*) (1525–1947) die Wikidata-Entität *Königreich Preußen* (*Q27306*) (1701–1918) ebenso wie das *Herzogtum Preußen* (*Q153091*) (1525–1618) oder den *Freistaat Preußen* (*Q161036*) (1918–1947) zeitlich. Welche dieser Entitäten sollte der Algorithmus verlinken, die umfassendste, die zeitlich am nächsten stehende oder nur die vollständig übereinstimmende Version?

5.5.2 Qualität des Korpus

Da die Verlinkung anhand einer Stichprobe von manuell korrigierten Gold-Standards der Entitätenauszeichnung (NER) durchgeführt wurden, war zwar eine geringere Fehlerrate im Korpus zu erwarten, diese ist aber nicht null, welches auch die Herausforderungen einer manuellen Annotation (NER) deutlich macht. Fehler im Korpus führen zu Folgefehlern in der automatischen Verlinkung. Beispiele für Korpusfehler sind:

- Inkorrekte Entitätenerkennung (Fehlertyp 2 in Ling et al. 2015) (z. B. inkorrekte Annotation des Tokens *Krone* als Organisation, welche in der Verlinkung nicht identifiziert wird)
- OCR-Fehler (z. B. erschwert die inkorrekte Erkennung von J und I in Fraktur-Texten die automatische Verlinkung)

5.5.3 Fehler in der Disambiguierung

Disambiguierungsfehler entstehen immer dann, wenn Entitäten semantisch schwer voneinander unterscheidbar sind. Beispiele für Disambiguierungsfehler sind:

- Koreferenzfehler durch Homonymie (Fehlertyp 3 in Ling et al. 2015) (z. B. Straßen mit identischen Namen in der Wissensbasis, die vom Verlinkungsalgorithmus auf die gleiche Entität verlinkt werden)
- Kontextfehler durch Homonymie (Fehlertyp 4 in Ling et al. 2015) (z. B. inkorrekte Verlinkung des hessischen Großherzogs *Q57507* als Badischer Großherzog *Q57483*)
- Spezifität (Fehlertyp 5 in Ling et al. 2015) (z. B. Verlinkung des Bezirks *Friedrichshain-Kreuzberg*, wenn im Text die Entität *Berlin* benannt ist)
- Wortformen (z. B. fehlende Verlinkung von *Französisch* zur Entität *Frankreich*)

6 Zusammenfassung

Die Studie zeigt, dass einerseits ein Potenzial für die inhaltliche Erschließung von Named Entities in historischen Texten vorhanden ist, denn immerhin sind über die Hälfte der in der Stichprobe genannten Entitäten auch vorhanden. Andererseits zeigt sie gleichermaßen, dass sowohl die Abdeckung als auch die Korrektheit der Verlinkung durchaus noch verbessert werden kann. Eine Ausnahme bilden Orte, die sowohl für die Abdeckung in beiden Wissensbasen als auch in der Korrektheit der Verlinkung signifikant bessere Ergebnisse zeigen als die anderen Entitätentypen. Ein überraschendes Ergebnis dieser Untersuchung war, dass sich die Wissensbasen GND und Wikidata, obwohl sehr verschieden in der Größe, in der Abdeckung der historischen Entitäten im untersuchten Korpus gleichen. Ein Grund dafür kann sein, dass die umfangreiche deutschsprachige Wikipedia auch in Wikidata repräsentiert ist und damit deutsche Entitäten – wie sie aus der Stichprobe deutscher Zeitungen zu erwarten sind – auch überdurchschnittlich oft in der globalen Wissensbasis vorkommen. Die GND, die für Personen ursprünglich auf Urheber:innen von publizierten Texten fokussiert war, steht der Wikidata mit einem allgemeinen Fokus allerdings auch für Personen für die untersuchte Stichprobe nicht hinterher.

Was die vorliegende Studie nicht vermochte, war die automatische Verlinkung mit beiden Wissensbasen zu vergleichen. Die durchschnittliche *precision*

für die automatische Verlinkung mit Wikidata lag bei unter 40 %, was u. a. auch an Qualitätsproblemen der Wissensbasis, also der Erfassung von Informationen in Wikidata, erklärt werden konnte (s. Abschnitt 5.5). Eine interessante, zu überprüfende Theorie ist, dass einige dieser Fehlerquellen in der intellektuell kuratierten GND nicht auftauchen würden und damit die automatische Verlinkung verbessert werden könnte. Andererseits enthalten Wikidata-Datensätze oftmals mehr und andere Informationen, die wiederum einem Algorithmus bei der korrekten Verlinkung durch ihren Kontext unterstützen könnten.

Ein anderer Aspekt, der die Qualität der inhaltlichen Erschließung durch Verlinkungen mit Wissensbasen beeinflusst, ist sowohl die Beschreibungstiefe und -qualität als auch das interne Beziehungsnetzwerk von Datensätzen innerhalb der Wissensbasis. Wenn man von einer Entität in einem digitalisierten Text auf die entsprechende Beschreibung der Entität in einer Wissensbasis weitergeleitet wird, wie viele Informationen enthält der entsprechende Datensatz in der Wissensbasis und mit wie vielen anderen Datensätzen ist er verlinkt? Auch diese Art der Untersuchung wartet auf zukünftige Analysen.

7 Danksagung

Die Erstellung dieses Beitrags wurde unterstützt mit Mitteln des von der Deutschen Forschungsgemeinschaft (DFG) geförderten Projekts *SoNAR (IDH)* (Projektnr. 414792379) sowie des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projekts *QURATOR* (Unternehmen Region, Wachstumskern, Projektnr. 03WKDA1A).

8 Literaturverzeichnis

Balog, Krisztian: Entity Linking. In: Entity-Oriented Search. Hrsg. v. Krisztian Balog. Cham: Springer International Publishing 2018. S. 147–188. https://doi.org/10.1007/978-3-319-93935-3_5.

Balzer, Detlev, Barbara K. Fischer, Jürgen Kett, Susanne Laux, Jens M. Lill, Jutta Lindenthal, Mathias Manecke, Martha Rosenkötter und Axel Vitzthum: Das Projekt „GND für Kulturdaten“; (GND4C). In: o-bib. Das Offene Bibliotheksjournal (2019) Bd. 6 Nr. 4. S. 59–97. https://doi.org/10.5282/o-bib/2019H4S59-97.

Batista, David S.: Named-Entity evaluation metrics based on entity-level. 2018. http://www.davidsbatista.net/blog/2018/05/09/Named_Entity_Evaluation/.

Berners-Lee, Tim: Linked Data. W3C 2006. https://www.w3.org/DesignIssues/LinkedData.html.

Bludau, Mark-Jan, Marian Dörk, Heiner Fangerau, Thorsten Halling, Elena Leitner, Sina Menzel, Gerhard Müller, Vivien Petras, Georg Rehm, Clemens Neudecker, David Zellhöfer und Julián Moreno Schneider: SoNAR (IDH): Datenschnittstellen für historische Netzwerkanalyse. In: DHd 2020 Spielräume: Digital Humanities zwischen Modellierung und Interpretation. Konferenzabstracts. Tagung des Verbands Digital Humanities, March 2–6, Paderborn, Germany. Verband Digital Humanities im deutschsprachigen Raum e. V. Hrsg. v. Christof Schöch. 2020. S. 360–362. https://doi.org/10.5281/zenodo.3666690.

Bruce, Thomas und Diane I. Hillmann: Metadata Quality in a Linked Data Context. In: VoxPopuLII. New voices in legal information. 2013. https://blog.law.cornell.edu/voxpop/2013/01/24/metadata-quality-in-a-linked-data-context/.

Bunescu, Razvan und Marius Pasca: Using Encyclopedic Knowledge for Named entity Disambiguation. In: 11th Conference of the European Chapter of the Association for Computational Linguistics. Hrsg. v. Diana McCarthy und Shuly Wintner. Association for Computational Linguistics 2006. S. 9–16. https://www.aclweb.org/anthology/E06-1002.

Cucerzan, Silviu: Large-scale named entity disambiguation based on Wikipedia data. In: Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL). Hrsg. v. Jason Eisner. Association for Computational Linguistics 2007. S. 708–716. https://www.aclweb.org/anthology/D07-1074.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee und Kristina Toutanova: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics 2019. S. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423.

Ehrmann, Maud, Matteo Romanello, Alex Flückiger und Simon Clematide: Extended overview of CLEF HIPE 2020: named entity processing on historical newspapers. In: CLEF 2020 Working Notes. Conference and Labs of the Evaluation Forum. CEUR-WS 2020. http://dx.doi.org/10.5281/zenodo.4117566.

Ganea, Octavian-Eugen und Thomas Hofmann: Deep joint entity disambiguation with local neural attention. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Hrsg. v. Martha Palmer, Rebecca Hwa und Sebastian Riedel. Association for Computational Linguistics 2017. S. 2619–2629. https://www.aclweb.org/anthology/D17-1277.

Hachey, Ben, Will Radford, Joel Nothman, Matthew Honnibal und James R. Curran: Evaluating Entity Linking with Wikipedia. In: Artificial Intelligence (2013) Bd. 194. S. 130–150. https://doi.org/10.1016/j.artint.2012.04.005.

Hochstein, Juliane: „Ihr Bibliothekare habt doch jetzt …“. Ein Jahr „Gemeinsame Normdatei“. In: Theke aktuell 2013 Bd. 20 Nr. 1. S. 19–23. https://journals.ub.uni-heidelberg.de/index.php/ThekeAkt/article/download/11337/5198.

Kolitsas, Nikolas, Octavian-Eugen Ganea und Thomas Hofmann: End-to-end neural entity linking. In: Proceedings of the 22nd Conference on Computational Natural Language Learning. 2018. S. 519–529. https://arxiv.org/abs/1808.07699.

Labusch, Kai, Clemens Neudecker und David Zellhöfer: BERT for Named Entity Recognition in Contemporary and Historic German. In: Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019). 2019. https://konvens.org/proceedings/2019/papers/KONVENS2019_paper_4.pdf.

Labusch, Kai und Clemens Neudecker: Named Entity Disambiguation and Linking Historic Newspaper OCR with BERT. In: CLEF 2020 Working Notes. Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum. Hrsg. v. Linda Capellato, Carsten Eikhoff, Nicola Ferro und Aurélie Névéoll. CEUR-WS 2020. http://ceur-ws.org/Vol-2696/paper_163.pdf.

Leitner, Elena, Julián Moreno-Schneider, Georg Rehm, Sina Menzel, Vivien Petras, Mark-Jan Bludau und Marian Dörk: Graphtechnologien für die Analyse historischer Netzwerke mit heterogenen Datenbeständen. In: Proceedings of Graph Technologies in the Humanities 2020, February 21–22, Vienna, Austria. Hrsg. v. T. Andrews, F. Diehr, T. Efer, A. Kuczera und J. v. Zundert. Verband Digital Humanities im deutschsprachigen Raum e. V 2020.

Ling, Xiao, Sameer Singh und Daniel S. Weld: Design Challenges for Entity Linking. In: Transactions of the Association for Computational Linguistics (2015) Bd. 3. S. 315–328. https://doi.org/10.1162/tacl_a_00141.

Linhares Pontes, Elvys, Ahmed Hamdi, Nicolas Sidere und Antoine Doucet: Impact of OCR Quality on Named Entity Linking. In: Digital Libraries at the Crossroads of Digital Information for the Future. ICADL 2019. Lecture Notes in Computer Science, Bd, 11853. Hrsg. v. Adam Jatowt, Akira Maeda und Sue Yeon Syn. Cham: Springer 2019. S. 102–115. https://doi.org/10.1007/978-3-030-34058-2_11.

Moreno-Schneider, Julián, Peter Bourgonje, Florian Kintzel und Georg Rehm: A Workflow Manager for Complex NLP and Content Curation Pipelines. In: Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020, co-located with LREC 2020), Marseille, France, 2020. Hrsg. v. Georg Rehm, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis und Andrejs Vasiljevs. 2020. S. 73–80. https://arxiv.org/abs/2004.14130.

Neudecker, Clemens: An open corpus for named entity recognition in historic newspapers. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). European Language Resources Association 2016. S. 4348–4352. https://www.aclweb.org/anthology/L16-1689.

Neudecker, Clemens, Konstantin Baierer, Maria Federbusch, Kay-Michael Würzner, Matthias Boenig, Elisa Herrmann und Volker Hartmann: OCR-D: An end-to-end open-source OCR framework for historical documents. In: Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage (DATeCH2019), Brüssel 09.05.2019. New York, NY: Association for Computing Machinery 2019. S. 53–58. https://doi.org/10.1145/3322905.3322917.

Ohlig, Jens: Gemeinsam wieder Neuland betreten: Die Deutsche Nationalbibliothek und Wikimedia Deutschland. Wikimedia Blog 2018. https://blog.wikimedia.de/2018/11/02/gemeinsam-wieder-neuland-betreten-die-deutsche-nationalbibliothek-und-wikimedia-deutschland/.

Piscopo, Alessandro und Elena Simperl: What we talk about when we talk about Wikidata quality: a literature survey. In: Proceedings of the 15th International Symposium on Open Collaboration (OpenSym '19). New York, NY: Association for Computing Machinery 2019. S. 1–11. https://doi.org/10.1145/3306446.3340822.

Provatorova, Vera, Svitlana Vakulenko, Evangelos Kanoulas, Koen Dercksen und Johannes M. van Hulst: Named Entity Recognition and Linking on Historical Newspapers: UvA.ILPS & REL at CLEF HIPE 2020. In: CLEF 2020 Working Notes. Working Notes of CLEF 2020 – Conference and Labs of the Evaluation Forum. Hrsg. v. Linda Capellato, Carsten Eikhoff, Nicola Ferro und Aurélie Névéoll. CEUR-WS 2020. http://ceur-ws.org/Vol-2696/paper_209.pdf.

Rao, Delip, Paul McNamee und Mark Dredze: Entity linking: Finding extracted entities in a knowledge base. In: Multi-source, multilingual information extraction and summarization. Hrsg. v. Thierry Poibeau, Horacio Saggion, Jakub Piskorski und Roman Yangarber. Berlin, Heidelberg: Springer 2013. S. 93–115. https://doi.org/10.1007/978-3-642-28569-1_5.

Rehm, Georg, Peter Bourgonje, Stefanie Hegele, Florian Kintzel, Julián Moreno Schneider, Malte Ostendorff, Karolina Zaczynska, Armin Berger, Stefan Grill, Sören Räuchle, Jens Rauenbusch, Lisa Rutenburg, André Schmidt, Mikka Wild, Henry Hoffmann, Julian Fink, Sarah Schulz, Jurica Seva, Joachim Quantz, Joachim Böttger, Josefine Matthey, Rolf Fricke, Jan Thomsen, Adrian Paschke, Jamal Al Qundus, Thomas Hoppe, Naouel Karam, Frauke Weichhardt, Christian Fillies, Clemens Neudecker, Mike Gerber, Kai Labusch, Vahid Rezanezhad, Robin Schaefer, David Zellhöfer, Daniel Siewert, Patrick Bunk, Lydia Pintscher, Elena Aleynikova und Franziska Heine: QURATOR: Innovative Technologies for Content and Data Curation. In: Qurator 2020 – Conference on Digital Curation Technologies. Proceedings of the Conference on Digital Curation Technologies, Berlin, 2020. Hrsg. v. Adrian Paschke, Clemens Neudecker, Georg Rehm, Jamal Al Qundus und Lydia Pintscher. CEUR Workshop Proceedings. Bd. 2535. 2020. http://ceur-ws.org/Vol-2535/paper_17.pdf

Sevgili, Ozge, Artem Shelmanov, Mihhail Arkhipov, Alexander Panchenko und Chris Biemann: Neural Entity Linking: A Survey of Models based on Deep Learning. 2020. https://arxiv.org/abs/2006.00575.

Shen, Wei, Jianyong Wang und Jiawei Han: Entity linking with a knowledge base: Issues, techniques, and solutions. In: IEEE Transactions on Knowledge and Data Engineering (2014) Bd. 27 H. 2. S. 443–460. https://doi.org/10.1109/TKDE.2014.2327028.

Springmann, Uwe, Christian Reul, Stefanie Dipper und Johannes Baiter: Ground Truth for training OCR engines on historical documents in German Fraktur and Early Modern Latin. Journal for Language Technology and Computational Linguistics (2018) Bd. 33 H. 1. S. 97–114. https://jlcl.org/content/2-allissues/2-heft1-2018/jlcl_2018-1_5.pdf.

Tamper, Minna, Eero Hyvönen und Petri Leskinen: Visualizing and Analyzing Networks of Named Entities in Biographical Dictionaries for Digital Humanities Research. In: Proceedings of the 20th International Conference on Computational Linguistics and Intelligent Text Processing (CICling 2019). Germany: Springer-Verlag 2019. Preprint: https://easychair.org/publications/preprint_download/7xBL.

Varma, Vasudeva, Vijay Bharat, Sudheer Kovelamudi, Praveen Bysani, Santosh GSK, Kiran Kumar N, Kranthi Reddy, Karuna Kumar und Nitin Maganti: IIIT Hyderabad at TAC 2009. https://www.researchgate.net/publication/242545144_IIIT_Hyderabad_at_TAC_2009.

Voß, Jakob, Susanna Bausch, Julian Schmitt, Jasmin Bogner, Viktoria Berkelmann, Franziska Ludemann, Oliver Löffel, Janna Kitroschat, Maiia Bartoshevska und Katharina Seljuzki: Normdaten in Wikidata – Handbuch. Version 1.0 (2014). https://hshdb.github.io/normdaten-in-wikidata/.

Yamada, Ikuya, Koki Washio, Hiroyuki Shindo, und Yuji Matsumoto: Global Entity Disambiguation with Pretrained Contextualized Embeddings of Words and Entities. 2019. https://arxiv.org/abs/1909.00426.

Elektronische Quellen zuletzt geprüft am 08.06.2021.