Lydia Pintscher, Peter Bourgonje, Julián Moreno Schneider,
Malte Ostendorff, Georg Rehm

Wissensbasen für die automatische Erschließung und ihre Qualität am Beispiel von Wikidata

1 Einführung

Wikidata[1] ist eine freie Wissensbasis, die allgemeine Daten über die Welt zur Verfügung stellt. Sie wird von Wikimedia entwickelt und betrieben, wie auch das Schwesterprojekt Wikipedia. Die Daten in Wikidata werden von einer großen Community von Freiwilligen gesammelt und gepflegt, wobei die Daten sowie die zugrundeliegende Ontologie von vielen Projekten, Institutionen und Firmen als Basis für Applikationen und Visualisierungen, aber auch für das Training von maschinellen Lernverfahren genutzt werden. Wikidata nutzt MediaWiki[2] und die Erweiterung Wikibase[3] als technische Grundlage der kollaborativen Arbeit an einer Wissensbasis, die verlinkte offene Daten für Menschen und Maschinen zugänglich macht.

Ende 2020 beschreibt Wikidata über 90 Millionen Entitäten (siehe Abb. 1) unter Verwendung von über 8 000 Eigenschaften, womit insgesamt mehr als 1,15 Milliarden Aussagen über die beschriebenen Entitäten getroffen werden. Die Datenobjekte dieser Entitäten sind mit äquivalenten Einträgen in mehr als 5 500 externen Datenbanken, Katalogen und Webseiten verknüpft, was Wikidata zu einem der zentralen Knotenpunkte des Linked Data Web macht. Mehr als 11 500 aktiv Editierende[4] (siehe Abb. 2) tragen neue Daten in die Wissensbasis ein und pflegen sie. Diese sind in Wiki-Projekten organisiert, die jeweils bestimmte Themenbereiche oder Aufgabengebiete adressieren. Die Daten werden in mehr als der Hälfte der Inhaltsseiten in den Wikimedia-Projekten genutzt und unter anderem mehr als 6,5 Millionen Mal am Tag über den SPARQL-Endpoint[5] abgefragt, um sie in externe Applikationen und Visualisierungen einzubinden.

1 https://www.wikidata.org (17.12.2020).
2 https://www.mediawiki.org (17.12.2020).
3 https://wikiba.se (17.12.2020).
4 Aktiv Editierende sind Editierende, die in den letzten 30 Tagen fünf oder mehr Änderungen vorgenommen haben.
5 https://query.wikidata.org (17.12.2020).

https://doi.org/10.1515/9783110691597-005

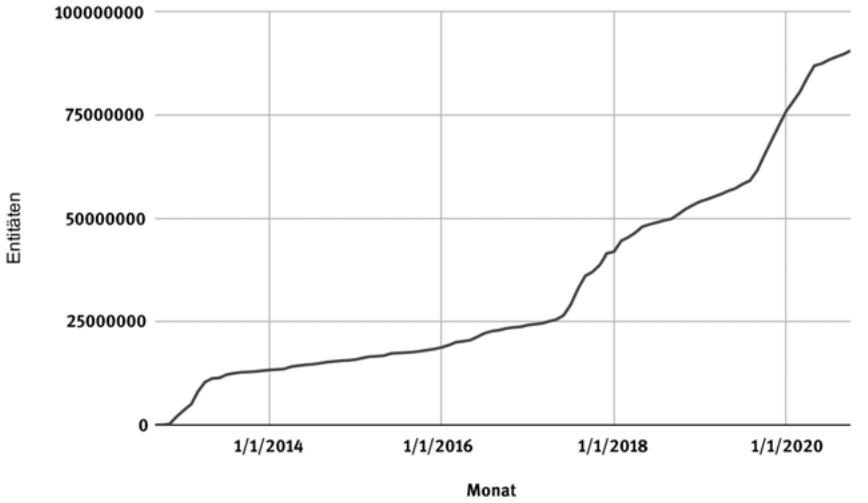

Abb. 1: Entwicklung der Anzahl der in Wikidata beschriebenen Entitäten

Abb. 2: Entwicklung der Anzahl der aktiv Editierenden in Wikidata

Wikidata wurde 2012 ins Leben gerufen. Das primäre Ziel war, die Wikipedia durch die zentrale Verwaltung der Links zu unterschiedlichen Sprachversionen

der Wikipedia[6] und durch die Pflege allgemeiner Daten an einem zentralen Ort zu verbessern (Vrandečić und Krötzsch 2014). Bis zu diesem Zeitpunkt pflegte jede Sprachversion der Wikipedia ihre Daten eigenständig und unabhängig. Dies führte zu zahlreichen Inkonsistenzen und Benachteiligungen kleinerer Wikipedien. Wenig später wurden auch die anderen Wikimedia-Projekte[7] unterstützt. Gleichzeitig wuchs das Interesse an Wikidata auch außerhalb der Wikimedia-Projekte, und die Daten wurden von Projekten, Institutionen und Firmen – von kleinen Betrieben bis hin zu großen Technologie-Unternehmen – in einer Vielzahl von Anwendungsfällen genutzt. Durch diese Erweiterung der Zielgruppen von Wikidata erweiterte sich auch dessen Inhalt. Anfangs war er sehr stark am enzyklopädischen Inhalt von Wikipedia ausgerichtet, wurde dann durch die Inhalte der anderen Wikimedia-Projekte erweitert (Daten zu Mediendateien sowie Sprachdaten) und später erneut ausgebaut, geprägt durch die Bedürfnisse von externen Nutzenden der Daten nach einer noch breiteren Abdeckung. Aber nicht alle Inhalte können in Wikidata gespeichert werden. Dies wäre durch die Wikidata-Community nicht beherrschbar und auch technisch nicht wünschenswert. Als Lösung wurde Wikibase, die technische Basis von Wikidata, auch für Dritte nutzbar gemacht. Die Vision dabei ist, dass ein mit Wikidata eng verknüpftes Netzwerk von Wikibase-Installationen entsteht, die jeweils spezialisierte Daten vorhalten. Innerhalb dieses als *Wikibase Ecosystem* bezeichneten Netzwerks können Daten untereinander leicht verlinkt und ausgetauscht werden.

1.1 Funktionsweise von Wikidata

Eine Gemeinschaft von Freiwilligen sammelt in und für Wikidata Daten, die strukturiert, verknüpft, multilingual und maschinenlesbar sind, wobei Wikidata eng mit den anderen Wikimedia-Schwesterprojekten[8] verknüpft ist. Viele Daten werden inzwischen zentral in Wikidata gespeichert, wo sie über eine grafische Schnittstelle angereichert und verlinkt werden können. Die Daten in Wikidata sind aber nicht nur für die Wikimedia-Projekte verfügbar, sondern sie werden an vielen anderen Stellen weiterverwendet.

6 Verlinkung zwischen Artikeln zum gleichen Thema in unterschiedlichen Sprachversionen eines Wikimedia-Projekts.

7 Beispielsweise Wikivoyage, Wikiquote, Wikimedia Commons und Wikisource.

8 https://wikimediafoundation.org/our-work/wikimedia-projects/ (17.12.2020).

1.2 Besonderheiten von Wikidata

Wikidata ist mit der Vision verbunden, mehr Menschen mehr Zugang zu mehr Wissen zu geben. Zur Erreichung dieses Ziels bietet Wikidata eine Plattform, die es Menschen aus aller Welt erlaubt, ihre Daten zu teilen, anzureichern, zu verlinken und auch unmittelbar weiterzuverwenden, wobei der Zugang sowohl für Menschen als auch Maschinen möglich sein soll. Aus dieser Zielsetzung ergeben sich verschiedene essenzielle Anforderungen.

Frei: Die Daten von Wikidata werden unter CC0[9] veröffentlicht, d. h. sie sind für jede Person und für jeden Zweck frei verfügbar.

Kollaborativ: Die Inhalte von Wikidata werden gemeinschaftlich von einer weltweiten Community gesammelt und gepflegt. Um diese Kollaboration zu ermöglichen und bestmöglich zu unterstützen, stehen diverse Werkzeuge zur Verfügung, wie etwa Diskussionsseiten und Versionshistorien.

Breite Themenabdeckung: Wikidata ist keine Spezialwissensbasis, sondern deckt vor allem ein breites Themenspektrum von allgemeinen Daten über die Welt ab. Der genaue Zuschnitt der Daten ist dabei stark von den Bedarfen der Wikimedia-Projekte und ihrer Partner geprägt.

Enge Verbindung zu Wikipedia: Wikidata ist ein eigenständiges Projekt, profitiert aber von seiner engen Verbindung zu Wikipedia, sowohl inhaltlich als auch in den überlappenden Communitys der jeweils Editierenden.

Flexibles Datenmodell: Wikidata liegt ein flexibles und mächtiges Datenmodell (siehe Abb. 3) zugrunde. Es ermöglicht, Datenobjekten Bezeichnungen in vielen Sprachen zu geben, um den Inhalt jeder Person zugänglich zu machen, unabhängig von der gesprochenen Sprache. Zu jeder Entität können Aussagen gemacht werden. Jede Aussage kann außerdem qualifiziert werden, um sie in ihrem jeweiligen Kontext einzuordnen. Ferner kann jede Aussage referenziert werden, um sie zu belegen. Dies ist besonders wichtig, da Wikidata eine sekundäre Wissensbasis ist, also in erster Linie Daten enthält, deren Primärquelle sich an anderer Stelle befindet. Da sich unterschiedliche Primärquellen widersprechen können, ermöglicht es das Datenmodell, auch widersprüchliche Aussagen einzutragen. Die Komplexität der Welt kann dadurch in einem Maße abgebildet werden, das einem globalen Projekt gerecht wird. Die Ontologie von Wikidata ist emergent, das heißt sie entsteht aus Beziehungen, die in die Aussagen eingetragen werden (zum Beispiel *ist ein* oder *Unterklasse von*), die für spezifische Datenobjekte gelten.

9 https://creativecommons.org/publicdomain/zero/1.0/ (17.12.2020).

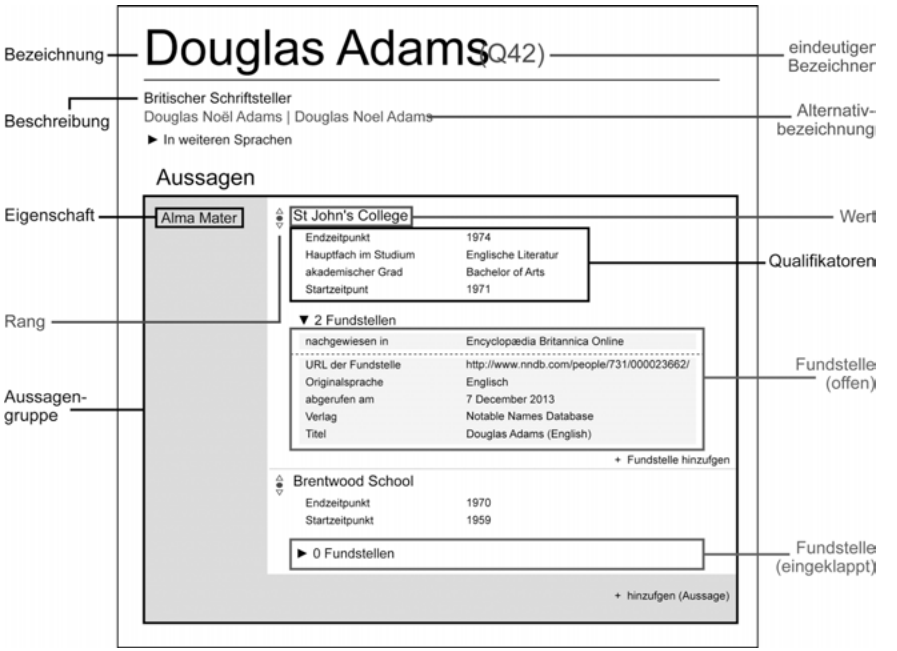

Abb. 3: Das Datenmodell von Wikidata an einem Ausschnitt des Datenobjekts zu Douglas Adams

1.3 Beispielanwendungen von Wikidata

Die Daten in Wikidata werden bereits jetzt auf verschiedenste Art und Weise genutzt. Nachfolgend werden schlaglichtartig einige Beispiele aufgeführt.

- Als Quelle von strukturiertem Grundlagenwissen: Infoboxen in Wikipedia, digitale persönliche Assistenten und andere Anwendungen nutzen Wikidata als Quelle strukturierten Grundlagenwissens, um z. B. Faktenfragen zu beantworten wie beispielsweise nach der Anzahl der Einwohner:innen einer Stadt.
- Als Entitätsprovider und Schlagwortvokabular: Die persistenten IDs für eine Vielzahl von Datenobjekten in Wikidata werden vielseitig genutzt, um Entitäten eindeutig und sprachunabhängig zu referenzieren. Dies wird unter anderem in Verfahren zur automatischen Eigennamenerkennung sowie bei der Verschlagwortung von Inhalten (Nachrichtenartikel, Bilder etc.) ge-

nutzt, um sie leichter auffindbar und durchsuchbar zu machen, z. B. bei Institutionen wie dem finnischen Radio- und Fernsehsender YLE.[10]
- Als Übersetzungsgrundlage: MapBox und andere Institutionen verwenden die Bezeichnungen der Datenobjekte von Wikidata, um z. B. Namen von Städten in ihren Anwendungen in verschiedenen Sprachen und Schriftsystemen anzeigen zu können.
- Als Ontologieprovider: Die Ontologie von Wikidata wird von Quora[11] und anderen Online-Plattformen genutzt, um ihre eigene Ontologie anzureichern und zu verbessern.
- Als Hub: Wikidata ist durch seine zahlreichen Verknüpfungen zu anderen Datenbanken ein zentraler Knotenpunkt im Linked Data Web. Dies macht sich z. B. die Browsererweiterung EntityExplosion zunutze. Beim Besuch einer Webseite listet sie Links zu anderen Webseiten auf, die eine Seite zur gleichen Entität haben, und gibt auf diese Weise den Nutzenden Zugang zu mehr Informationen.

2 Datenqualität in Wikidata

2.1 Bedeutung der Datenqualität in Wikidata

Das tägliche Leben von immer mehr Menschen wird grundlegend von Informationstechnologie beeinflusst – von Infografiken über Suchmaschinen bis hin zu digitalen persönlichen Assistenten. Wikidata liegt die Überzeugung zugrunde, dass die Daten, die die Basis für diese Technologien darstellen, frei und offen für alle sein müssen – und zwar sowohl der Zugang zu als auch die Bearbeitung der Daten. Je mehr Wikidata in und von derartigen Technologien genutzt wird, umso wichtiger wird es, dass die Daten in Wikidata von hoher Qualität sind. Mit steigender Durchdringung und Nutzung von Wikidata steigt aber auch das Interesse und die Motivation von böswilligen Agierenden, die Daten in ihrem Interesse zu beeinflussen (Oboler et al. 2010). Dies kann sich unter anderem in Vandalismus und der gezielten Streuung von Misinformation ausdrücken. Hinzu kommt, dass die Menge der Inhalte in Wikidata signifikant schneller wächst als die Community der Editierenden, was zur Folge hat, dass jede:r Editierende theoretisch für mehr und mehr Inhalt verantwortlich ist. Damit steigt die Ver-

10 https://yle.fi (17.12.2020).
11 https://www.quora.com (17.12.2020).

antwortung der Wikidata-Community und gleichzeitig wird die Aufgabe, die Datenqualität hoch zu halten, immer schwieriger.

2.2 Aspekte der Datenqualität in Wikidata

Datenqualität in Wikidata hat verschiedene Aspekte. Piscopo et al. (2019) fassen diese Qualitätsaspekte in ihrer Literaturübersicht in vier Gruppen zusammen:

- *Intrinsische Aspekte:* Diese beinhalten den Daten inhärente Eigenschaften wie Richtigkeit (sind die Daten korrekt?), Vertrauenswürdigkeit (kann ich den Daten vertrauen?) und Konsistenz (sind die Daten konsistent modelliert und eingetragen?).
- *Kontextuelle Aspekte:* Diese beinhalten Eigenschaften, die vom Kontext der Nutzung abhängen. Dazu gehören Relevanz (sind alle für mich wichtigen Daten vorhanden?), Vollständigkeit (sind die Daten, die ich brauche, komplett?) und Aktualität (sind die Daten für meinen Zweck aktuell genug?).
- *Aspekte der Repräsentation:* Diese betreffen die Form, in der die Daten verfügbar sind. Dazu gehören die Verständlichkeit der Darstellung (können sowohl Menschen wie auch Maschinen die Daten leicht und eindeutig interpretieren?) und die Interoperabilität (kann ich die Daten leicht in andere Systeme integrieren?).
- *Aspekte der Zugänglichkeit:* Diese Aspekte betreffen die Zugänglichkeit auch von Linked Data (verknüpfte Daten). Dazu gehören die Abrufbarkeit (sind die Daten schnell und persistent verfügbar?), Grad der Vernetzung (sind die Entitäten mit äquivalenten Einträgen anderer Datenquellen verlinkt?) und Fragen der Lizenzierung (darf ich die Daten in verbundenen Datenquellen für meine Zwecke einsetzen?).

2.3 Arten von Qualitätsproblemen in Wikidata

Angelehnt an die in Abschnitt 2.2 aufgeführten Qualitätsaspekte lassen sich Probleme in den Daten von Wikidata beschreiben und wie folgt klassifizieren:

- *Inkorrekte Daten:* Daten, die allgemein als nicht korrekt angesehen werden und berichtigt oder um Kontext ergänzt werden sollten.
- *Unbelegte Daten:* Daten, für die keine Quellenangabe vorliegt und die damit weder nachvollziehbar noch prüfbar sind.

- *Inkonsistente Modellierung:* Identische Typen von Daten, die auf verschiedenen Datenobjekten unterschiedlich modelliert werden und damit die Prüfung und Nachnutzung erschweren.
- *Ontologieprobleme:* Klassenhierarchien, die lokal sinnvoll erscheinen, aber global zu sinnfreien Beziehungen führen und damit die Nachnutzung erschweren.
- *Unvollständige Daten:* Daten, die teilweise oder gar vollständig fehlen und damit für Nachnutzende nur einen Teil der Realität widerspiegeln.

Diese Probleme können entweder absichtlich (Vandalismus) oder unabsichtlich (Versehen) entstehen. Unabsichtlich entstehen diese Probleme in den Daten durch Unkenntnis von Wikidata, fehlende Grundlagen der Wissensmodellierung oder fehlendes Wissen im speziellen Themengebiet der Daten.

2.4 Besonderheit der Datenqualität in Wikidata

Die Besonderheit von Wikidata liegt in der inhärenten Offenheit, die von der Wikimedia-Bewegung als hohes Gut betrachtet wird (sichtbar z. B. in der Projektvision „Stell dir eine Welt vor, in der jeder einzelne Mensch frei an der Summe allen Wissens teilhaben kann. Das ist unsere Verpflichtung.“[12]). Jeder Person soll es möglich sein, etwas beizutragen und somit die Wissensbasis vollständiger, akkurater und insgesamt nützlicher zu machen. Gleichzeitig macht es die Offenheit böswilligen Agierenden einfacher, die Datenqualität durch Vandalismus und falsche Aussagen zu beeinträchtigen. Es müssen also Werkzeuge und Prozesse entwickelt werden, die es der Editierenden-Community ermöglichen, die hohe Qualität der Daten in Wikidata zu erhalten, ohne die Offenheit zu kompromittieren.

Diese Offenheit als Grundprinzip von Wikipedia funktioniert bereits seit zwanzig Jahren. Wikipedia macht sich außerdem das Viele-Augen-Prinzip zunutze. Es besagt, dass es umso wahrscheinlicher ist, dass Fehler gefunden und behoben werden, je mehr Menschen sich etwas anschauen (Brändle 2005). Die weltweite Nutzendenschaft der Wikipedia hilft tagtäglich, Fehler zu beheben, die sie zufällig beim Lesen eines Wikipedia-Artikels bemerken. Dies lässt sich allerdings nicht ohne Weiteres auf Wikidata übertragen. Für Wikidata kommt erschwerend hinzu, dass ein Großteil der Menschen, die mit Daten aus Wikidata in Berührung kommen, nicht weiß, dass diese Daten aus Wikidata stammen. Die Daten von Wikidata sind unter CC0 frei verfügbar und auch für Drittanbie-

12 https://meta.wikimedia.org/wiki/Vision/de (17.12.2020).

tende frei nutzbar – ohne Verpflichtung, Wikidata als Quelle zu nennen. Daraus folgt, dass viele dieser Drittanbietenden ihre Nutzenden nicht in die Lage versetzen, gefundene Fehler in Wikidata zu beheben.

Als Lösung können den Nachnutzenden der Daten einfach zu implementierende Werkzeuge und Prozesse angeboten werden, die Fehlerkorrekturen und Ergänzungen ermöglichen. Abb. 4 zeigt 1) den bidirektionalen Kommunikationsfluss aus Informationen und Änderungen bei Wikipedia, 2) den bislang nur unidirektionalen Kommunikationsfluss bei der Nachnutzung von Wikidata sowie 3) den angestrebten zukünftigen Zustand.

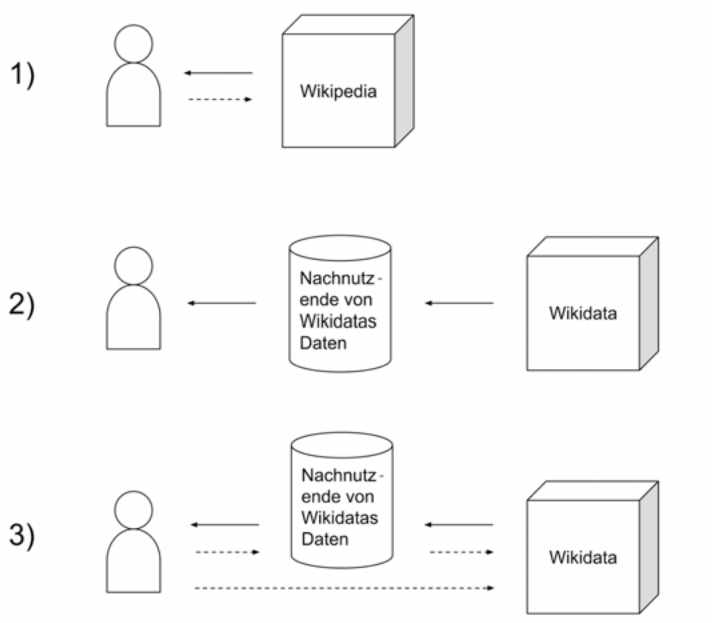

Abb. 4: Informations- und Änderungsfluss zwischen Wikidata und den Nutzenden der Daten aus Wikidata

Auf der anderen Seite kommt Wikidata zugute, dass es aus strukturierten Daten besteht. Hierdurch sind automatisierte Prüfungen und Fehlerbehebungen deutlich leichter möglich als zum Beispiel in den unstrukturierten Freitexten der Wikipedia.

2.5 Werkzeuge und Prozesse zum Finden und Beheben von Qualitätsproblemen in Wikidata

2.5.1 Grundprinzipien

Allen Qualitätswerkzeugen und -prozessen für Wikidata liegen verschiedene Prinzipien zugrunde. Diese werden nachfolgend verdeutlicht.

Qualität messbar machen: Qualität muss messbar sein, um einschätzen zu können, wo Wikidata aktuell steht und wie sich die Qualität über die Zeit verändert. Um ein umfassendes Bild zu erhalten, muss dies für einzelne Datenobjekte (zum Beispiel Marie Curie), für bestimmte Teilbereiche der Daten (zum Beispiel alle niederländischen Maler des 19. Jahrhunderts) sowie für Wikidata insgesamt möglich sein. Angesichts der Größenordnung von Wikidata mit derzeit mehr als 90 Millionen Datenobjekten ist eine manuelle Qualitätsbestimmung unmöglich, stattdessen muss sie automatisiert erfolgen.

Fehler automatisch finden und sichtbar machen: Viele Fehler in den Daten können automatisiert gefunden und sichtbar gemacht werden. Hierzu können Regeln und Heuristiken sowie maschinelle Lernverfahren eingesetzt werden. Um nicht versehentlich ungewöhnliche, aber legitime Änderungen zu verbieten, sollte dies allenfalls in Ausnahmefällen dazu führen, dass Änderungen bereits bei der Eingabe komplett verboten werden. Der beste Einsatzzweck ist vielmehr, Editierenden und Datennachnutzenden das Auffinden und Beheben von Fehlern zu erleichtern. Die ultimative Entscheidung sollte bei einem Menschen bleiben.

Mehr Augen auf die Daten lenken: Je mehr Menschen Daten aus Wikidata begegnen, umso wahrscheinlicher ist es, dass Fehler bemerkt und behoben werden. Dies geschieht nur zu geringem Maße in Wikidata selbst, sondern bisher vor allem in Wikipedia und den anderen Wikimedia-Schwesterprojekten, sowie mit Hilfe einer Vielzahl von Anwendungen und Visualisierungen außerhalb von Wikimedia. Zukünftig sollten auch bei Nachnutzenden geeignete Rückkanäle zu Wikidata aufgebaut werden.

Daten referenzieren, verlinken und vergleichen: Die Daten in Wikidata sollten mit Referenzen belegt und mit anderen Datenquellen verlinkt werden. Später können diese Referenzen und Verlinkungen dann genutzt werden, um Daten händisch oder automatisiert zu vergleichen und Diskrepanzen aufzuzeigen.

Hochqualitative Daten hervorheben: Daten, die bereits eine besonders hohe Qualität haben, sollten positiv hervorgehoben werden. Dies kann unter anderem durch die prominente Platzierung von Listen mit hochwertigen Inhalten, durch das Teilen von besonders vollständigen Abfrageergebnissen in sozialen

Medien oder die Nutzung der Daten in besonders wichtigen und weit verbreiteten Applikationen geschehen. Dadurch entstehen Anreize, andere Daten weiter zu verbessern und ebenfalls auf ein hohes Qualitätsniveau zu bringen.

2.5.2 Existierende Werkzeuge und Prozesse

Basierend auf diesen Prinzipien wurden bereits verschiedene Werkzeuge und Prozesse mit unterschiedlichem Reifegrad entwickelt, um die Datenqualität in Wikidata zu verbessern und hochzuhalten. Im Folgenden wird eine Auswahl vorgestellt.

Beobachtungsliste, Letzte Änderungen und Versionshistorie: Als Wiki-System umfasst MediaWiki standardmäßig eine Reihe von Werkzeugen, um Änderungen an den Inhalten leicht nachvollziehbar zu machen. Jeder Änderung einer Seite wird in der dazugehörigen Versionshistorie festgehalten und es ist erkennbar, welche Änderungen wann und von wem vorgenommen wurden. Editierende können einzelne Seiten auf eine Beobachtungsliste setzen, um leichter verfolgen zu können, welche Änderungen an Seiten, die sie interessieren, vorgenommen wurden. Zusätzlich werden auf der Seite *Letzte Änderungen* alle Änderungen der letzten Tage im Wiki aufgelistet, um einen globalen Überblick zu ermöglichen. Wikidata stellt all diese Funktionen zur Verfügung, hat aber mittlerweile eine Größe erreicht, bei der diese Standardwerkzeuge nicht mehr gut funktionieren und überdacht werden müssen.

Missbrauchsfilter: MediaWiki bietet auch bei Wikidata die Möglichkeit, Regeln zu definieren, aufgrund derer jede neue Änderung geprüft wird. Es ist unter anderem möglich, solche Änderungen mit einem Schlagwort zu versehen, um sie leichter auffindbar zu machen, die Änderung abzulehnen oder die Editiergeschwindigkeit der Nutzenden zu beschränken. Dies ist zum Beispiel hilfreich, um das Auffinden potenziell destruktiver Änderungen von unerfahrenen Nutzenden zu vereinfachen.

Automatisierte Einschätzung einzelner Änderungen: Für alle Änderungen in Wikidata wird mit Hilfe des Wikimedia-eigenen maschinellen Lernsystems ORES (Sarabadani et al. 2017) prognostiziert, wie hoch die Wahrscheinlichkeit ist, dass es sich um Vandalismus handelt. Diese Einschätzung wird genutzt, um die Aufmerksamkeit der Editierenden besonders auf solche Änderungen zu lenken, die eine hohe Vandalismuswahrscheinlichkeit haben, damit dieser zeitnah wieder rückgängig gemacht werden kann. Dies geschieht zum Beispiel in der Beobachtungsliste. Die vergleichsweise kleinteiligen Änderungen auf Wikidata mit wenig Kontext erschweren allerdings bisher gute Prognosen.

Automatisierte Qualitätsbewertung von Datenobjekten: Auch die Qualität aller Datenobjekte in Wikidata wird mit Hilfe von ORES bewertet. Sie werden dabei automatisiert in eine von fünf Qualitätsklassen von A (enthält alle relevanten Inhalte und Belege) bis E (ohne jegliche Inhalte und Belege) eingestuft. Diese Einstufung wird verwendet, um die Qualitätsentwicklung von Wikidata als Ganzes oder auch von Teilbereichen zu beobachten, sowie gezielt Datenobjekte mit niedriger Qualität zu finden und zu verbessern. Wichtige Qualitätsaspekte wie die Richtigkeit der Inhalte liegen jedoch außerhalb der Fähigkeiten von ORES.

Arbeitslisten: Die Editierenden von Wikidata können basierend auf einer Abfrage (meist in SPARQL formuliert) Arbeitslisten erstellen. Diese beinhalten entweder eine Datenmenge, die vervollständigt bzw. berichtigt werden muss (z. B. Nobelpreisträger:innen und zugehörige Portraitfotos, um fehlende Bilder zu ergänzen) oder eine Aufstellung von inkonsistenten Daten, die berichtigt werden müssen (z. B. Menschen, die gestorben sind, bevor sie geboren wurden, und die keine Zeitreisenden sind). Diese Arbeitslisten werden dann auch unter Hinzunahme der Versionskontrolle von MediaWiki über die Zeit auf ungewollte Veränderungen hin beobachtet, um einmal verbesserte Daten leichter in einem guten Zustand zu halten.

Visualisierung von Lücken und Tendenzen: Visualisierungen haben sich als besonders hilfreich erwiesen, um Lücken und Tendenzen in Wikidata sichtbar zu machen und Editierende in die Lage zu versetzen, Gegenmaßnahmen zu ergreifen. Erfolgreiche Beispiele hierfür sind Kartendarstellungen, die die geografische Verteilung der Datenobjekte visualisieren, sowie detaillierte Statistiken zur Geschlechterverteilung über die Datenobjekte (aufgeschlüsselt nach Beruf, Geburtsdekade, Land etc.). Diese helfen den Editierenden, konkrete Ansatzpunkte zu finden, um unerwünschte Lücken und Tendenzen zu beheben. Visualisierungen sollten auf weitere Bereiche ausgeweitet werden.

Constraint-Checks: Mit Hilfe von Constraint-Checks ist es den Editierenden auf Wikidata möglich, bestimmte Regeln für Eigenschaften zu definieren. Verletzungen dieser Regeln werden im Anschluss anderen Editierenden in der grafischen Oberfläche angezeigt, um sie auf eventuelle Fehler aufmerksam zu machen. Die Regeln können eine Vielzahl von Fällen abdecken. Mit ihnen kann man z. B. festlegen, dass es in Wikidata einen bestimmten Wert für eine ID in einer anderen Datenbank nur einmal geben sollte, um Duplikate leichter auffinden zu können. Oder man kann festlegen, dass eine Eigenschaft symmetrisch sein soll, um fehlende Verlinkungen aufzuzeigen. Des Weiteren kann man festlegen, dass bestimmte Eigenschaften nur mit Werten aus einem kontrollierten Vokabular genutzt werden sollten oder eine Eigenschaft nur bei Instanzen bestimmter Klassen verwendet werden sollte.

Schemata: Editierende können in Wikidata mit Hilfe der Strukturschema-Sprache Shape Expressions (ShEx) Schemata für Klassen festlegen.[13] Sie können somit einzelne Datenobjekte oder Teilbereiche von Wikidata gegen ein Schema prüfen und eventuelle Modellierungsfehler aufdecken. Dies hilft den Editierenden Teilbereiche von Wikidata konsistent zu modellieren und Abweichungen zu finden. Schemata können in ihrem Geltungsbereich überlappen und müssen nicht universell gültig sein. Die Nutzenden eines Schemas müssen allerdings immer den jeweiligen Kontext verstehen und es entsprechend anwenden. Deshalb werden Strukturschemata als Prüfwerkzeug verstanden, aber nicht zur automatisierten Verhinderung von Änderungen verwendet.

2.5.3 Zukünftige Werkzeuge und Prozesse

Mit dem fortschreitenden Wachstum von Wikidata müssen auch die Prozesse und Werkzeuge mitwachsen sowie neue entwickelt werden. Es folgt ein Ausblick auf wünschenswerte zukünftige Entwicklungen im Gebiet der Datenqualitätssicherung und -verbesserung.

Vergleiche mit anderen Datenbanken: Ein Großteil der Datenobjekte in Wikidata ist mit anderen Datenbanken, Katalogen und Webseiten verlinkt. Diese Links können genutzt werden, um die Daten in Wikidata mit den dort vorhandenen Daten zu vergleichen und eventuelle Diskrepanzen aufzuzeigen.

Feedbackschleifen mit Datennachnutzenden: Es muss den Nutzenden der Daten ermöglicht werden, Fehler zu melden oder direkt zu beheben. Das bedeutet, dass die Barriere zwischen der Person, die die Daten am Ende konsumiert (z. B. in einer Visualisierung auf einer Webseite), und Wikidata aufgebrochen werden muss. Dies wird abhängig von der Nachnutzung auf drei Wegen geschehen:

1. Nachnutzende leiten einen sinnvollen Teil der Fehlermeldungen, die sie von ihren Nutzenden erhalten, an Wikidata weiter, wo sie von den Editierenden bearbeitet werden. Dies hat den Vorteil, dass Fehlermeldungen gefiltert und priorisiert werden können, um die Editierenden nicht zu überlasten. Dieser Ansatz hat den Nachteil, dass er Personen, die Fehler finden, keinen leichten Pfad aufzeigt, selbst zur bzw. zum Editierenden zu werden. Die eigentliche Editierenden-Community vergrößert sich dadurch also nicht.
2. Nachnutzende machen Wikidata als Quelle ihrer Daten sichtbarer, um ihren Nutzenden zu ermöglichen, Fehler an der Quelle selbst zu beheben.

13 http://shex.io (17.12.2020).

Das hat den Vorteil, potenzielle neue Editierende direkt zu Wikidata zu führen. Dieser Ansatz hat den Nachteil, dass er potenziell viele Menschen ohne Erfahrung mit Wissensmodellierung oder offenem kollaborativem Arbeiten direkt zu Wikidata führt und die bestehende Community im schlimmsten Fall überwältigt.

3. Nachnutzende erlauben ihren Nutzenden über eine in das Produkt integrierte grafische Oberfläche, Wikidata zu bearbeiten. Der Vorteil dabei ist, dass Nachnutzende die volle Kontrolle über das Editieren und ihren speziellen Themenbereich haben und hilfreiche dedizierte grafische Oberflächen entwickeln können, die das Editieren stark vereinfachen. Der Nachteil ist, dass den Nutzenden dieser Oberflächen der weitere Kontext für ihre Änderung in Wikidata fehlt und ihnen kein leichter Pfad aufgezeigt wird, selbst zur bzw. zum Editierenden zu werden. Auch hier vergrößert sich die eigentliche Editierenden-Community also nicht.

Automatisierte Prüfung von Belegen: Jede Aussage in Wikidata kann mit einer Quellenangabe versehen werden, um sie zu belegen. Im Laufe der Zeit kann sich entweder ein Inhalt in Wikidata ändern oder der Inhalt der zugehörigen Quelle. Ein neues System könnte es erlauben, Inhalte regelmäßig und automatisiert daraufhin zu prüfen, ob sie immer noch durch die angegebene Quelle belegt werden. Dies würde einem unbegründeten Vertrauen in falsche Belege entgegenwirken und gleichzeitig helfen, veraltete Daten aufzufinden.

Signierte Aussagen: Für einen Teil der Aussagen in Wikidata gibt es Organisationen, die als autoritative Quelle für diese Daten fungieren. Dies sind z. B. Identifier in Normdatenbanken. Wenn diese in Wikidata eingetragen werden, könnten sie inklusive ihrer Quellenangabe signiert werden. Änderungen in den Daten, die die Signatur brechen, könnten den Editierenden angezeigt werden, um zu bewerten, ob die Änderung gerechtfertigt war und neu signiert werden muss oder zurückgesetzt werden sollte. Die Signatur würde Nachnutzenden der Daten außerdem einen weiteren Anhaltspunkt für die Vertrauenswürdigkeit der Daten geben.

Auffinden von Lücken und Tendenzen: Bisherige Analysen und Visualisierungen der Daten in Wikidata konzentrieren sich auf Lücken und Tendenzen in Bezug auf geografische Abdeckung und Geschlechterrepräsentation. Dies kann und sollte auf weitere Dimensionen ausgeweitet werden. Detaillierte und aktionsfokussierte Visualisierungen und Applikationen sollten den Editierenden zur Verfügung stehen, um unerwünschte Lücken und Tendenzen gezielt bearbeiten zu können.

Arbeitslistensystem: Es gibt eine Vielzahl an Werkzeugen innerhalb und außerhalb von Wikidata, die Probleme in den Daten finden. Dies führt dazu, dass

Editierende oft nur einen Bruchteil davon im Blick haben, und gleichzeitig erschwert es die Integration und Verbreitung von neuen Werkzeugen. Es sollte also in Zukunft ein zentrales System geben, das aus verschiedenen Quellen mit Datenproblemen befüllt wird und diese dann verschiedenen anderen Werkzeugen zentral zur Verfügung stellt. Eine Forscherin oder ein Forscher könnte dann z. B. einen neuen Algorithmus entwickeln, der eine bestimmte neue Art von Fehler entdeckt und diese in das System speisen. Die Fehler könnten dann von verschiedenen anderen Werkzeugen genutzt werden, um die Abarbeitung dieser Fehler zu vereinfachen.

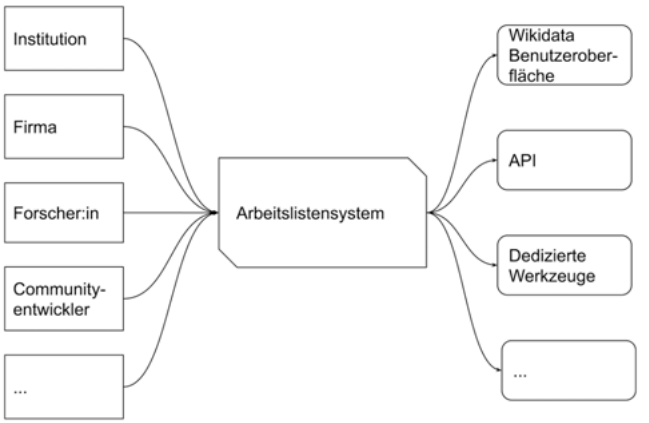

Abb. 5: Aufbau des zukünftigen Arbeitslistensystems für Wikidata, das aus verschiedenen Quellen Aufgaben (potenzielle Fehler, etc.) aufnehmen kann und diese in verschiedenen Werkzeugen für die Editierenden zugänglich macht

3 Praxisbeispiel: Erfahrungen bei der Nachnutzung von Wikidata durch das Deutsche Forschungszentrum für Künstliche Intelligenz (DFKI)

Die automatische Erkennung von Eigennamen (Named Entity Recognition, NER) ist eine wichtige Komponente zahlreicher Anwendungen im Bereich Natural Language Processing, Sprachtechnologie und Information Retrieval. Sie stellt für viele Downstream-Anwendungen eine wichtige Grundlage dar (wie

z. B. Suchmaschinen und Information Retrieval einschließlich automatischer Erschließung, Chatbots, automatische Textzusammenfassung, maschinelle Übersetzung, Sentimentanalyse und viele mehr).

Als sehr große und frei verfügbare Datensätze mit allgemeinem Weltwissen werden Wikidata sowie auch Wikipedia oft benutzt, um Systeme auf eine Eigennamenerkennung zu trainieren und deren Ergebnisse zu evaluieren (Nothman et al. 2013; Ghaddar und Langlais 2017; Li et al. 2019). Wir verfolgen ebenfalls diesen Ansatz, wobei wir ihn speziell für diesen Beitrag mit Fragen bzgl. der Datenqualität verbunden haben. Zu diesem Zweck haben wir fünf unterschiedliche Systeme für die Erkennung von Eigennamen in englischsprachigen Texten auf Basis der fünf ORES-Qualitätsklassen trainiert (A–E, siehe Abschnitt 2.5.2). Der Qualitäts-Score bezieht sich dabei auf die Wikidata-Datenobjekte. Systeme für Eigennamenerkennung müssen mit Fließtexten trainiert werden, bei denen die Eigennamen in den Texten explizit annotiert sind (Rehm 2020). Dafür nutzen wir jeweils den Fließtext des Wikipedia-Artikels, der dem Wikidata-Datenobjekt des Qualitäts-Scores zugeordnet ist.

Für jeden der fünf Qualitäts-Scores liegt eine Liste von Wikidata-Datenobjekten vor, die den entsprechenden Qualitäts-Score besitzen. Anhand dieser fünf Listen erstellen wir fünf Sammlungen, die die zugehörigen Wikipedia-Artikel umfassen. Diese fünf Sammlungen von Wikipedia-Artikeln werden im Anschluss benutzt, um jeweils ein Modell für die Eigennamenerkennung zu trainieren und zu evaluieren. Die Grundidee dabei ist es zu überprüfen, ob die Qualität eines Wikidata-Datenobjekts mit der Performanz des zugehörigen NER-Modells korreliert, ob sich also ein geringer Qualitäts-Score in einer eher niedrigen Performanz des NER-Modells manifestiert. Es handelt sich demnach um eine Art extrinsische Evaluation des Wikidata Qualitäts-Scores.

3.1 Technische Umsetzung

Basierend auf jeder der fünf Sammlungen von Wikipedia-Artikeln trainieren wir ein Modell, das vier unterschiedliche Typen von Eigennamen erkennen kann: Personen (PER), Organisationen (ORG), Orte (LOC) und Sonstiges (MISC). Der allgemeine Ansatz basiert auf BERT (Devlin et al. 2019), einem Transformer-basierten neuronalen Sprachmodell, das derzeit in einer Reihe von NLP-Standard-Tasks den Stand der Forschung definiert. Die Implementierung unseres Systems ist frei zugänglich.[14]

14 https://gitlab.com/qurator-platform/dfki/srv-bertner-en (17.12.2020).

Wir benutzen die Qualitäts-Scores von September 2020[15] und verknüpfen die ID des jeweiligen Wikidata-Datenobjekts mit dem jeweils entsprechenden Wikipedia-Artikel. Die Inhalte der Wikipedia-Seite erfassen wir dann wiederum mit einer Python-Bibliothek[16], die gleichzeitig die Verknüpfungen im Fließtext verarbeiten kann. Wir überprüfen den Typ einer jeden Verknüpfung mittels einer SPARQL-Anfrage: Wenn es sich um eine http://dbpedia.org/ontology/Person, http://dbpedia.org/ontology/Organisation oder http://dbpedia.org/ontology/Location handelt, annotieren wir im Fließtext jeweils eine PER-, ORG- oder LOC- Instanz, die anschließend für das Training der NER-Modelle verwendet werden und zwar jeweils für die fünf unterschiedlichen Qualitäts-Scores.

Der Vorteil dieses Verfahrens, das auch als *distantly supervised machine learning* bezeichnet wird, ist, dass wir in relativ kurzer Zeit viele annotierte Trainingsdaten erhalten. Der entscheidende Nachteil ist, dass die Daten nicht von Menschen überprüft wurden und evtl. fehlerhafte Annotationen enthalten. Deswegen benutzen wir für die Evaluation der fünf Modelle eine Teilmenge des englischen Datensatzes von Nothman et al. (2013), weil diese bereits überprüft und validiert worden sind. Wir evaluieren die fünf NER-Modelle jeweils anhand der ersten 50 000 Instanzen des Datensatzes und berechnen die Anzahl korrekt klassifizierter Entitäten.

3.2 Datenanalyse

Die Wikidata-Datenobjekte sind bzgl. ihrer jeweiligen Anzahl in den fünf Qualitätsklassen (von A bis E) sehr unterschiedlich verteilt. Um den Effekt dieser Ungleichverteilung auf die Performanz zu minimieren, begrenzen wir die Anzahl der Instanzen pro Qualitätsklasse auf 15 000 (siehe Tabelle 1). Somit erreichen wir eine balancierte Verteilung bzgl. der fünf Qualitätsklassen, jedoch bleibt eine geringe Diskrepanz zwischen den Entitätsklassen erhalten. Die PER- und LOC-Klassen besitzen eine ähnliche Anzahl Instanzen, während die Anzahl Instanzen für die ORG-Klasse sehr klein ist (siehe Tabelle 1, letzte Zeile).

15 https://analytics.wikimedia.org/published/datasets/wmde-analytics-engineering/Wikidata/WD_QualitySnapshots/wikidata_quality_snapshot_202009.tsv.gz (17.12.2020).
16 https://pypi.org/project/wikipedia/ (17.12.2020).

Tab. 1: Anzahl der Entitäten, aufgeteilt nach Qualitäts- und Entitätsklassen.

Qualitätsklasse	PER	ORG	LOC	Anzahl Instanzen (Qualitätsklasse)
A	9 119	2 842	3 039	15 000
B	5 205	3 560	6 235	15 000
C	4 648	3 521	6 831	15 000
D	5 352	3 017	6 631	15 000
E	5 192	2 902	6 906	15 000
Anzahl Instanzen (Entitätsklasse)	29 516	15 842	29 642	

3.3 Experimentelle Ergebnisse

Basierend auf den Daten (siehe Abschnitt 3.2) haben wir fünf NER-Modelle trainiert. Tab. 2 stellt die Ergebnisse dar. Die Ergebnisse des Experiments legen keinen Zusammenhang zwischen Wikidata-Qualitätsklassen und der Genauigkeit der Eigennamenerkennung nahe. Alle Klassen erzielen eine vergleichbar hohe Genauigkeit von etwa 88 %.

Tab. 2: Genauigkeit der Eigennamenerkennung pro Qualitätsklasse

Qualitätsklasse	Genauigkeit (*accuracy*) in Prozent
A	87,95
B	88,05
C	88,07
D	88,03
E	88,11

Für eine weitere Analyse betrachten wir die Verteilung der Trainingsdaten nicht in Bezug auf die Anzahl der Instanzen pro Qualitätsklasse, sondern bezüglich der verfügbaren Wörter in den Trainingsdaten. Zwar haben wir durch die Vorverarbeitung eine Gleichverteilung auf Qualitätsklassenebene erreicht, jedoch bleibt die Anzahl der Wörter davon unberührt. Gleichwohl kann die Wortmenge die Performanz beeinflussen. Tab. 3 stellt die Verteilung der Wörter pro Qualitätsklasse dar.

Tab. 3: Verteilung der Wörter pro Qualitätsklasse

Qualitätsklasse	Wörteranzahl
A	729 779
B	1 083 633
C	674 798
D	671 038
E	770 544

Die gleiche Anzahl von Entitäten verteilt über eine große Anzahl von Wörtern bedeutet entweder eine geringe Anzahl von im Text verlinkten Entitäten, die ihren eigenen Wikipedia-Artikel haben, oder viele Entitäten sind nicht als solche im Text verlinkt. Auf Basis dieser Zahlen lässt sich dies nicht abschließend beurteilen. Eine aufwendige manuelle Analyse wäre notwendig.

Sollten viele Verlinkungen zu Entitäten fehlen, spräche dies dafür, dass viele Annotationen im Datensatz nicht vollständig sind. Fehlende Annotationen beeinflussten den Trainingsprozess und die Performanz insofern, dass die Falsch-Negativ-Fehlerquote erhöht wäre. Diese Vermutung wird jedoch nicht von den empirischen Ergebnissen bestätigt. Die Qualitätsklasse B hat die meisten Wörter, während D die wenigsten hat. Gleichzeitig ist die Performanz für B und D vergleichbar (88,05 % und 88,03 % Genauigkeit). Des Weiteren können wir keinen Anstieg im Vergleich zwischen der höchsten (A) und der niedrigsten (E) Qualitätsklasse in Bezug auf die Qualität und die Wahrscheinlichkeit für fehlende Annotationen feststellen.

Insgesamt zeigt unser Experiment keinen Zusammenhang zwischen den Wikidata-Qualitätsklassen und der Performanz der NER-Modelle, die mit Wikipedia-Artikeln und entsprechenden Annotationen trainiert wurden. Bei dem Experiment sollte aber beachtet werden, dass wir es auf einem Silberstandard[17] trainiert und ausgewertet haben. Die zugrundeliegenden Daten wurden also nicht intellektuell kuratiert, sondern automatisch erzeugt. Diese Vorgehensweise hat den Nachteil einer potenziell hohen Falsch-Negativ-Fehlerquote, die durch unvollständige Verlinkung von Entitäten verursacht werden kann. Ferner haben wir eher geringe Mengen von Daten für das Training der Modelle verwendet. Abschließend kann beobachtet werden, dass nicht alle Faktoren, die sich positiv auf die Wikidata-Qualitätsklassen auswirken (z. B. Multilingualität, d. h. Verlinkung von Wikidata-Datenobjekten mit korrespondierenden Wikipedia-Ar-

17 In vergleichbaren Experimenten konnte anstatt von 88 % eine Genauigkeit von 95 % erzielt werden, vorausgesetzt die Modelle werden nach Nothman et al. (2013) trainiert und ausgewertet.

tikeln in unterschiedlichen Sprachen), zugleich einen Einfluss auf die Verwendung von Wikidata-Daten für das Training maschineller Verfahren besitzen, d. h. es existiert zumindest für die automatische Erkennung von Eigennamen gerade *keine* Korrelation zwischen der Datenqualität von Wikidata und der Performanz von Modellen, die mit Wikidata-Daten trainiert wurden.

4 Zusammenfassung

Wissensbasen sind der Grundpfeiler einer Vielzahl technologischer Anwendungen wie persönlicher digitaler Assistenten oder Suchmaschinen. Wissensbasen stellen systematisch gepflegte Sammlungen von Entitäten bzw. semantischen Konzepten zur Verfügung, wobei die Qualität der Wissensbasis von entscheidender Bedeutung für die Qualität der Inhaltserschließung ist.

Weil sie von Freiwilligen betrieben und gepflegt wird, ist die Bemessung und Sicherstellung der Datenqualität bei Wikidata eine große Herausforderung. In diesem Beitrag erläutern wir aktuelle Verfahren zur Qualitätsmessung, die bei Wikidata eingesetzt werden, welche Aspekte bei der Qualität berücksichtigt werden und inwiefern die Datenqualität für nachgelagerte Anwendungen eine Rolle spielt. Bei der Messung der Qualität wird zum einen auf die Community und deren Feedback vertraut, zum anderen werden aber auch automatische Verfahren genutzt. Des Weiteren zeigen wir auf, welche Maßnahmen zukünftig eingesetzt werden könnten, um eine weitere Verbesserung der Datenqualität zu erreichen.

Neben der Wikimedia-Sicht betrachten wir die Datenqualität von Wikidata auch aus der Anwendungssicht. In einem Experiment untersuchen wir empirisch den Zusammenhang zwischen der Datenqualität und der Performanz von Eigennamenerkennung (NER) als einem sprachtechnologischen Anwendungsbeispiel. Wir trainieren ein NER-Modell auf einem speziell für dieses Experiment erzeugten Datensatz, der Wikipedia-Artikel in fünf unterschiedliche Klassen gruppiert und zwar auf Basis der korrespondierenden Wikidata-Qualitätsklasse. Anschließend wird die Performanz des Modells auf Basis eines kuratierten Goldstandards gemessen. Die empirischen Ergebnisse zeigen keinen unmittelbaren Zusammenhang zwischen den Wikidata-Qualitätsklassen und der Performanz der Eigennamenerkennung, was insbesondere an den Kriterien liegt, die an *intellektuell* kuratierte Wissensbasen einerseits und *maschinell* trainierte Verfahren andererseits gelegt werden, denn diese unterscheiden sich in fundamentaler Weise. Während für die von Freiwilligen gepflegte Wissensbasis Wikidata Verlinkungen, Evidenzen und Kontextualisierungen eine große Rolle spielen,

ist für den Einsatz von Wikidata in einem maschinellen Lernverfahren lediglich die Kategorisierung eines Datenobjekts sowie die Menge annotierter Beispiele von Bedeutung Dieses Spannungsfeld zwischen der *intellektuellen* Bewertung der Datenqualität von Wikidata-Datenobjekten und der Nutzung von Wikidata-Daten für *maschinelle* Lernverfahren werden wir im Rahmen von zukünftigen Arbeiten genauer untersuchen.

5 Danksagung

Dieser Beitrag wurde im Rahmen des vom Bundesministerium für Bildung und Forschung (BMBF) geförderten Projektes QURATOR (Unternehmen Region, Wachstumskern, Projektnr. 03WKDA1A) erstellt.

6 Literaturverzeichnis

Brändle, Andreas: Zu wenig Köche verderben den Brei: eine Inhaltsanalyse der Wikipedia aus Perspektive der journalistischen Qualität, des Netzeffekts und der Ökonomie der Aufmerksamkeit. Lizentiatsarbeit. Universität Zürich 2005.

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). Minneapolis, Minnesota: Association for Computational Linguistics 2019. S. 4171–4186. http://dx.doi.org/10.18653/v1/N19-1423.

Ghaddar, Abbas und Phillippe Langlais: WiNER: A Wikipedia Annotated Corpus for Named Entity Recognition. In: Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Asian Federation of Natural Language Processing 2017. S. 413–422. https://www.aclweb.org/anthology/I17-1042 (4.1.2021).

Li, Maolong, Qiang Yang, Fuzhen He, Zhixu Li, Pengpeng Zhao, Lei Zhao und Zhigang Chen: An Unsupervised Learning Approach for NER Based on Online Encyclopedia. In: Web and Big Data. Third International Joint Conference, APWeb-WAIM 2019, Proceedings Part I. Hrsg. v. Jie Shao, Man Lung You, Masashi Toyoda, Dongxiang Zhang, Wei Wang und Bin Cui. Cham: Springer 2019. S. 329–344. https://doi.org/10.1007/978-3-030-26072-9_25.

Nothman, Joel, Nicky Ringland, Will Radford, Tara Murphy und James R. Curran: Learning multilingual named entity recognition from Wikipedia. In: Artificial Intelligence (2013) Bd. 194. S. 151–175. https://doi.org/10.1016/j.artint.2012.03.006.

Oboler, Andre, Gerald Steinberg und Rephael Stern: The Framing of Political NGOs in Wikipedia through Criticism Elimination. In: Journal of Information Technology & Politics (2010) Bd. 7 H. 4. S. 284–299. https://doi.org/10.1080/19331680903577822.

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

Rehm, Georg: Observations on Annotations. In: Annotations in Scholarly Edition and Research. Functions, Differentiation, Systematization. Hrsg v. Julia Nantke und Frederik Schlupkothen. Berlin, Boston: De Gruyter 2020. S. 299–324. https://doi.org/10.1515/9783110689112-014.

Sarabadani, Amir, Aaron Halfaker und Dario Taraborelli: Building automated vandalism detection tools for Wikidata. CoRR abs/1703.03861, 2017. http://arxiv.org/abs/1703.03861.

Vrandečić, Denny und Markus Krötzsch: Wikidata: a free collaborative knowledgebase. In: Communications of the ACM (2014) Bd. 57 Nr. 10. S. 78–85. https://doi.org/10.1145/2629489.