Linked Data & VGI – eine komparative Qualitätsanalyse für Deutschland, Österreich und die Schweiz auf Basis von Wikidata und OpenStreetMap

Linked Data & VGI – A Comparative Analysing of Quality for Germany, Austria and Switzerland Based on Wikidata and OpenStreetMap

Timo Homburg, Pascal Neis

Hochschule Mainz · timo.homburg@hs-mainz.de

Zusammenfassung: In diesem Beitrag werden die Ergebnisse einer vergleichenden Analyse zwischen verlinkten Wikidata- und OpenStreetMap(OSM)-Geodaten im Gebiet von Deutschland, Österreich und der Schweiz vorgestellt. Hierfür werden Metadaten von OSM und Wikidata aufgegriffen und gegenübergestellt. Anschließend werden mittels direkter Vergleiche der verlinkten Objekte verschiedene Beziehungen untersucht. Die Studie zeigt die Unterschiede in der Attributabdeckung der untersuchten Länder, typische Annotationen, typische Annotationsfehler und den Grad der Übereinstimmung zwischen Attributen und Geometrien.

Schlüsselwörter: Wikidata, OpenStreetMap, Qualitätsanalyse, Vergleich

Abstract: *In this publication we present results of a comparative study of Wikidata and OpenStreetMap (OSM) in the area of Germany, Austria and Switzerland. We include metadata of OSM and Wikidata, compare the two datasets on an object-by-object basis and on equivalent properties as defined by the respective communities. Our results give an indication about the tag coverage of the respective countries, which objects are typically associated with a wikidata tag, which mistakes are commonly made when annotating OSM objects with wikidata and the equality and equivalence of the respective Wikidata and OSM objects.*

Keywords: *Wikidata, OpenStreetMap, quality analysis, comparison*

1 Einleitung

Die Qualitätssicherung von Geodaten gewinnt durch die Nutzung von frei verfügbaren räumlichen Informationen in verschiedenen Anwendungsszenarien in immer mehr Applikationen an Relevanz. Neben der Positionsgenauigkeit spielt dabei der Umfang und die Richtigkeit der Attribute der genutzten Daten eine zunehmend größere Rolle. In den letzten Jahren wurde bereits vielfach die Qualität durch relative Vergleichsanalysen zwischen proprietäreren und freiverfügbaren oder offenen Datenquellen untersucht. Inzwischen gibt es aber durch das Web 3.0 immer mehr Informationen im World Wide Web, die durch eine Verlinkung miteinander verbunden werden. Diese neu vorhandenen Zusatzinformationen können auch für den erwähnten Anwendungsfall der Qualitätssicherung nützlich sein. Naheliegende Quellen für derartige Informationen sind neben eigens zu integrierenden Daten (z. B. von offiziellen staatlichen Stellen) vor allem (Linked) Open Data Repositories, welche bereits vorklassifizierte und vorkategorisierte Daten bereitstellen. In dieser Publikation untersuchen wir die

AGIT – Journal für Angewandte Geoinformatik, 5-2019, S. 141-155. © Wichmann Verlag, VDE VERLAG GMBH · Berlin · Offenbach. ISBN 978-3-87907-669-7, ISSN 2364-9283, eISSN 2509-713X, doi:10.14627/537669013.

Unterschiede und Gemeinsamkeiten zwei der größten Datenbanken für Linked Data und *Volunteered Geographic Information* (VGI): Wikidata und *OpenStreetMap* (OSM).

- Der Begriff VGI beschreibt Geoinformationen, die üblicherweise von Laien erfasst und in einer organisierten Form zum Beispiel im Internet für jedermann unter einer offenen Lizenz bereitgestellt werden. Goodchild (2007) gibt einen guten Überblick über die Entstehung von VGI. OSM ist dabei das wohl größte kollaborative Projekt zur Sammlung von offenen Geodaten. Ramm (2011) gibt eine Übersicht über die Geschichte und Historie von OSM sowie die praktische Bedeutung des Projektes in der Geocommunity.
- Ähnlich wie OSM ist Wikidata eine frei bearbeitbare Wissensdatenbank zur Unterstützung verschiedenster Wikimedia-Projekte (Wikipedia, Wikivoyage etc.; Erxleben, 2014; Vrandrevic, 2012, 2014) und enthält zum jetzigen Zeitpunkt mehr als 53 Millionen Instanzen. Im Gegensatz zum OSM-Projekt hat das Wikidata-Projekt nicht primär das Ziel, eine Geodatenbasis aufzubauen: Jedoch sind auch geographische Verortungen in vielen Konzepten in Wikidata enthalten, die die Arbeit mit den Daten erleichtern sollen.

2 Datenqualität

Die Literatur definiert Datenqualität auf verschiedene Weisen. Devillers (2005) definiert Datenqualität als Fitness For Use eines Datensets für einen spezifischen Anwendungsfall. Senaratre (2016) und Goodchild (2012) geben einen guten Überblick über Datenqualitätsanalysen im VGI-Bereich. Generell werden Datenqualitätsdimensionen und Kriterien definiert, welche mit Datenqualitätsmetriken überprüft werden können. Mocnik (2018) definiert eine Ontologie, in der Datenqualitätskriterien und Metriken festgehalten werden können und bietet somit eine Übersicht über die Möglichkeiten von Datenqualitätserfassungen. Einen ähnlichen Ansatz verfolgt die Datenqualitätsvokabular (DQV) des W3C. Die Relevanz von Datenqualitätskriterien zeigen die Analysen von Neis (2012) auf, welche die Veränderungen in OpenStreetMap über einen längeren Zeitraum beschreiben.

In der Vergangenheit wurden bereits zahlreiche Datenqualitätsanalysen mit OSM durchgeführt. Mit einer der ersten war Haklay (2008), der eine vergleichende Untersuchung von Straßen in England durchführte. Mooney (2010) führte Qualitätsmetriken über Attribut-, Abdeckung und Formvergleiche ein und evaluierte diese beispielhaft auf der Fläche von Irland. Neis (2010) verglich das deutsche OSM-Straßennetz mit proprietären Daten für Deutschland. Weitere Untersuchungen und Vergleiche mit anderen Datenquellen sind in Neis & Zielstra (2014) zu finden. Einen anderen Ansatz wählten Barron et al. (2014). Sie haben eine Anwendung für intrinsische Qualitätsanalysen von OSM-Daten vorgestellt.

Zusätzlich existieren verschiedene Studien, die sich mit Wikidata und OSM beschäftigen. Almeida (2016) untersuchte eine Methode, um Korrespondenzen zwischen OSM-Straßen in Rom und zugehörigen Wikidata-Straßeneinträgen halbautomatisiert zu matchen. Sie präsentierten die Ergebnisse ihrer Studie auf einer modifizierten OSM-Karte. Leyh (2017) stellte ein Projekt vor, welches OSM-Daten mit Citizen-Science-Daten in einer Open-Data-Cloud u. a. unter zu Hilfenahme von Wikidata-Daten matchte. Majic (2017) stellte einen Ansatz für das Matchen von semantisch äquivalenten Keys in OSM vor. Des Weiteren existieren eine Reihe von Anwendungen, die jedoch primär der Verlinkung von Wikidata-Objekten zu OSM dienen. OSM-Wikidata-Link (https://osm.wikidata.link) versucht, Wikidata-Items und OSM-

Items anhand ihrer räumlichen Distanz und anhand übereinstimmender Attribute zu matchen. Der Wikidata-OSM-Distance-Visualizer (https://osmlab.github.io/wikidata-osm) zeigt die räumliche Distanz zwischen in OSM annotierten Wikidata-Konzepten und der OSM-Geometrie an. Der Hintergrund hierbei ist, eine Fehlersuche bei der Annotation zu ermöglichen.

Bei den vorgestellten Studien wurde entweder Wert auf eine Integration und Nutzung von OSM- und Wikidata-Daten gelegt oder Methoden zur Verbesserung der Integration und der Linkings von Wikidata und OSM entwickelt. Die erwähnten OSM-Qualitätsuntersuchungen konzentrierten sich unter anderem auf den Vergleich der Geometrien zueinander und verwendeten jeweils entweder die gleiche Kartenressource oder eine Referenzkartenressource. Im Rahmen dieser Studie wird hingegen ein semantischer Vergleich der Elemente von OSM sowie Wikidata durchgeführt um Übereinstimmungen und Widersprüche aufzuzeigen.

2.1 Strukturelle Unterschiede zwischen Wikidata und OSM

Auch wenn OSM und Wikidata beide ähnliche Daten modellieren können, sind die strukturellen Unterschiede und die Motivation hinter den beiden Wissensbasen durchaus unterschiedlich.

2.2 Annotation

OSM folgt der in der GIS-Welt traditionellen Beschreibung eines geographischen Objektes durch eine Geometrie und einer Liste von Attributen (im OSM-Projekt als Key-Value-Paar oder auch als Tag bezeichnet). Hierbei hat sich die Community auf Empfehlungen für die Verwendung von weitverbreiteten Keys für spezielle Zwecke geeinigt und diese Empfehlungen werden mehr oder minder durch die Community befolgt. Sollen mehrere Values pro Key abgebildet werden, wie z. B. bei dem sprachlich differenzierten name-Tag, werden neue Keys erstellt, welche die jeweilige Variante des Values aufnehmen können (z. B. name:de, name:en). Im Gegensatz hierzu kann Wikidata für einen gleichen Key mehrere Values enthalten. Beispielsweise ist es durchaus möglich, für die Beschreibung P17 (country) mehrere Werte zu erhalten.

2.3 Klassifizierung

Eine Klassifizierung von Objekten findet in OSM zum Beispiel über das amenity-Tag statt, in dem die Nutzung des Objektes festgehalten wird. Alternativ gibt es jedoch weitere Möglichkeiten, die Funktion eines Objektes durch andere Tags klarzustellen. In Wikidata wird die Klassifizierung eines Objektes durch die Relation P31 (instance of) festgehalten. Wie im letzten Abschnitt beschrieben, kann diese Relation mehrere Werte enthalten, sodass eine Klassifizierung in mehreren Kontexten dargestellt werden kann.

2.4 Linked Data

Ein weiterer Unterschied besteht darin, dass Wikidata nach dem Linked-Data-Prinzip auf weitere Konzepte mit Zusatzinformationen verweisen kann. Hierbei ist es durchaus möglich, dass Informationen, die in OSM zwangsläufig am jeweiligen Objekt annotiert sind, in Wikidata in einem referenzierten Objekt gespeichert werden. Ein Beispiel hierfür ist die Modellierung von Adressen als eigene Instanzen in Wikidata. Da OSM keine Linked-Data-Res-

source ist, werden hier mehrere Tags mit einem ähnlichen Namen verwendet, um die Adresse zu modellieren (addr-Tags). Auch wenn dieser Umstand bekannt ist, wird in diesem Beitrag die weitergehende Objektstruktur nicht betrachtet.

2.5 Beziehungen zu WFS- und WMS-Services

Zur Zeit dieser Publikation ist die Nutzung von Geographic-Linked-Open-Data-Ressourcen im Vergleich zu geographischen Webservices, welche von der OGC standardisiert wurden, noch stark rückläufig. Dies mag mit fehlenden Möglichkeiten der Abfrage in der Sprache (Geo)SPARQL zusammenhängen, welche im Vergleich zu z. B. POSTGIS nur Beziehungen zwischen Geometrien abbilden kann und manipulierende, geometrieerstellende sowie Rasterdatenmanipulation in Queries nicht beherrscht. Dennoch ist außerhalb der Geocommunity ein Trend in verschiedenen Forschungsrichtungen zu einer vermehrten Nutzung von Linked Data erkennbar. Die Möglichkeiten einer schnellen, unkomplizierten, dezentralen und standardisierten Datenkombination machen Linked Data jedoch auch vermehrt für die Geocommunity attraktiv. Ein weiterer Aspekt, welcher eher für Linked-Data-Ressourcen als für traditionelle OGC-Webservices spricht, ist der Aspekt der Auffindbarkeit derselben. Zwar existieren mit den OGC Catalog Web Services Möglichkeiten, geographische Ressourcen auffindbar zu machen, jedoch werden diese oft nur unzureichend genutzt, und selbst die Catalog Web Services müssen wiederum gefunden werden. Hier bietet das Semantic Web durch die automatische Klassifizierung durch vereinheitlichte Vokabularien bessere Möglichkeiten der Auffindbarkeit für klassifizierte thematische Daten, welche durch Federated Queries ebenfalls einfach zugreifbar sind. Aus Sicht des Semantic Webs können OGC Webservices jedoch auch Potenziale bieten. Wenn OGC Webservices beispielsweise durch adequate RDF-Vokabularien im Semantic Web besser auffindbar und somit integrierbar wären, würde das Geospatial Semantic Web dadurch sehr gewinnen. Das Semantic Web bietet hier einen integrierenden Aspekt, welcher durch die aktuelle Struktur von OGC Webservices nur unzureichend dargestellt werden kann.

3 Datenaufbereitung

Die durchgeführte komparative Qualitätsanalyse zwischen den Daten des OSM- und Wikidata-Projektes wurde für diese Untersuchung auf die DACH-Region festgelegt. Die DACH-Region mit Deutschland, Österreich und der Schweiz zählt durch ihre im Vergleich zu anderen Ländern und Regionen hohen Anzahl an aktiven Mitgliedern zu den Gebieten mit einer der höchsten aktiven Mitglieder pro km^2 weltweit. Um die Daten aus beiden Projekten für den Vergleich entsprechend aufzubereiten, musste als erster Schritt jeweils ein Export der benötigten Daten aus beiden Projekten durchgeführt werden.

3.1 Wikidata im OSM-Projekt

Im OSM-Projekt werden korrespondierende Verweise auf Wikidata über das Tag (Attribut) „wikidata“ gepflegt. Der Wert des Tags enthält dabei die Identifikationsnummer eines Objektes in Wikidata. Diese Identifikationsnummer, die sogenannte Wikidata-ID, besteht aus dem Großbuchstaben Q und einer positiven ganzen Zahl ohne führende Nullen, zum Beispiel für Salzburg „Q34713“ (https://www.wikidata.org/entity/Q34713). Laut Tag-Info (https://

taginfo.openstreetmap.org) (Stand 02.01.2019) gibt es weltweit 1.232.324 mit Wikidata annotierte OSM-Objekte. Insgesamt besitzen diese 1.015.870 verschiedene Werte bzw. unterschiedliche Wikidata-IDs. Um alle OSM-Objekte (Nodes, Ways und Relations) mit dem beschriebenen Wikidata-Tag zu exportieren, wurde die Overpass-API (https://overpass-turbo.eu) verwendet. Auf Basis der Länder-Relationen der drei Länder wurden die IDs der Objekte und deren Koordinatenpaar (im Falle einer Way- oder Relation-Entität der Mittelpunkt) exportiert. Anschließend wurde mittels der Listen mit den jeweiligen IDs der OSM-Wikidata-Objekt und dem Datenbank-Dump mit der kompletten Historie des OSM-Projekts der für den Vergleich verwendete Datensatz erstellt. Insbesondere durch den letzten Schritt sind erst detaillierte Aussagen darüber möglich, wann und von wem ein OSM-Wikidata-Objekt erstellt wurde. Abbildung 1 zeigt zwei Diagramme für die drei Länder. Dabei ist die zeitliche Entwicklung (links) von Wikidata und die Verteilung auf die verschiedenen OSM-Objekte (rechts) im OSM-Projekt zu sehen.

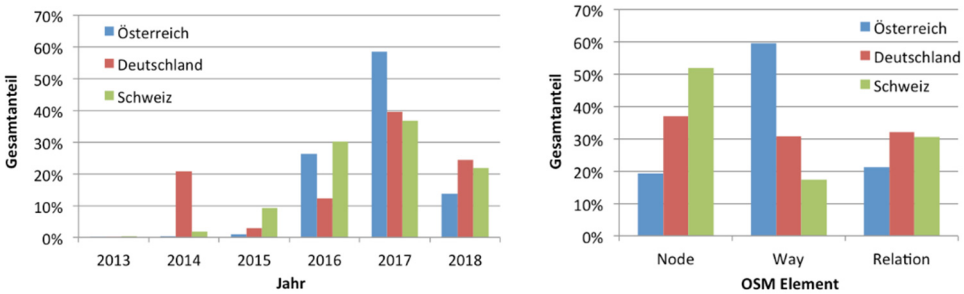

Abb. 1: Zuwachs des Anteils an Elementen mit einem Wikidata-Tag im OSM-Projekt (links) und Verteilung von Wikidata auf die drei OSM-Elemente (rechts)

In der zeitlichen Entwicklung (links) ist vor allem sichtbar, dass in allen drei Ländern der im Schnitt größte Anteil am Wikidata-Tag an verschiedenen Objekten im Jahr 2017 angelegt wurde. Im direkten Vergleich bezüglich der Verteilung (rechts), an welcher Art das Element (Node, Way oder Relation) das Wikidata-Tag vorhanden ist, wird ersichtlich, dass lediglich in Deutschland eine einigermaßen gleiche Verteilung über die drei Elemente vorhanden ist. In Österreich gibt es verhältnismäßig viele Ways und in der Schweiz verhältnismäßig viele Nodes mit einer Wikidata-ID.

Im direkten Vergleich der drei Länder zueinander haben die OSM-Elemente mit einem Wikidata-Tag in Österreich zu 95 %, in Deutschland zu 80 % und in der Schweiz zu 68 % auch ein Wikipedia-Tag. Weiterhin ist es interessant, dass in Österreich bei 27 % der Elemente weitere Attribute für Eisenbahnen-Informationen auftreten. In Deutschland treten dagegen im Vergleich viele Wikidata-Elemente in Kombination von Administrativen-Grenzen auf. In der Schweiz beinhalten viele Wikidata-Elemente zusätzlich einen ele-Key, der für die Angabe einer Höhe über den Meeresgrund verwendet wird, beispielsweise für Berggipfel. Tabelle 1 enthält die Top-5-Keys pro Land, mit welchen das Wikidata-Tag an erfassten Elementen in OSM für den Untersuchungsgebieten vorkommt.

Tabelle 1: Kombinationen von Wikidata-Elementen mit anderen OSM-Tags

	Österreich Key (Häufigkeit)	**Deutschland** Key (Häufigkeit)	**Schweiz** Key (Häufigkeit)
1.	wikipedia (94 %)	name (95 %)	name (97 %)
2.	name (94 %)	wikipedia (80 %)	wikipedia (68 %)
3.	operator (29 %)	type (32 %)	ele (39 %)
4.	railway (27 %)	boundary (25 %)	source (31 %)
5.	ref (27 %)	admin_level (23 %)	type (30 %)

3.2 Wikidata-Export und Statistiken

Nachdem eine Extrahierung der relevanten OSM-Objekte erfolgt war, wurde die Mediawiki-API (Ferschke 2011) verwendet, um die verlinkten Wikidata-Objekte sowie deren Metadaten (Revisionen und Benutzerinformationen) anhand ihrer Wikdata-IDs zu speichern. Hierbei ergaben sich 61.056 Wikidata-Objekte für Deutschland, 9.723 Wikidata-Objekte für Österreich und 11.655 Wikidata-Objekte für die Schweiz. Von diesen Objekten hatten 62 % (Schweiz), 93 % (Österreich) und 86 % (Deutschland) keine geographische Position annotiert. Mit zusammengenommen ca. 10 % waren sowohl in Deutschland als auch in Österreich (6 %) und der Schweiz (30 %) Gemeinderelationen am meisten mit Wikidata-Tags versehen. Eine Annotation des Landes in Wikidata konnte in ca. 14 % der Fälle in Deutschland, 8 % in Österreich und zu 37 % in der Schweiz festgestellt werden. Anhand der Property LocatedIn, die üblicherweise den Landkreis oder die Stadt angeben, kann gezeigt werden, dass in der Schweiz 55 %, in Deutschland 16 % und in Österreich 8 % der Objekte näher semantisch verortet wurden. Wikidata-Objekte weisen weitaus mehr Annotationen als OSM-Objekte auf. Dies ist zum einen damit verbunden, dass Äquivalenzen zwischen OSM-Tags sowie Keys und OSM-Instanzen/Klassen aktuell in Wikidata mit der Beziehung (OSM-Tag oder -Key) abgebildet werden. Durch das SPARQL (Harris 2013) Query in Listing 1 konnten die Äquivalenzen in JSON exportiert und als Basis für den Vergleich verwenden werden. Der JSON-Export enthielt dabei 1798 Key- und Tag-Äquivalenzen. Die Äquivalenzen bieten die Basis für den Vergleich zwischen äquivalenten Keys/Tags, um bei den Keys jeweils einen Vergleich der Attribute vornehmen können.

Listing 1: SPARQL-Query zur Ermittlung von OSM/Wikidata-Äquivalenzen

```
PREFIX bd: <http://www.bigdata.com/rdf#>
PREFIX wikibase: <http://wikiba.se/ontology#>
PREFIX wdt: <http://www.wikidata.org/prop/direct/>
SELECT DISTINCT ?wdclass ?wdclassLabel ?osm
WHERE {
 ?wdclass wdt:P1282 ?osm . SERVICE wikibase:label { bd:serviceParam wikibase:language "en" }
} ORDER BY ?wdclassLabel
```

4 Ergebnisse der Vergleichsanalyse

Wie bereits in der Einleitung beschrieben, wird mit diesem Beitrag insbesondere eine Vergleichsanalyse zwischen Wikidata und OSM durchgeführt. Dabei soll vorrangig keine Aussage getroffen werden, welche Datenquelle „qualitativ" besser geeignet ist, sondern wo evtl. Unterschiede in den Daten zu finden sind. In diesem Abschnitt sind dementsprechend verschiedene Metriken beschrieben, die OSM-Tags mit Wikidata-Informationen vergleichen. Tabelle 2 stellt die Anzahl der Attribute in Wikidata der Anzahl der Attribute in OSM gegenüber. Zudem zeigt die Tabelle die Anzahl verlinkter Attribute sowie die Anzahl der gematchten Tags, bei denen sowohl ein äquivalentes Attribut als auch ein gleicher bzw. äquivalenter Wert vorliegt.

Tabelle 2: Attributstatistiken in den Untersuchungsgebieten (Häufigkeiten gerundet in %)

Gebiet	Anzahl der Attribute (Wikidata)	Anzahl der Attribute (OSM)	Anzahl verlinkter Attribute	Anzahl gematchter Tags
Deutschland	0-10 (4 %) 11-20 (40 %) 21-30 (23 %) ≥ 31 (33 %)	0-5 (21 %) 6-10 (47 %) 11-15 (22 %) ≥ 16 (10 %)	0 (83 %) 1 (12 %) 2 (3 %) ≥ 3 (1 %)	0 (67 %) 1 (20 %) 2 (5 %) ≥ 3 (6 %)
Österreich	0-10 (3 %) 11-20 (56 %) 21-30 (18 %) ≥ 31 (23 %)	0-5 (14 %) 6-10 (42 %) 11-15 (16 %) ≥ 16 (28 %)	0 (67 %) 1 (30 %) 2 (2 %) ≥ 3(1 %)	0 (47 %) 1 (25 %) 2 (6 %) ≥ 3 (22 %)
Schweiz	0-10 (3 %) 11-20 (45 %) 21-30 (22 %) ≥ 31(30 %)	0-5 (21 %) 6-10 (35 %) 11-15 (32 %) ≥ 16 (12 %)	0 (52 %) 1 (24 %) 2 (11 %) ≥ 3 (14 %)	0 (64 %) 1 (21 %) 2 (8 %) ≥ 3 (5 %)

4.1 Positionsgenauigkeit

Um Unterschiede in der geographischen Position zwischen den beiden Projekten zu untersuchen, wurden jeweils die Mittelpunkte (Zentroid) der Geometrien in OSM mit der Koordinate im Wikidata Attribut P625 (coordinate location) verglichen. Dabei kann zwar keine Aussage über die Qualität der Positionsangabe getroffen werden, trotzdem gibt es Auskunft darüber, in welchem Rahmen die Daten zueinander passen. In Wikidata gibt es zwar weitere Attribute, wie z. B. geoshape, um Geometrien anzugeben; zum Zeitpunkt dieser Untersuchung war allerdings lediglich das Attribut P625 (coordinate location) nennenswert annotiert und kann daher nur als Referenz für einen geometrischen Vergleich verwendet werden. Abbildung 2 zeigt die Distanzen zwischen den jeweiligen Koordinaten in Deutschland, Österreich und der Schweiz. Hierbei muss bedacht werden, dass der Mittelpunkt nur bei Punktkoordinaten eine gute Vergleichsbasis bietet. Speziell bei LineStrings (Ways) muss das Ergebnis daher mit einer gewissen Vorsicht betrachtet werden, weil jeweils von OSM-Seite der Mittelpunkt der Linie mit der Punktkoordinate aus Wikidata verglichen wurde. Ebenfalls kommt es durchaus häufig vor, dass in Wikidata keine P625 (coordinate location) Annotation vorhanden ist. Die Ergebnisse beziehen sich somit nur auf die Fälle in denen die entsprechende Annotation vorlag.

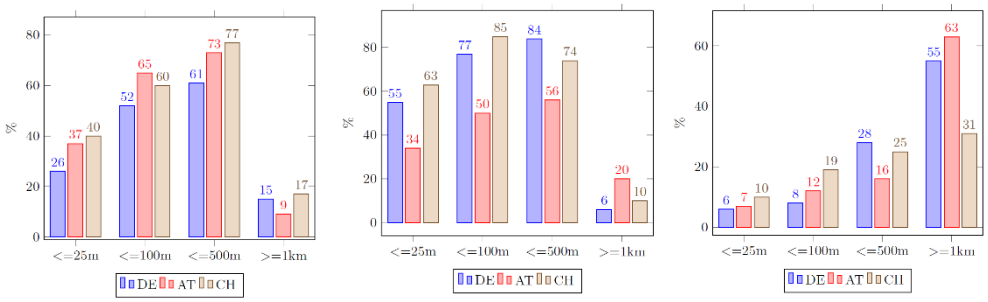

Abb. 2: Objektdistanzen [m] aufgebrochen nach Node, Way und Relation, (Häufigkeiten gerundet in %) von links nach rechts (Nodedistanzen, Waydistanzen, Relationdistanzen)

Hierbei wiesen 13 % (10219) der WD in Deutschland, 5287 (30 %) der WD-Objekte in Österreich und 633 (5 %) der WD-Objekte in der Schweiz keine Koordinate auf.

Als Ergebnis lässt sich feststellen, dass ca. $^1/_3$ der OSM-Nodes in einer Distanz von 25 m zur Wikidata-Koordinate stehen. Abgesehen von Deutschland hat die Hälfte der Nodes eine Nodedistanz von < 100 m. Ähnliche Ergebnisse, wenn auch etwas schlechter aufgrund der Centerpointanalyse, weisen die Waydistanzen auf. Für die Relationdistanzen lässt sich ein schlechteres Ergebnis festhalten. Centerpoints von Relationen haben naturgemäß eine größere Distanz zu den jeweiligen Wikidata-Koordinaten, da sich diese über eine größere Fläche erstrecken.

4.2 Logische Konsistenz: Objektäquivalenz

Neben der Möglichkeit, Wikidata-Konzepte in OSM zu annotieren, existiert auch die Möglichkeit in Wikidata, OSM-Relationen mit P402 (OSM Relation ID) zu annotieren. Wikidata verlinkt nur zu Relationen, da die Wikidata Community Node und Way-IDs als zu instabil ansieht, um diese direkt zu verlinken. Die Löschung eines OSM-Elements und sein erneutes Anlegen (z. B. im Falle einer erneuten Erfassung von einem Satellitenbild oder GPS) erzwingt die Nutzung einer neuen ID, sodass ein Objekt potenziell jederzeit umbenannt und nicht mehr aufgelöst werden kann. Im Gegensatz hierzu existieren Vorschläge in der OSM-Community Wikidata-IDs als stabile IDs für die Annotation von Objekten zu nutzen. Tabelle 3 zeigt, in wie vielen Fällen eine in Wikidata annotierte OSM-Relation mit den Wikidata-Tags in OSM übereinstimmt.

Tabelle 3: Statistiken über verlinkte OSM-Relations

Gebiet	Relation IDs	korrekte IDs	falsche IDs	nicht verlinkte Relations	Anzahl Relations
Deutschland	22249 (28 %)	14647 (65 %)	702 (3 %)	9465 (37 %)	24.924
Österreich	275 (7 %)	177 (64 %)	46 (16 %)	3487 (93 %)	3.710
Schweiz	4334 (37 %)	2637 (73 %)	110 (3 %)	824 (23 %)	3.571

Die Auswertung der Wikidata-Verlinkungen in Tabelle 3 zeigt, dass in der Schweiz die meisten Wikidata-Relationen verlinkt sind. Ebenfalls weist die Schweiz die höchste Korrektheit der Verlinkungen mit 73 % auf. In Österreich sind verhältnismäßig wenige OSM-Relation-IDs in Wikidata vorhanden und dementsprechend wenige OSM-Relationen zu Wikidata verlinkt.

4.3 Thematische Genauigkeit: Äquivalenz des Label- und Name-Attributes

In Wikidata werden Labels mit der rdfs:label-Beziehung beschrieben. Labels dienen in Wikidata der Beschreibung von Konzepten in verschiedenen Sprachen, der Definition von Synonymen und werden üblicherweise für die Suche nach Konzepten in Suchmaschinen verwendet. Labels werden des Weiteren mit einer Annotation im Value versehen, um die Sprache zu kennzeichnen (z. B. Berlin@de für Deutsch). In OSM ist diese Annotation am jeweiligen Key des Tags (name:de,name:en) annotiert. Jedoch gibt es auch den name-Key, der das Label üblicherweise in der jeweiligen Landessprache beschreibt. Die Landessprache selbst ist jedoch nicht immer eindeutig definiert. So finden sich in Hongkong überwiegend name-Tags mit Englisch sowie Chinesisch und in dem Anwendungsfall in der Schweiz kann ebenfalls das name-Tag je nach Schweizer Region in Deutsch, Französisch, Italienisch oder Rätoromanisch erwartet werden. Ein Labelvergleich muss deshalb auf den Labels der gleichen Sprache, also name:de sowie rdfs:label „Berlin"@de erfolgen, und alle Labels müssen mit dem generalisierten name-Tag verglichen werden, um Gleichheit bzw. Äquivalenzen aufzuzeigen. Neben der Gleichheit wurde eine Substring-Metrik und eine Metrik der Editierdistanz (Wagner 1974) angewendet, um die Ähnlichkeit der jeweiligen Labels aufzuzeigen. Die Tabelle 4 zeigt die Labelstatistiken kumuliert für das jeweilige name- bzw. name:locale-Tag.

Tabelle 4: Verhältnis der Labeläquivalenzen in Wikidata und OSM für Labels der jeweiligen Landessprache (name bzw. name:de, name:fr, name:it)

Gebiet	Gleiche Labels	Substring-Matches	Editierdistanzen
Deutschland	6623 (OSM 10 %) (WD 16 %)	0.2 %	0 (58 %), <=1 (72 %)
Österreich	820 (OSM 6 %) (WD 5 %)	0.1 %	0 (52 %), <=1 (65 %)
Schweiz	2872 (OSM 33 %) (WD 36 %)	0.7 %	0 (66 %), <=1 (77 %)

Bei diesem Vergleich zeigte sich, dass in Wikidata im Schnitt mehr Labels als in OSM existieren. Wenn Labels gematcht werden können, haben sie im Schnitt eine hohe Übereinstimmung, da die Hälfte aller Matchings in der vorherrschenden Sprache funktioniert. Anschließend weist ca. $^1/_4$ der Nicht-Matches eine Editierdistanz von 1 auf. Dies könnte auf eine andere Buchstabierung oder einen Tippfehler hindeuten. Ca. $^1/_3$ der Labels haben jedoch eine höhere Editierdistanz. Dies weist auf Unstimmigkeiten in der Bezeichnung des Objektes hin. Denkbar ist hier die Verwendung von Abkürzungen anstatt einer ausgeschriebenen Bezeichnung oder eine komplett andere Bezeichnung für das jeweilige Objekt. Der relativ geringe Anteil von Substring-Matches zeigt, dass voneinander verschiedene Labels üblicherweise keine Bestandteile voneinander sind. Somit sind Fälle wie „Vienna – Wien", d. h. Dopplungen von Namensbeschreibungen in einem Label bzw. Schreibvarianten mit dem gleichen Bestandteil, weniger häufig anzutreffen.

Tabelle 5: Labelstatistiken für Wikidata und OSM (Häufigkeiten gerundet in %)

Gebiet	Anzahl Labels Wikidata	Anzahl Labels OSM	Anzahl fehlender Labels (OSM)	Anzahl fehlender Labels (Wikidata)
Deutschland	0 (30 %) 1 (57 %) 2 (1 %) ≥ 3 (12 %)	0 (4 %) 1 (74 %) 2 (14 %) ≥ 3 (8 %)	0 (6 %) 1 (74 %) 2 (13 %) ≥ 3 (7 %)	0 (29 %) 1 (57 %) 2 (1 %) ≥ 3 (13 %)
Österreich	0 (33 %) 1 (56 %) 2 (1 %) ≥ 3 (11 %)	0 (6 %) 1 (66 %) 2 (22 %) ≥ 3 (6 %)	0 (6 %) 1 (66 %) 2 (22 %) ≥ 3 (6 %)	0 (38 %) 1 (52 %) 2 (1 %) ≥ 3 (10 %)
Schweiz	0 (18 %) 1 (52 %) 2 (1 %) ≥ 3 (29 %)	0 (1 %) 1 (66 %) 2 (18 %) ≥ 3 (15 %)	0 (1 %) 1 (70 %) 2 (17 %) ≥ 3 (12 %)	0 (17 %) 1 (44 %) 2 (1 %) ≥ 3 (38 %)

In dieser Auswertung zeigt sich, dass die Mehrheit der OSM-Objekte mit mindestens einem Label beschriftet ist. Bei Wikidata hingegen sind öfter mehr Labels als in OSM annotiert. Dies wird besonders in der Auswertung der Schweiz deutlich, in der 29 % der Wikidata-Objekte mehr als drei Label aufweisen. Während der Untersuchung sind bei der Annotation von Labels folgende Fehler aufgefallen:

- Ortsnamen sind unklar/mehrfach belegt: way/93777199 ist in OSM als Fischbach annotiert, in Wikidata als Raperswilen.
- Abkürzungen werden in OSM als name-Tag eingetragen, jedoch die ausgeschriebene Variante in Wikidata als Label (Wikidata pflegt Abkürzungen in anderen Annotationen); Beispiel: way/255675972 und Q675983.

Beide Beispiele zeigen, dass in Wikidata üblicherweise mehrere Varianten des gleichen Namens teilweise unter unterschiedlichen Annotationen verfügbar sind. In OSM ist üblicherweise nur einer dieser Namen im name-Tag annotiert. Zwar existieren alternative Namenstags in OSM, jedoch sind diese zumindest in diesem Versuchsset üblicherweise nicht vorhanden oder unterrepräsentiert.

4.4 Vollständigkeit: Adressangaben

Eine Vollständigkeit von Attributen für OSM respektive Wikidata zu definieren, ist ein schwieriger Prozess, da in beiden Communities keine Richtlinien dafür existieren. Wikidata erlaubt das Interlinken von beliebigen Attributen zu beliebigen Klassen und Instanzen, sodass einzig häufig mit der jeweiligen Instanz verlinkten Attribute als Vollständigkeit angesehen werden können. Im OSM-Projekt existieren zwar Empfehlungen, jedoch ist die Umsetzung dieser Empfehlungen sehr lückenhaft und nicht bindend. Dennoch kann man für geometrische Objekte im Allgemeinen eine Adresse erwarten und die Elemente einer Ortsbeschreibung in OSM sowie in Wikidata prüfen. Tabelle 6 zeigt äquivalente Adress-Keys mit ihren Wikidata-Property-Gegenstücken, ermittelt durch das Query in Abschnitt 3.2.

Tabelle 6: Äquivalente Adress-Keys (OSM) respektive Wikidata-Properties für die Ermittlung der Adressvollständigkeit

OSM-Keys	Wikidata-Property
addr:city	P131
addr:country	P17
addr:housenumber	P670
addr:postcode	P281
addr:street	P969

Für die Country- sowie LocatedIn-Annotation der Wikidata-Objekte wurde analysiert, wie viele mit diesen annotiert sind. In Deutschland sind dies bei beiden Attributen 89 %. In Österreich hatten 92 % der WD-Objekte die Annotation country und 60 % die Annotation LocatedIn. In der Schweiz waren 94 % der WD-Objekte mit country und 98 % mit LocatedIn annotiert.

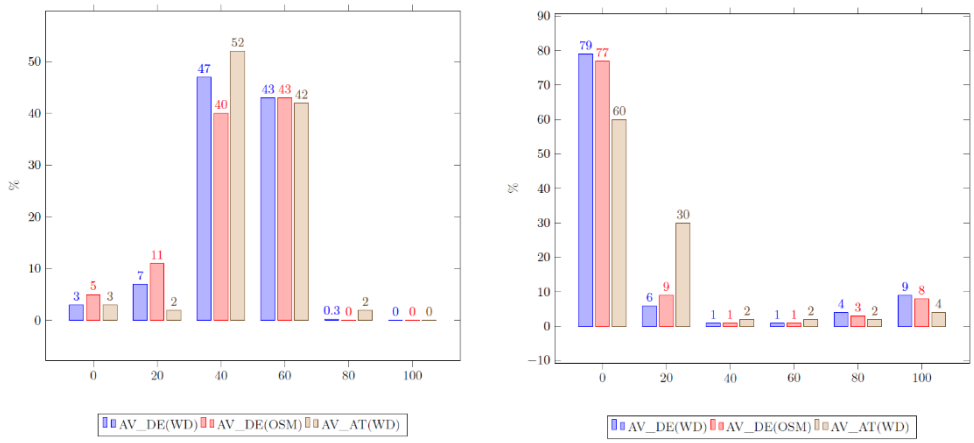

Abb. 3: Adressvollständigkeit in OSM und Wikidata (Häufigkeiten gerundet in %)

Abbildungen 3 zeigt, dass in OSM ein Großteil der untersuchten Objekte nicht mit vollständigen Adressdaten versehen ist. In Wikidata hingegen ist im Allgemeinen eine höhere Annotationsdichte an den gegebenen Adressdaten. Eine Kombination der in Wikidata annotierten Adressdaten mit OSM kann also durchaus zu einer Vollständigkeit auch der OSM-Daten beitragen. Auffällig ist in Wikidata die außergewöhnlich hohe Annotationsdichte von Country- und LocatedIn-Annotationen, welche beispielsweise in OSM für die Zugehörigkeitsbeziehungen zu Gemeinden genutzt werden könnten.

5 Diskussion der Ergebnisse

Beim Vergleich der beiden Datensätze ist für den Endanwender natürlich potenziell interessant, welche Arten von Informationen in Wikidata respektive OSM nicht vorkommen. Durch die Zusammenführung oder gemeinsame Darstellung der Daten kann ein möglicher Gewinn besser abgeschätzt werden. Mit diesem Beitrag wurde gezeigt, dass in Wikidata bedeutend mehr Labels zur Verfügung stehen, sodass ein idealer Anwendungsfall die Erstellung von mehrsprachigen Karten mit Wikidata als Wissensbasis unterstützt werden könnten. Wikidata kann somit also zu einer Ergänzung möglicher fehlender OSM-Tags beitragen. Gleichermaßen können die höhere Adressvollständigkeit in OSM als Anreiz dienen, diese in Wikidata zu überführen. Neben diesem ergänzenden Effekt einer Kombination von beiden Ressourcen sind jedoch auch einige typische Fehler bei der Annotation von Wikidata-IDs in OSM aufgefallen:

- Nicht ausdifferenzierte Objektbeschreibung: Das Konzerthaus Mozarteum in Salzburg besitzt einen Ableger in Innsbruck node/5010318287. In Innsbruck wird jedoch auf das Q871369 (Mozarteum in Salzburg) verwiesen.
- Zeitbedingte Aussagen: OSM unterstützt im Gegensatz zu Wikidata konzeptuell keine Zeitdimension. Dies hat zur Folge, dass in Wikidata bspw. Konzentrationslager mit dem Country Tag Nazi Germany gekennzeichnet sind, die es in OSM aufgrund der Aktualität nicht geben kann.
- Mehrfaches Tagging von untergeordneten Objekten: Bei Untergebäuden einer Gebäudeeinheit kommt es öfters vor, dass alle mit der gleichen Wikidata-ID ausgezeichnet werden. Dies ist sowohl semantisch als auch nach OSM-Standards falsch. Dieses Problem tritt ebenfalls bei Straßen auf. Die Augustinerstraße in Mainz besteht aus den zwei Segmenten way/179133904 und way/15971420. Eine Straßenrelation existiert nicht und der Wikidata-Tag mit Wert Q14531614 (Augustinerstraße) ist an beiden Straßensegmenten annotiert.
- Tagging von Hauptsitzen von Unternehmen: Das Unternehmen Bauhaus betreibt verschiedene Märkte in Deutschland, Österreich und der Schweiz. Der Bauhaus-Baumarkt way/85521406 in Linz/Österreich ist hierbei mit dem Wikidata-Tag Q672043 (Bauhaus) annotiert, welcher als P17 (country) Annotation jedoch die Q39 (Schweiz) angibt (Hauptsitz des Unternehmens Bauhaus)
- Denkmäler als Events der Zeitgeschichte: Das Denkmal der Flugkatastrophe von Hochwald node/2632327301 ist mit dem Event der Flugkatastrophe (Q1671791) in Wikidata annotiert.

Ein weiterer interessanter Aspekt für Anwender ist die Anzahl an Konflikten bzw. Übereinstimmungen von äquivalenten Wikidata-Tags. Konflikte von Values zwischen äquivalenten Wikidata- und OSM-Attributen waren in unserem Versuchsset meistens auf andere Schreibweisen beispielsweise der Bezeichner der jeweiligen Objekte zurückzuführen.

Für eine Demonstration der Unterschiede, wie im letzten Abschnitt erläutert, ist eine Web-Anwendung erstellt worden, die Wikidata- und OSM-Objekte anzeigt, vergleicht und Unterschiede farbig hervorhebt. Im Hintergrund wird dabei eine Kombination aus Overpass-API und Wikidata-Queries verwendet, um die verlinkten Objekte anzuzeigen und mittels eines Popups zu vergleichen (http://i3mainz.github.io/semgistestbench/wikidataview.html).

6 Zusammenfassung und Ausblick

Die Ergebnisse dieser Studie zeigen die Unterschiede der verlinkten Wikidata- und OSM-Objekte in Deutschland, Österreich und der Schweiz. Zusammenfassend kann festgehalten werden, dass zumindest im Untersuchungsgebiet im Wikidata-Projekt bedeutend mehr Labels zur Verfügung stehen als im OSM-Projekt. Würden die Daten des OSM-Projekts mit den Labels aus Wikidata angereichert, wäre eine vollständigere Erstellung von mehrsprachigen Karten möglich. Weiterhin konnte gezeigt werden, dass sich Wikidata und OSM im Bereich der Adressannotationen zu einem gewissen Grad ergänzen können. Dies lässt eine Kombination von Wikidata- und OSM-Daten auch in diesem Aspekt sinnvoll erscheinen. In Bezug auf die Qualität der Annotationen konnte jedoch auch an verschiedensten Beispielen die Probleme und Missverständnisse der OSM-User in Bezug auf Wikidata dargestellt werden. Annotationen, die nicht das zu annotierende Objekt betrachten, sondern eine übergeordnete Entität (wie z. B. die Zentrale eines Baumarktes anstatt die Filiale), sind mögliche typische semantische Fehlerquellen, die durch fehlerbehandelnde Tools in Zukunft evtl. vermieden werden könnten. Hier besteht noch großes Potenzial für die Entwicklung von ähnlichen Anwendungen, wie z. B. Osmose(http://osmose.openstreetmap.fr/de/map/) für die semantische Validierung von annotierten Wikidata-Objekten. Neben einer noch nicht überall hinreichend gegebenen (bei Ways und Relations jedoch auch nur angenäherten) Positionsgenauigkeit und nur einem gewissen Grad von noch nicht äquivalenten Key-Value-Kombinationen gibt es also noch genug Verbesserungspotenzial bei der Annotation.

In der Zukunft sollen weitere Analysen durchgeführt werden, um zusätzliche Informationen über die verwendeten Tools und die Community, die Wikidata in ihrem und im OSM-Projekt erfasst, zu erhalten. Dadurch werden weitere Erkenntnisse gewonnen, ob tendenziell eine automatisierte oder manuelle Verlinkung der Daten Realität ist. Aufgrund mancher Ergebnisse und gefundener Fehler dieser Studie könnten einige Informationen nicht von Mitgliedern händisch hinzugefügt worden sein. Weiterhin soll das Untersuchungsgebiet auf weitere Länder ausgebreitet werden um, ähnlich wie für das OSM-Projekt, gültige Aussagen über die Quantität und Qualität treffen zu können. Gerade beim OSM-Projekt kann die Datendichte von Land zu Land relativ unterschiedlich sein. Ein weiterer zu untersuchender Aspekt könnten Tag-Varianten wie brand:wikidata, operator:wikidata oder artist:wikidata sein, die in dieser Publikation nicht betrachtet wurden. Laut Tag-Info gibt es weltweit mehr als 100.000 solcher weiterer Wikidata-Annotationen, also genug, um eine Analyse der Korrektheit auch dieser Verlinkungen durchzuführen.

Literatur

Almeida, P. D., Rocha, J. G., Ballatore, A., & Zipf, A. (2016). Where the streets have known names. *International Conference on Computational Science and Its Applications* (pp. 1–12). Springer.

Auer, S., Lehmann, J., & Hellmann, S. (2009). Linkedgeodata: Adding a spatial dimension to the web of data. *International Semantic Web Conference* (pp. 731–746). Berlin/Heidelberg: Springer.

Barron, C., Neis, P., & Zipf, A. (2013). Towards intrinsic quality analysis of openstreetmap datasets. *Online proceedings of the international workshop on action and interaction in volunteered geographic information (ACTIVITY)*. AGILE.

Barron, C., Neis, P., & Zipf, A. (2014). A comprehensive framework for intrinsic openstreetmap quality analysis. *Transactions in GIS, 18*(6), 877–895.

Basiri, A., Jackson, M., Amirian, P., Pourabdollah, A., Sester, M., Winstanley, A., Moore, T., & Zhang, L. (2016). Quality assessment of openstreetmap data using trajectory mining. *Geo-spatial information science*, *19*(1), 56–68.

Brassel, K., Bucher, F., Stephan, E. M., & Vckovski, A. (2015). Completeness. *Elements of spatial data quality* (pp. 81–108). Elsevier.

Chaudhuri, G., & Clarke, K. C. (2014). Temporal accuracy in urban growth forecasting: A study using the sleuth model. *Transactions in GIS*, *18*(2), 302–320.

Devillers, R., Bédard, Y., & Jeansoulin, R. (2005). Multidimensional management of geospatial data quality information for its dynamic use within GIS. *Photogrammetric Engineering & Remote Sensing*, *71*(2), 205–215.

Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D. (2014). Introducing wikidata to the linked data web. *International Semantic Web Conference* (pp. 50–65). Springer.

Fan, H., Zipf, A., Fu, Q., & Neis, P. (2014). Quality assessment for building footprints data on openstreetmap. *International Journal of Geographical Information Science, 28*(4), 700–719.

Ferschke, O., Zesch, T., & Gurevych, I. (2011). Wikipedia revision toolkit: efficiently accessing wikipedia's edit history. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations* (pp. 97–102). Association for Computational Linguistics.

Goodchild, M. F. (2007). Citizens as sensors: the world of volunteered geography. *Geo-Journal*, *69* (4), 211–221.

Goodchild, M. F., & Li, L. (2012). Assuring the quality of volunteered geographic information. *Spatial statistics*, *1*, 110–120.

Goodchild, M. F., & Hunter, G. J. (1997). A simple positional accuracy measure for linear features. *International Journal of Geographical Information Science*, *11*(3), 299–306.

Gruenreich, D. (1992). Atkis-a topographic information system as a basis for GIS and digital cartography in germany. From Digital Map Series to Geo-Information Systems. *Geologisches Jahrbuch, Series A*. Hannover, Germany: Federal Institute of Geosciences and Resources.

Haklay, M., & Weber, P. (2008). Openstreetmap: User-generated street maps. *IEEE Pervasive Computing*, *7*(4), 12–18.

Harris, S., Seaborne, A., & Prud'hommeaux, E. (2013). Sparql 1.1 query language. *W3C recommendation*, *21*(10).

Homburg, T., & Boochs, F. (2019). Situation-dependent data quality analysis for geospatial data using semantic technologies. In: W. Abramowicz, & A. Paschke (Eds.), *Business Information Systems Workshops* (566–578). Cham: Springer International Publishing.

ISO 19113 (2002). *Geographic information – quality principles*. Genf: International Organization for Standardization.

Kainz, W. (1995). Logical consistency. *Elements of spatial data quality, 202*, 109–137.

Leyh, W., & Fonseca Filho, H. (2017). Interlinking standardized openstreetmap data and citizen science data in the opendata cloud. *International Conference on Applied Human Factors and Ergonomics* (pp. 85–96). Springer.

Majic, I., Winter, S., & Tomko, M. (2017). Finding equivalent keys in openstreetmap: semantic similarity computation based on extensional definitions. *Proceedings of the 1st*

Workshop on Artificial Intelligence and Deep Learning for Geographic Knowledge Discovery (24–32). ACM.

Masek, W. J., & Paterson, M. S. (1980). A faster algorithm computing string edit distances. *J. Comput. Syst. Sci.*, *20*(1), 18–31.

Mocnik, F. B., Mobasheri, A., Griesbaum, L., Eckle, M., Jacobs, C., & Klonner, C. (2018). A grounding-based ontology of data quality measures. *Journal of Spatial Information Science*, (16), 1–25.

Mooney, P., Corcoran, P., & Winstanley, A. C. (2010). Towards quality metrics for openstreetmap. *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems* (pp. 514–517). ACM.

Neis, P., Zielstra, D., Zipf, A., & Strunck, A. (2010). Empirische Untersuchungen zur Datenqualität von Openstreetmap – Erfahrungen aus zwei Jahren betrieb mehrerer Online-Dienste. In: J. Strobl et al. (Eds.), *Angewandte Geoinformatik 2010.* Berlin/Offenbach: Wichmann.

Neis, P., & Zipf, A. (2012). Analyzing the contributor activity of a volunteered geographic information project – the case of openstreetmap. *ISPRS International Journal of Geo-Information*, *1*(2), 146–165.

Neis, P., Zielstra, D., & Zipf, A. (2012). The street network evolution of crowdsourced maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, *4*(1), 1-21.

Neis & Zielstra (2014). Recent Developments and Future Trends in Volunteered Geographic Information Research – The Case of OpenStreetMap. *Future Internet 2014, 6*(1), 76–106. https://doi.org/10.3390/fi6010076.

Ramm, F., Topf, J., & Chilton, S. (2011). *OpenStreetMap: using and enhancing the free map of the world*. Cambridge: UIT Cambridge.

Salgé, F. (1995). *Semantic accuracy. Elements of spatial data quality* (pp. 139–151).

Senaratne, H., Mobasheri, A., Ali, A. L., Capineri, C., & Haklay, M. (2017). A review of volunteered geographic information quality assessment methods. *International Journal of Geographical Information Science*, *31*(1), 139–167.

Schoof, M., Behncke, K., & Ehlers, M. (2012). *Atkis-basis-dlm und openstreetmap – ein datenvergleich anhand ausgewählter Gebiete in niedersachsen. Untersuchung der Nutzung von OpenStreetMap Daten zur Darstellung von TMC Verkehrsmeldeinformation.*

Vrandečić, D. (2012). Wikidata: A new platform for collaborative data collection. *Proceedings of the 21st International Conference on World Wide Web* (pp. 1063–1064). ACM.

Vrandečić, D., & Krötzsch, M. (2014). Wikidata: a free collaborative knowledgebase. *Communications of the ACM, 57*(10), 78–85.

Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM)* 21.1, 168–173.