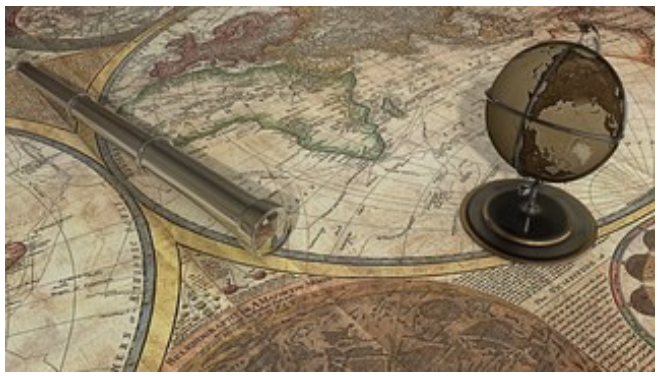
Universität Bielefeld
Institute for Interdisciplinary Studies of Science (I²SoS)

kompetenzzentrum bibliometrie

Geokodierung von Autorenadressen in Publikationsdatenbanken

Abschlussbericht einer Untersuchung
für das Kompetenzzentrum Bibliometrie

Christine Rimmert, Matthias Winterhager



Universität Bielefeld
Institute for Interdisciplinary Studies of Science (I²SoS)

Bielefeld, 31.03.2017

Der vorliegende Bericht dokumentiert die Ergebnisse eines Projekts, das 2016 am I²SoS der Universität Bielefeld im Auftrag des nationalen Kompetenzzentrums Bibliometrie durchgeführt wurde. Das Verfahren wird in Nachfolgeprojekten an der Universität Bielefeld weiterentwickelt.

Inhalt

1. Einleitung

Die in wissenschaftlichen Publikationen enthaltenen Angaben zu den Autorenadressen („Affiliations“) erlauben neben organisationsbezogenen Auswertungen auch solche, die auf Fragestellungen mit geographischen Bezügen zielen. Im Science Citation Index hat diese Tatsache schon früh Ausdruck gefunden, indem der *Corporate Index* jeweils in zwei separaten Teilen ausgeliefert wurde: einer *Organization Section* und einer *Geographic Section*.

Für organisationsbezogene Analysen stehen die Ergebnisse der Institutionenkodierung (vgl. Winterhager et al. 2014) zur Verfügung, die in die Datenbanken des Kompetenzzentrums Bibliometrie eingespeist werden. Und mit der zusätzlich bereitgestellten Länderkodierung[1] gibt es bereits auch eine wichtige Ressource für die Arbeit in Projekten mit Geographiebezug. Damit kann jedoch nur auf der sehr hohen Aggregationsebene der Nationen gearbeitet werden. Unterhalb dieser Ebene werden Analysen schwierig. Hier gibt es aber einen steigenden Bedarf: Raumbezogene bibliometrische Untersuchungen haben in den letzten Jahren unter dem Stichwort *spatial scientometrics* erheblich an Bedeutung gewonnen (vgl. etwa: Frenken et al. 2009, Bornmann & Leydesdorff 2011, Gao 2014, Xuemei et al. 2014, Abramo et al. 2015).

Für eine verlässliche räumliche Zuordnung (Geokodierung) von Publikationen mittels Autorenadressen gibt es eine Reihe von Problemen, die nicht einfach zu lösen sind.[2] Die Probleme sind teilweise analog zu denen der Institutionenkodierung. Die Schreibweisen für die verschiedenen geografischen Entitäten können unterschiedlich sein (spelling variants, englisch/deutsch), Bezeichnungen können uneindeutig sein (Frankfurt), Postleitzahlen können variieren (auch über die Zeit) oder fehlerhaft sein. Hinzu kommen Fragen bei der Aggregation zu größeren räumlichen Einheiten.

Ein naiver Betrachter könnte annehmen, dass sich die Geokodierung automatisch aus der Institutionenkodierung ergibt, nicht zuletzt deshalb, weil bei großen Institutionen der Name öfters eine scheinbar eindeutige Bezeichnung des Standorts beinhaltet. In der Realität zeigen sich jedoch viele Fälle, in denen es zu Abweichungen kommt (Beispiel: die organisatorisch zur Universität Heidelberg gehörende Medizinische Fakultät Mannheim). Namentlich im Bereich der Helmholtz-Gemeinschaft Deutscher Forschungszentren (HGF) gibt es eine Reihe von Institutionen, deren Publikationen ganz unterschiedlichen Standorten zuzuordnen sind. Ein prominentes Beispiel dafür ist das Deutsche Zentrum für Luft- und Raumfahrt (DLR), das zwar seinen Hauptsitz in Köln hat, aber zugleich an vielen anderen Standorten aktiv ist.

Hinzu kommt, dass in raumbezogenen Analysen häufig der organisationale Aspekt der Zuordnung gar nicht relevant ist, sondern stattdessen die rein geographische Lokalisation (und ggf. Aggregation) der Publikationen gefordert ist. Es wäre deshalb von großem Wert, neben der Institutionenkodierung eine eigenständige Geokodierung bereitstellen zu können (einschließlich

1 Vgl. die Dokumentation im Abschnitt „KB-Tabellen zur Länderkodierung“ im Wiki des Kompetenzzentrums Bibliometrie (auf Anfrage erhältlich).

2 Vgl. Frenken et al. 2009, S. 226.

verschiedener Aggregationsmöglichkeiten wie z.B. für Städte, Regionen, Bundesländer). Die vorliegende Studie dient dem Ziel, die Grundlagen dafür zu schaffen.

Dass es für eine valide Geokodierung nicht ausreicht, Adressdatensätze aus Publikationsdatenbanken schlicht als Masseninput in sog. Geotools einzuspeisen und dort verarbeiten zu lassen, ist bekannt (vgl. z.B. Bornmann & Leydesdorff 2011, S.1956). Eigene Vorarbeiten zur Geokodierung haben das bestätigt (Rimmert 2012 und 2013). Da bei der Nutzung von Diensten kommerziell agierender Produzenten wie Yahoo oder Google auch lizenzrechtliche Probleme bestehen, wurde das Projekt von vornherein auf die Untersuchung der Verfügbarkeit und Eignung anderer Ressourcen ausgerichtet. Ziel war die Entwicklung eines Ansatzes, der Anwendungen ohne zusätzliche Lizenzkosten ermöglicht (auf der Basis von open access).

Dazu wurde zunächst eine statistische Auswertung zu Art und Vollständigkeit der in den Adressdatensätzen von Web of Science (WoS)[3] und Scopus[4] enthaltenen (und extrahierbaren) Geoinformationen durchgeführt. Die multidisziplinären Zitationsdatenbanken WoS und Scopus wurden als Ausgangspunkte gewählt, da sie zu den meistgenutzten für bibliometrische Auswertungen zählen.

In einem weiteren Schritt wurden geeignete Quellen (im Sinne von open access) für den Bezug von Daten zu geografischen Daten wie Ortsnamen, Geokoordinaten, Postleitzahlen und Hierarchiebeziehungen zwischen den geografischen Einheiten recherchiert und in Bezug auf ihren Nutzen im spezifischen Kontext der Autorenadressen evaluiert. Darüber hinaus wurden Möglichkeiten für den Bezug weiterer Daten zu geografischen Einheiten aus diesen freien Quellen geprüft (zum Beispiel aus der Wikipedia).

Mit den so gewonnenen Erkenntnissen und Daten wurden Grundzüge einer Methode für die Geokodierung entwickelt sowie eine geeignete Tabellenstruktur entworfen, mit der die durch dieses Verfahren erzielbaren Kodierungen zukünftig ausgeliefert bzw. in die Datenbanken des Kompetenzzentrums Bibliometrie eingespeist werden könnten.

Die Ergebnisse der genannten Schritte werden in den folgenden Abschnitten dargestellt. Zuvor wird noch in einem kurzen Abschnitt erläutert, dass die Verfahren zur Geokodierung und Institutionenkodierung vom Grundsatz her analog angelegt sind.

3 Die für die statistischen Auswertungen zum Web of Science (WoS) genutzten Daten stammen aus dem Science Citation Index Expanded (SCIE), dem Social Sciences Citation Index (SSCI), dem Arts & Humanities Citation Index (AHCI) sowie dem Conference Proceedings Citation Index (CPCI-S und CPCI-SSH), als Rohdaten (in der XML-Version) bereitgestellt durch Thomson Reuters (Scientific) Inc, (TR©), Philadelphia, Pennsylvania, USA: © Copyright Thomson Reuters (Scientific) 2016. Alle Rechte vorbehalten.

4 Die für die statistischen Auswertungen zu Scopus genutzten Daten stammen aus den XML-basierten Rohdaten von Scopus, Copyright © 2016 Elsevier B.V. Alle Rechte vorbehalten.

Geokodierung & Institutionenkodierung

Autorenadressen können einerseits organisationellen Einheiten/Institutionen (hier als Org-Unit bezeichnet) und andererseits auch geografischen Einheiten (hier als Geo-Unit bezeichnet) zugeordnet werden. Abb. 1 zeigt Adressen, Org-Units und Geo-Units als die zentralen Entitäten im Datenschema.

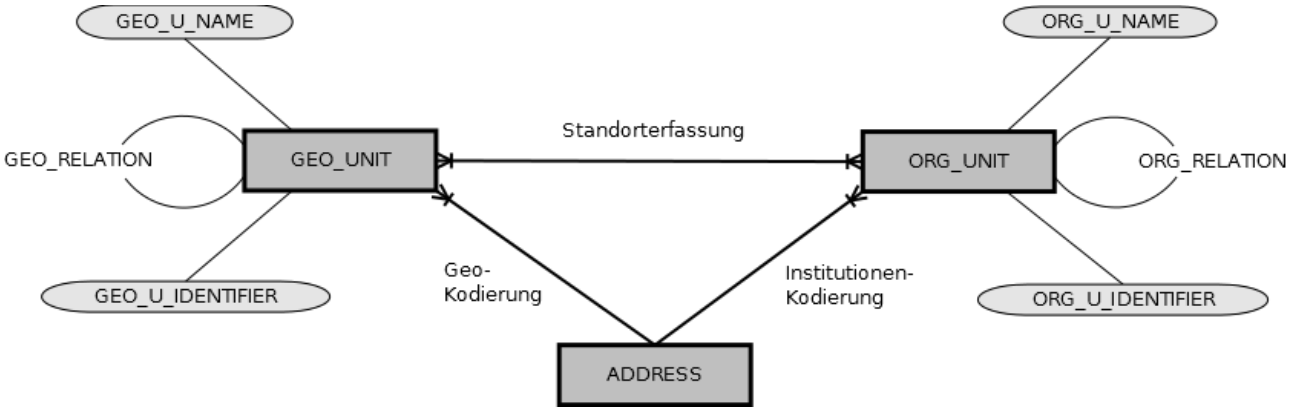


Abbildung 1: Geokodierung und Institutionenkodierung

Org-Units und Geo-Units haben strukturelle Gemeinsamkeiten: in beiden Fällen existieren Einheiten auf verschiedenen Aggregationsebenen, im Fall der Org-Unit zum Beispiel Fakultäten und Universitäten, im Fall der Geo-Unit Städte und Länder. Org-Units und Geo-Units haben ähnliche Attribute (wie beispielsweise Namen und Identifier) und für beide können Beziehungen zwischen mehreren Ausprägungen der Entitäten bestehen. Solche Beziehungen können hierarchischer Natur sein, müssen es aber nicht. Im Fall der Org-Units können beispielsweise Fakultäten in einer Teileinheitenbeziehung zu einer Universität stehen, im Fall der Geo-Units können Ortsteile in einer hierarchischen Beziehung zu einem Ort oder auch Orte zu einem Bundesland, Bundesländer zu einem Land usw. stehen.

Die drei Entitäten stehen jeweils miteinander in Beziehung – so kann eine Adresse einer Org-Unit zugeordnet werden; diese Beziehung wird durch die Institutionenkodierung erfasst. Andererseits kann die Adresse aber auch (unabhängig von einer Zuordnung zu einer Org-Unit) einer Geo-Unit zugeordnet werden, was durch die Geokodierung erfasst wird. Schließlich stehen auch Org-Unit und Geo-Unit miteinander in Beziehung (unabhängig von Adressen auf Publikationen): eine Org-Unit hat einen oder mehrere Standorte, Außenstellen usw. und eine Geo-Unit kann eine oder mehrere Org-Units beherbergen.

2. Statistik

Informationen zu Geo-Units sind in WoS und Scopus in unterschiedlichen Komponenten der Adressen und auf unterschiedlichen Ebenen verfügbar. Sowohl WoS als auch Scopus bieten separate Felder für Attribute wie country, city, plz, state und street mit entsprechend strukturierter Geoinformation, die jedoch unterschiedlich häufig verfügbar sind. Die Statistiken wurden auf Basis der Rohdatenbanken[5] des Kompetenzzentrums Bibliometrie erstellt, um den aktuellen Stand[6] zu erfassen.

Für das WoS sind die Produkte SCIE, SSCI, AHCI, CPCI-S und CPCI-SSH enthalten. In beiden Fällen (WoS und Scopus) wurden alle verfügbaren Jahrgänge[7] betrachtet.

Attribute zu Geoinformationen im WoS

Für das WoS wurde die Rohdatenbank im XML-Format verwendet. Hier sind folgende Attribute vorhanden, die Geoinformationen enthalten können:

- country
- state
- city
- street
- postal code
- fulladdress
- province
- post_num
- url

Dabei sind province, post_num und url immer leer, die übrigen Attribute eignen sich prinzipiell als Quellen für Geoinformationen zur Adresse. Fulladdress enthält die gesamte Adresse als String; also ist hier nicht nur die geographische, sondern auch die institutionelle Information enthalten.

Attribute zu Geoinformationen in Scopus

In Scopus sind die Geoinformationen in zwei Tabellen enthalten: affiliations und addresses – wobei bei einem Adressrecord die Werte für bestimmte Attribute aus affiliations und aus addresses gleich sind (z.B. entspricht der Inhalt der Spalte affiliations.address_part der Spalte addresses.address_part), so dass hier nicht beide Attribute benötigt werden. Für Scopus wurden folgende Attribute einbezogen:

- address_part
- city_group
- postalcode
- state
- country_code

Im Folgenden werden die Attribute als street, city, plz, state und country bezeichnet, um nicht jeweils beide Attributnamen (WoS und Scopus) nennen zu müssen. Die folgende Tabelle enthält die Bezeichnungen der Attribute im WoS mit den entsprechenden Bezeichnungen der verwendeten Attribute in Scopus.

5 Die im Folgenden verwendeten Bezeichnungen der Datenfelder entsprechen diesen Rohdatenbanken.
6 Recherchedatum: 12.04.2016
7 WoS: 1980 - 04/2016; Scopus: 1995 - 04/2016,

	Bezeichnung WoS (XML)	Bezeichnung Scopus
street	Street (NN im tagged format)	address_part
plz	postalcode	postalcode
city	city (NY im tagged format)	city_group
state	state	state
country	country	country_code

Tabelle 1: Bezeichnungen für Geo-Attribute in WoS und Scopus

Wenn im Folgenden von 'distinkten Adressen' die Rede ist, sind hier im Fall von WoS 'distinct fulladress' gemeint. Im Fall von Scopus existiert kein Attribut, das die Adresse als Gesamtstring enthält, daher wurde hier für die Zählung der distinkten Adressen nach organization1-3, country_code, text, address_part, city_group, postalcode gruppiert. Unter 'Adressen' wurden alle erfassten Adress-Records in der jeweiligen Datenbank gezählt (nicht distinkt).

Die Prozentangaben in den folgenden Tabellen beziehen sich jeweils auf die in der ersten Zeile genannte Grundgesamtheit und sind auf ganze Zahlen gerundet.

Die Tabellen 2 und 3 enthalten zunächst Informationen über die Vollständigkeit der Informationen zu den entsprechenden Attributen.

Anzahl Adressen	WoS	Scopus
Total	106.818.053 (100 %)	100.292.560 (100 %)
… davon mit country	100 %	96 %
… davon mit state	49 %	1 %
… davon mit city	~ 100 %	89 %
… davon mit plz	78 %	3 %
… davon mit street	27 %	32 %

Tabelle 2: Vollständigkeit der Attribute bezogen auf die Anzahl der Adressen insgesamt

Anzahl distinkte Adressen	WoS	Scopus
Total	28.891.625 (100 %)	39.505.921 (100 %)
… davon mit country	100 %	96 %
… davon mit state	41 %	2 %
… davon mit city	~ 100 %	89 %
… davon mit plz	72 %	3 %
… davon mit street	40 %	39 %

Tabelle 3: Vollständigkeit der Attribute bezogen auf die Anzahl der distinkten Adressen insgesamt

Für die Betrachtung der 'deutschen' Adressen im WoS wurden alle Adressen einbezogen, die über die Länderkodierung dem deutschen ISO-3-Code 'DEU' zugeordnet werden können. Es kommen folgende Länderbezeichner vor:

Country	Anzahl Adressen
Germany	4.708.411
GERMANY	690.711
FED REP GER	511.950
GER DEM REP	78.442
WEST GERMANY	159
FED RER GER	51
EAST GERMANY	10
Germany;	4
DEUTSCH DEM REP	3
Deutschland	2
GER DEM	1
DEUTSCH DEM REP	1
W GER	1
Summe	5.989.746

Tabelle 4: Häufigkeiten von country – Ausprägungen für deutsche Adressen im WoS

Im Fall der deutschen Adressen konzentrieren sich die country-Ausprägungen also auf sehr wenig verschiedene Werte. In Scopus ist keine Länderkodierung erforderlich, da bereits ISO3-Codes in den Rohdaten gegeben sind.

Die so ermittelten deutschen Adressen enthalten die Attribute mit folgenden Häufigkeiten:

Anzahl deutsche Adressen	WoS	Scopus
Total	5.989.746 (100 %)	5.359.612 (100 %)
… davon mit state	< 1 %	~ 0 %
… davon mit city	~ 100 %	91 %
… davon mit plz	79 %	3 %
… davon mit street	34 %	52%

Tabelle 5: Vollständigkeit der Attribute bezogen auf die Anzahl der deutschen Adressen

Anzahl distinkte deutsche Adressen	WoS	Scopus
Total	1.921.380 (100 %)	2.458.468 (100 %)
… davon mit state	1 %	~ 0 %
… davon mit city	~ 100 %	91 %
… davon mit plz	76 %	4 %
… davon mit street	44 %	52 %

Tabelle 6: Vollständigkeit der Attribute bezogen auf die Anzahl der distinkten deutschen Adressen

Das Attribut city ist also im WoS sowohl insgesamt als auch in den deutschen Adressen flächendeckend vorhanden. Postleitzahlen sind zwar weniger häufig, aber doch auch in der überwiegenden Zahl der Adressen verfügbar. Das state Attribut ist insgesamt bei 41% der distinkten Adressen verfügbar, jedoch für deutsche Adressen fast nie vorhanden. Das street Attribut ist zwar bei den distinkten Adressen mit 40 % (insgesamt) bzw. 44 % (deutsche Adressen) vorhanden, lt. Anbieterdokumentation des WoS wurden zu diesem Attribut ('NN' im tagged format) bis 1995 (Jahr der Prozessierung) Informationen zu Teileinheiten (z.B. Fakultäten zu Universitäten) erfasst:

Street address/other additional address information. Contains department, division name 1995 & earlier. Mixed case Science Citation Index Expanded™ & Social Sciences Citation Index®, 1996 & later, A&H 1997 & later; upper case previously. Not available 1986 & earlier.

Für Fälle vor dem Prozessierungsjahr 1996 sind dort also überwiegend gar keine Straßen- (und vermutlich auch keine sonstigen Geo-) Informationen zu finden. Auch für spätere Prozessierungjahre findet man dort nicht unbedingt nur Straßenangaben, sondern auch weitere postalische Informationen wie z.B. Postfächer. Dieses Attribut eignet sich also nur bedingt für die Extraktion von Geoinformationen.

Neben der Vollständigkeit der Werte für die einzelnen Attribute ist auch interessant, wie viele Varianten jeweils vorkommen. Die folgenden Tabellen zeigen die Attribute mit der Anzahl ihrer distinkten Ausprägungen. ‚PLZ-City' bezeichnet dabei PLZ-City-Kombinationen.

WoS	State	City	Street	PLZ	PLZ-City
Insgesamt	36.606	280.000	4.767.158	447.176	850.456
Deutschland	1.301	19.344	304.049	23.444	41.279

Tabelle 7: Anzahl distinkter Ausprägungen für Attribute, WoS

Scopus	State	City	Street	PLZ	PLZ-City
Insgesamt	7.405	3.029.719	3.488.541	133.632	-
Deutschland	222	110.053	192.565	6.573	-

Tabelle 8: Anzahl distinkter Ausprägungen für Attribute, Scopus

Auffällig ist hier vor allem der große Unterschied in der Anzahl der distinkten City-Ausprägungen, auf die im Folgenden noch näher eingegangen wird. Der Grund dafür ist, dass in Scopus die City-Angabe häufig eine Postleitzahl enthält, also im Prinzip als PLZ-City Kombination vorhanden ist. Aus diesem Grund ist eine Zählung der PLZ-City Kombinationen wie für das WoS nicht sinnvoll.

Struktur der Attribute und Werte

Street

Während in der Dokumentation vom WoS ein Hinweis darauf vorhanden ist, dass das Street-Attribut im WoS in den Jahren vor 1996 als Attribut für Teileinheiten diente, ist bei Scopus ein derartiges Problem nicht erkennbar. In Scopus heißt das entsprechende Attribut 'address_part': hier sind nicht nur Straßennamen, sondern auch andere Bestandteile einer postalischen Adresse enthalten (beispielsweise Postfächer). Auch im WoS enthält das Street-Attribut (auch für Jahre ab 1996) nicht ausschließlich Straßennamen.

City & PLZ

Während in Scopus 85.687 (77,9 %) der distinkten Ausprägungen des City-Attributs dem regulären Ausdruck '[0-9]{4,5}' ('enthält 4-5 Ziffern') entsprechen, ist das im WoS bei nur 113 (1 %) der distinkten Ausprägungen des City-Attributs der Fall. In Scopus ist also sehr oft eine PLZ im City-Wert enthalten, im WoS dagegen nicht. Das erklärt die großen Unterschiede in der Zählung der distinkten Ausprägungen für City-Werte in WoS und Scopus. Scopus enthält damit nicht nur mehr Varianz (allein schon durch die Darstellung von Sonderzeichen), sondern es existieren auch strukturelle Unterschiede. Das bedeutet zum einen, dass die sehr niedrige Anzahl von PLZ-Werten

in Scopus täuscht – diese Informationen fehlen nicht, sondern sind im City-Attribut enthalten – und zum anderen, dass für eine weitere Verarbeitung eine Extraktion der PLZ vorgenommen werden muss, da sie nicht in der passenden Struktur vorliegt.

Stichproben

Attribute für Geoinformationen

Für jedes Attribut wird (für WoS und Scopus) eine Stichprobe von jeweils 100 distinkten Ausprägungen gezogen und manuell daraufhin überprüft, ob inhaltlich zum Attribut passende Werte angegeben sind (ist beispielsweise der Wert des City-Attributs tatsächlich ein Bezeichner für eine Stadt, der Wert im street-Attribut ein Bezeichner für eine Straße?).

Diese Stichproben sollen einen Hinweis darauf geben, wie verlässlich die Attribute zu Geoinformationen auch tatsächlich die passenden Werte enthalten.

Street:

Da bekannt ist, dass im WoS in diesem Attribut vor 1996 noch Bezeichnungen für Teileinheiten enthalten waren, wurde hier die Stichprobe auf Publikationsjahre ab 1996 beschränkt.

Enthaltene Information	**WoS**	**Scopus**
Straße	50	88
Postfach/Postbox	8	3
Straße und weitere Elemente	24	5
Postfach/Postbox und weitere Elemente	1	-
Gebäudebezeichnung	5	3
Keine Geoinformation	12	-
Vmtl. Ausland	-	1

Tabelle 9: Street Attribut, enthaltene Geo-Informationen

PLZ:

Hier wurde keine inhaltliche Prüfung vorgenommen (also nicht geprüft, ob die eingetragene PLZ tatsächlich eine PLZ ist oder war), sondern lediglich die Struktur/Form der Werte erfasst.

Enthaltene Information	WoS	Scopus
Keine PLZ	-	1
5-stellige PLZ	-	52
'D-' und 5stellige PLZ	65	23
'DE-' und 5stellige PLZ	7	9
'D ' und 5stellige PLZ	-	1
'D' und 5stellige PLZ	-	2
Weitere Form	-	4
5stellige PLZ und weitere Elemente	-	3
Mehr als 5 Stellen	3	3
Formfehler Buchstabe statt Zahl, z.B. O statt 0	-	1
4stellige PLZ	22	1
Weniger als 4 Stellen	1	-
Erkennbar Ausland (z.B. CH-...)	2	-

Tabelle 10: PLZ Attribut, enthaltene Geo-Informationen

City:

Enthaltene Information	WoS	Scopus
Stadt	81	14
PLZ und Stadt	-	60
Stadt und weitere Elemente	6	13
Nur PLZ	-	3
Ausland	3	9
Keine Geoinformation zur Stadt oder PLZ	10	1

Tabelle 11: City Attribut, enthaltene Geo-Informationen

State:

Enthaltene Information	WoS	Scopus
Bundesland	22	18
Stadt/PLZ	46	61
Abkürzung	11	9
Straße	2	7
Institutionsbezeichnung	-	5
Sonstige Geoinformation (Region, See, …)	8	-
Ausland	6	-
Ohne weitere Informationen keine Interpretation möglich	1	-
Keine Geoinformation	4	-

Tabelle 12: State Attribut, enthaltene Geo-Informationen

Weitere Attribute

Eine weitere wertvolle Information für ein Verfahren zur Zuordnung von Adressen zu geografischen Einheiten ist, in welchen Attributen Geoinformation zu finden ist. Dies kann – außer in den in der Struktur der Datenbankanbieter für Geoinformationen vorgesehenen Attributen wie Land, Stadt, Straße usw. – auch in Attributen zur Institution der Fall sein. Beispielsweise enthält der Eintrag 'Univ Bielefeld' in Organization1 eine Geoinformation ('Bielefeld'), die bei der Zuordnung genutzt werden kann. So kann entschieden werden, welche Attribute in das Verfahren zur Zuordnung von Adressen zu geografischen Einheiten sinnvollerweise einbezogen werden sollten.

Für eine Zufallsauswahl von jeweils 1.000 Adressen in WoS und Scopus wurde manuell überprüft, in welchen Attributen (die nicht direkt dafür vorgesehen sind) sich Geoinformationen befindet. Dabei kann eine Geoinformation die Bezeichnung eines Landes, einer Stadt, einer Straße oder eine Postleitzahl sein. Die Bewertung (Geoinformation enthalten oder nicht) ist dabei unabhängig davon, ob die Geoinformation Bestandteil des Namens ist. Die Geoinformation muss direkt enthalten sein (also ohne die Hinzunahme externer Informationen). So enthält der String 'Univ Mainz' die Geoinformation 'Mainz', der String 'Johannes Gutenberg Universität' aber nicht, da hier die externe Information 'Johannes Gutenberg Universität befindet sich in Mainz' erforderlich wäre, um eine Geoinformation zu erkennen.

WoS:

Organization sind hier die Werte, die nicht als 'Suborganization' geflaggt sind (neue Struktur des XML-Formats, im tagged Format sind die Daten in einer anderen Struktur gegeben):

Enthaltene Information	Organization	Suborganization
Keine Geoinformation	256	862
Stadt	720	25
Land (Deutschland)	16	4
Straße	-	-
Bundesland	14	2
Postfach	-	-
Sonstige Geoinformation (Region, …)	5	5
Ortsteil	3	4

Tabelle 13: Enthaltene Geo-Informationen (WoS)

Damit scheint ein Organization-Eintrag ohne Suborganization-Flag eine wertvolle Informationsquelle zu sein, da in 72% der Fälle eine Geoinformation auf Stadtebene enthalten ist, ein Eintrag mit Suborganization-Flag dagegen enthält nahezu nie verwertbare Geoinformationen.

Scopus:

Enthaltene Information	Organization1	Organization2	Organization3
Keine Geoinformation	464	573	164
Stadt	482	83	28
Land (Deutschland)	15	5	-
Straße	6	-	-
Bundesland	12	7	2
Postfach	1	-	-
Sonstige Geoinformation (Region, …)	4	5	1
Ortsteil	-	-	-
NULL (kein Wert für das Attribut)	15	327	805
Ausland	1	-	-

Tabelle 14: Enthaltene Geo-Informationen (Scopus)

Hier enthält also Organization1 häufig verwertbare Informationen, Organization2 in einigen Fällen (ca. 8%), Organization3 dagegen nahezu nie.

3. Tabellenstruktur Basistabellen

Analog zur Institutionenkodierung werden Basisdaten (Geo-Einheiten und ihre Attribute und Relationen) in Tabellen erfasst ('Basistabellen'). Diese Basistabellen werden im Verfahren benötigt und enthalten unter anderem auch die Geo-Einheiten als Klassifikationsziele. Die Basisdaten sind außerdem (wie auch in der Institutionenkodierung) eine wichtige Grundlage der in die Bibliometriedatenbanken zu integrierenden Datenlieferungen.

Anforderungen

Geo-Einheiten, denen die Adressen (und damit Publikationen) aus WoS und Scopus zugeordnet werden, sollen mit ihren Attributen und Relationen in Tabellen einer relationalen Datenbank erfasst werden, deren Schema den folgenden Anforderungen genügen soll:

- verschiedene Typen von Geo-Units müssen aufgenommen werden können
- Flexibilität (verschiedene Strukturen in verschiedenen Ländern)
- nachträgliches Einfügen von Attributen muss problemlos möglich sein
- Hierarchiebeziehungen müssen abgebildet und abgefragt werden können (z.B. Stadt-Bundesland-Land)
- die Struktur sollte anschlussfähig an die Institutionenkodierung sein (vgl. oben Abschnitt „Geokodierung & Institutionenkodierung“)
- sie soll eine sinnvolle Extraktion der Ergebnisse der Geokodierung für die Einspeisung in die Bibliometriedatenbanken ermöglichen

Struktur

Die folgende Basistabellenstruktur kann in der Tabelle GEO_UNIT Geo-Einheiten unterschiedlicher Typen und Hierarchieebenen aufnehmen: Länder, Bundesländer, Städte, Postleitzahlengebiete usw. Der hier vergebene Primärschlüssel (Primary Key, PK) identifiziert die Geo-Unit. Außer diesem enthält die Tabelle noch die Geokoordinaten der Einheit und den Typ (in TYPE). Die zugehörige Lookup-Tabelle ist GEO_TYPE und enthält Bezeichnungen wie 'PLZ-Gebiet', 'Bundesland', 'Gemeinde' usw. Weitere Arten von Geo-Einheiten (beispielsweise Kantone) können hier einfach durch Ergänzungen in der Tabelle GEO_TYPE aufgenommen werden, ohne dass eine Änderung des Tabellenschemas notwendig würde.

Namen von Geo-Einheiten sind (wie auch in der Institutionenkodierung) in einer separaten Tabelle (GEO_U_NAME) enthalten. In der Tabelle können auch Start- und Enddaten für Namen angegeben werden (im Fall von Namensänderungen). In dieser Tabelle werden Namen für alle Typen von Geo-Einheiten erfasst – so kann z.B. der Name '33039' für eine Postleitzahl enthalten sein, aber auch Bezeichnungen für Städte (wie 'Frankfurt am Main') oder Länder (wie 'Deutschland').

In der Tabelle GEO_U_IDENTIFIER können unterschiedliche Identifier für eine Geo-Einheit erfasst werden – dabei werden für jeden Typ von Geo-Einheiten die Identifier erfasst, die für diesen

sinnvoll sind. Beispiele sind ISO3-Codes für Länder, NUTS-Codes[8] und AGS (amtliche Gemeindeschlüssel) für Gemeinden. Die Spalte ID_TYPE enthält dabei die Art des Identifiers (Lookup-Tabelle für die Identifier-Typen ist GEO_U_ID_TYPE), während ID_VALUE den Identifier selbst enthält. Beispiel: für Deutschland kann hier für den Typ ISO-3-Code der Wert 'DEU' erfasst werden.

Die Tabelle GEO_RELATION enthält – analog zur Tabelle RELATION in den Basistabellen zur Institutionenkodierung – Beziehungen zwischen Geo-Einheiten, die mit TYPE (Lookup-Tabelle GEO_RELATION_TYPE) näher bezeichnet werden und ein Start- und Enddatum erhalten. Mit dieser Tabelle ist es möglich, Geo-Einheiten auf bestimmten Ebenen zu aggregieren und so zum Beispiel alle Städte eines Landes oder alle Postleitzahlengebiete zu einer Stadt (oder einem Land) zu finden.

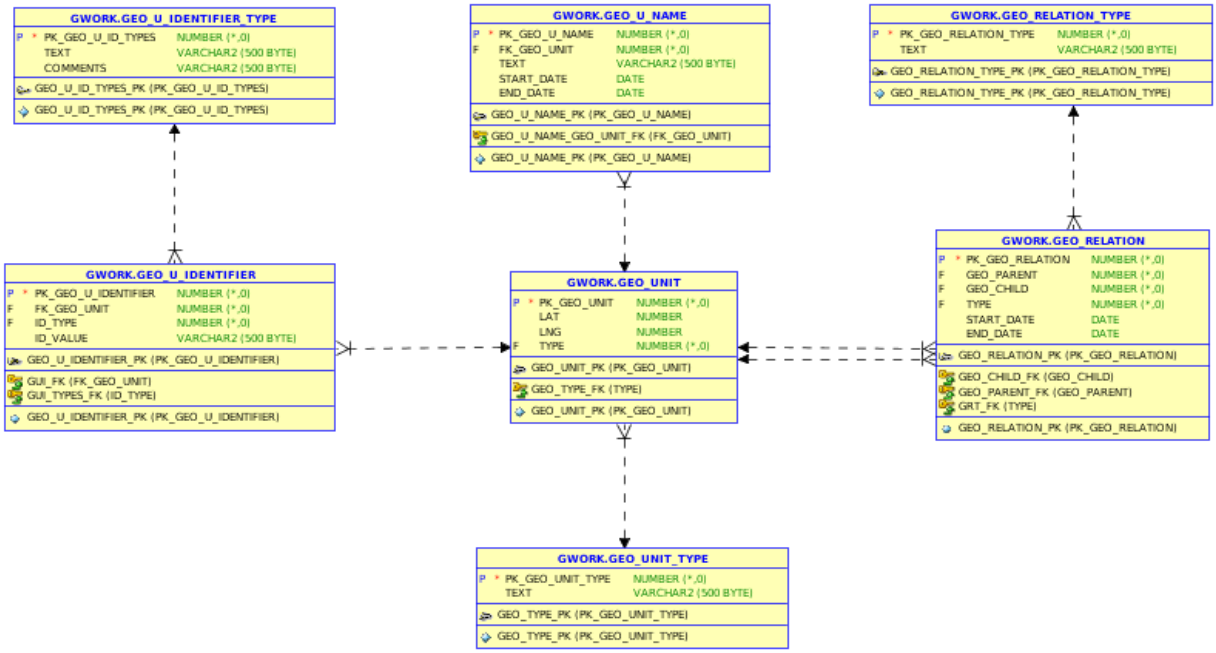


Abbildung 2: Tabellenstruktur Basistabellen

8 NUTS (Nomenclature des Unités territoriales statistiques): Klassifikation der Gebietseinheiten für die Statistik, gemäß der das Gebiet der Europäischen Union in eine geografische Systematik in drei Hierarchiestufen eingeteilt ist.

4. Datenquellen

Anforderungen

Die Datenquellen für Geodaten, die hier verwendet werden sollen, müssen bestimmten Anforderungen genügen. Zum einen sollen nur frei verfügbare Quellen verwendet werden, um bei der späteren Nutzung lizenzrechtliche Probleme und Kosten zu vermeiden. Zum anderen sind formale und inhaltliche Mindestanforderungen zu erfüllen: zur sinnvollen Nutzung ist eine einfache Downloadmöglichkeit in einem passenden Format (oder in einem Format, das mit entsprechenden Tools einfach geparst werden kann) unerlässlich.

Es sollte mindestens Deutschland abgedeckt sein (idealerweise mehr, um den Ansatz unproblematisch auf weitere Länder erweitern zu können) und für den jeweils abgedeckten Bereich sollten möglichst viele der relevanten Daten (insbesondere Städte und Postleitzahlen, die im WoS oder Scopus zu finden sind) mit zugehörigen Geokoordinaten vorhanden sein. Außerdem sollten Relationen vorhanden sein, die eine Aggregation erlauben (z.B.: alle Postleitzahlen zu einer Stadt, alle Städte eines Bundeslandes usw.).

Verfügbare Quellen

Es stehen verschiedene Quellen für Geodaten um Internet zur Verfügung. Dabei scheiden einige aufgrund der oben genannten Anforderungen aus – beispielsweise ‚Geodaten Deutschland'[9] aufgrund der Nutzungsbedingungen, das ‚Geoportal RLP'[10] aufgrund der Beschränkung auf ein einzelnes Bundesland usw.

Nach einer Vorauswahl anhand der oben genannten Mindestanforderungen scheinen folgende Datenquellen besonders geeignet:

- Geonames
- OpenGeoDB
- OpenStreetMap
- Wikipedia

Im Folgenden werden die einzelnen Quellen in Bezug auf Datenverfügbarkeit und -struktur kurz vorgestellt.

GeoNames

GeoNames[11] ist eine frei verfügbare Geodatenbank (creative commons attribution license) – die Daten sind nicht auf bestimmte Länder beschränkt. Neben der Möglichkeit einer Online-Suche stehen extract files zum Download zur Verfügung (tab-delimited text, utf8 encoding). Neben der frei verfügbaren Version ist es auch möglich, kostenpflichtig modifizierte und geprüfte Daten (‚Premium Data'[12]) zu beziehen.

9 https://www.geodaten-deutschland.de/
10 http://www.geoportal.rlp.de/portal/informationen/open-source-geoportal.html
11 http://www.geonames.org/
12 http://www.geonames.org/products/premium-data.html

GeoNames enthält nach Angaben auf der Homepage[13] (Stand 05.10.2016) aktuell über 10 Millionen geografische Namen. Die Daten werden aus verschiedenen Quellen bezogen und können von Nutzern über ein Wiki-Interface manuell geändert, korrigiert und ergänzt werden.

Die Daten liegen in der folgenden Struktur vor: es ist eine Haupttabelle (‚main geoname table') verfügbar, die die Geoentitäten mit Identifiern, Namen, Geokoordinaten und weiteren Attributen enthält, sowie mehrere Lookup-Tabellen (beispielsweise Kodierungen für ‚administrative divisions'), weitere Tabellen (mit beispielsweise Postleitzahlen oder alternativen Namen) und eine Tabelle namens ‚hierarchy', die hierarchische Beziehungen zwischen Geoentitäten enthält. Zum Download stehen sowohl der Gesamtdatenbestand als auch Teilmengen für einzelne Länder zur Verfügung.

Für die vorliegende Untersuchung wurden die Daten für Deutschland extrahiert und in eine relationale Datenbank geladen. Die Daten lassen sich einfach importieren und die Struktur eignet sich gut, um sie in das gewünschte Format zu überführen. Ein weiterer Vorteil ist die Verfügbarkeit von Daten über Deutschland hinaus.

OpenGeoDB

OpenGeoDB ist ein Projekt zum Aufbau einer Geodatenbank, das sich bisher auf die Länder Österreich, Belgien, die Schweiz, Deutschland und Liechtenstein beschränkt:

*‚Im Mittelpunkt des Projektes **OpenGeoDB** steht der Aufbau einer möglichst vollständigen Datenbank mit Geokoordinaten zu allen Orten und Postleitzahlen (bisher: A,B,CH,D und FL). Dies soll vor allem durch die Beteiligung von möglichst vielen Personen geschehen, die diese zentrale Datenbank pflegen.'*[14]

Die Daten sind frei verfügbar (im Rahmen der Gemeinfreiheit) und stehen als SQLDumps zur Verfügung[15], die einen einfachen Import ermöglichen. Abbildung 3 zeigt das Datenbankschema[16].

13 http://www.geonames.org/about.html
14 http://opengeodb.giswiki.org/wiki/OpenGeoDB
15 http://opengeodb.org/wiki/OpenGeoDB_Downloads
16 http://opengeodb.org/wiki/Datenbank

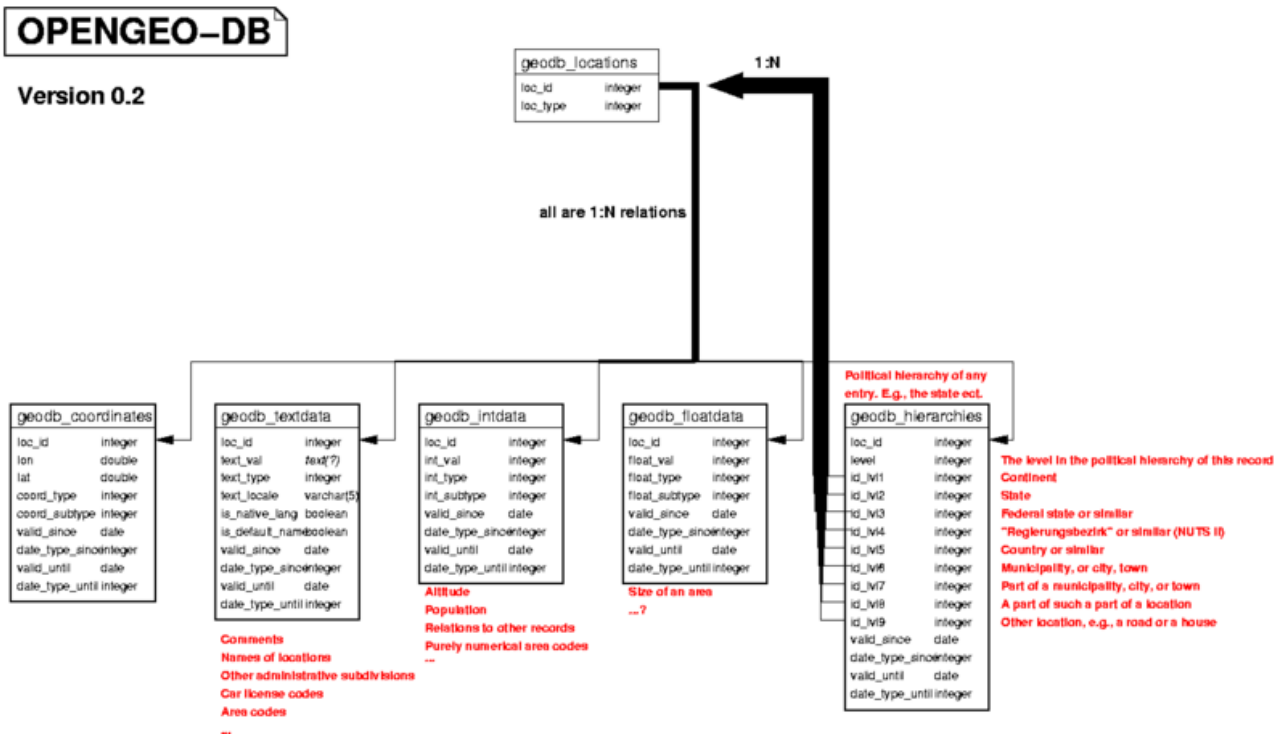


Abbildung 3: Datenbankschema open GeoDB

Die Geoentitäten werden mit einem Identifier (loc_id) und einem Typ in der Tabelle geodb_locations erfasst. Entitäten können dabei von beliebiger Art sein, zum Beispiel Städte, Ortsteile, Seen, Postleitzahlengebiete oder Kontinente. Über den Typ lässt sich die Art der Geoentität bestimmen. Attribute sind in weitere Tabellen ausgelagert: Geokoordinaten befinden sich in geodb_coordinates, andere Attribute werden je nach Datentyp der Werte in den Tabellen geodb_textdata, geodb_intdata und geodb_floatdata vorgehalten (Beispiele sind Namen, Identifier, Einwohnerzahlen usw.). Dabei ist jeweils in einer Typ-Spalte erkennbar, um welches Attribut es sich handelt, während eine Spalte namens ‚val' den Wert für dieses Attribut enthält. Der text_typ 500100000 steht beispielsweise für ‚Namen', so dass der Name der Geoentität mit loc_id 399 wie folgt abgefragt werden kann:

```
SELECT TEXT_VAL
FROM   GEODB_TEXTDATA
WHERE  TEXT_TYPE= 500100000 AND LOC_ID=399
```

```
→ Bielefeld
```

Hierarchiebeziehungen werden in einer weiteren Tabelle namens geodb_hierarchies abgelegt. Diese Tabelle enthält Hierarchiebeziehungen auf allen Ebenen (z.B. Bundesländer oder Länder zu Städten), wobei die Ebene mit angegeben ist. Diese Tabelle ist im Standarddump nicht enthalten, da sich die Daten aus den in den anderen Tabellen enthaltenen Informationen erzeugen lassen (so enthält zum Beispiel die Tabelle geodb_intdata hierarchische Relationen zwischen loc_ids unter der Typ-ID 400100000, bezeichnet mit ‚Teil von').

Damit sind Aggregationen wie gewünscht möglich und die Datenstruktur ist zur Überführung in das gewünschte Format gut geeignet. Der Nachteil an OpenGeoDB ist, dass zwar Daten für ganz Deutschland verfügbar sind, darüber hinaus aber nur für wenige weitere Länder.

Damit ist die Datenquelle zwar für Deutschland (und Österreich, Belgien, die Schweiz und Liechtenstein) nutzbar, erlaubt aber keine Erweiterung auf z.B. die EU oder gar alle Länder.

OpenStreetMap (OSM)

OpenStreetMap ist ein internationales Projekt zur Schaffung einer freien Weltkarte (keine Einschränkung auf bestimmte Länder)[17], Daten werden von einer offenen Community gesammelt, korrigiert und ergänzt. Entitäten, die hier aufgenommen werden, sind nicht notwendigerweise Städte, Ortsteile usw., sondern es werden beispielsweise auch Bäume, Parkbänke usw. mit zugehörigen Geokoordinaten und Attributen als Entitäten aufgenommen – sowie auch Geokoordinaten von Punkten, die z.B. zur Definition einer Straße erforderlich sind.

Die Daten sind frei nutzbar, verschiedenen Ausschnitte (beispielsweise nach Ländern) stehen zum Download zur Verfügung[18]. Die Daten werden im OSM-XML-Format[19] zur Verfügung gestellt. Ein OSM-Element enthält drei Blöcke:

- einen ‚nodes'-Block
- einen ‚ways'-Block
- und einen ‚relations'-Block

Ein ‚node' definiert einen Punkt. Nodes sind mindestens mit einem Identifier und Geokoordinaten (und einigen Metadaten wie version, changset, user, uid, visible und timestamp) versehen, optional mit Tags, die die Geoentität näher beschreiben. Beispiel[20]:

```
 <node id="1831881213" version="1" changeset="12370172" lat="54.0900666"
lon="12.2539381" user="lafkor" uid="75625" visible="true" timestamp="2012-07-
20T09:43:19Z">
  <tag k="name" v="Neu Broderstorf"/>
  <tag k="traffic_sign" v="city_limit"/>
 </node>
```

Für diesen Anwendungsfall sind nur Nodes mit mindestens einem Tag interessant. Ways definieren Wege, Straßen, Buslinien usw., die hier keine Rolle spielen. Relations können verschiedene Beziehungen zwischen zwei oder mehr Nodes oder Ways enthalten. Sie enthalten jeweils einen Identifier, die bereits oben genannten Metadaten, zwei oder mehr ‚members' (die mit type, reference und role angegeben werden) und optional Tags, die die Relation näher beschreiben.

17 https://www.openstreetmap.org/about
18 http://www.geofabrik.de/de/data/
19 http://wiki.openstreetmap.org/wiki/OSM_XML
20 http://wiki.openstreetmap.org/wiki/OSM_XML

Ein Beispiel:

```
 <relation id="56688" user="kmvar" uid="56190" visible="true" version="28"
changeset="6947637" timestamp="2011-01-12T14:23:49Z">
  <member type="node" ref="294942404" role=""/>
  ...
  <member type="node" ref="364933006" role=""/>
  <member type="way" ref="4579143" role=""/>
  ...
  <member type="node" ref="249673494" role=""/>
  <tag k="name" v="Küstenbus Linie 123"/>
  <tag k="network" v="VVW"/>
  <tag k="operator" v="Regionalverkehr Küste"/>
  <tag k="ref" v="123"/>
  <tag k="route" v="bus"/>
  <tag k="type" v="route"/>
 </relation>
```

Hier werden die Daten für Deutschland genutzt. Mit Hilfe eines OSM-Parsers für Python werden alle Nodes mit mindestens einem Tag und alle Relations in csv-Files geschrieben und in eine relationale Datenbank importiert. Der Import ist damit für OSM etwas aufwändiger.

In den Bezeichnern (Keys und Values) der Tags (sowohl für Nodes als auch für Relations) ist Varianz vorhanden, so dass es nicht einfach möglich ist, anhand der Tags die Nodes zu identifizieren, die in diesem Zusammenhang relevant sind (z.B. Städte, Ortsteile und Gemeinden im Gegensatz zu Bäumen und Straßenschildern). Gleiches gilt für Aggregationen: es ist nicht ganz einfach, die relevanten Informationen extrahieren, auch wenn sie im Prinzip enthalten sind.

Wikipedia

Für die englischsprachige Wikipedia besteht die Möglichkeit, Teile der Wikipedia herunterzuladen[21], unter anderem Seiteninformationen mit Titel und auch Geo-Tags zu den Seiten. Die Daten sind als Dump in SQL und XML erhältlich .

Geeigneter als der Datenbestand der englischsprachigen Wikipedia erscheint jedoch für deutsche Adressen die deutschsprachige Wikipedia, da hier in Bezug auf Deutschland mehr Informationen enthalten sind.

DBPedia bietet die Möglichkeit, viele Informationen aus der Wikipedia in strukturierter Form zu extrahieren[22] – diese können über einen Endpoint[23] mit der Sprache SPARQL abgefragt oder (für viele Sprachen) als Download[24] von beispielsweise RDF-Triples in Turtle Syntax[25] bezogen werden. Hier wird die letztgenannte Möglichkeit für den Bezug von Daten aus der deutschsprachigen Wikipedia genutzt. Zunächst

21 https://en.wikipedia.org/wiki/Wikipedia:Database_download
22 http://wiki.dbpedia.org/about
23 z.B.: https://dbpedia.org/sparql
24 http://wiki.dbpedia.org/Downloads2015-10
25 https://www.w3.org/TR/turtle/

werden alle relevant erscheinenden Tabellen in eine relationale Datenbank geladen, um einen Überblick über die enthaltenen Entitäten und die Nutzbarkeit zu erhalten.

Die RDF-Triples haben folgende Form:

<Ressource> <Property> <Value>

Das Triple

<http://de.dbpedia.org/resource/Bielefeld> <http://de.dbpedia.org/property/art> "Stadt"@de

beinhaltet also beispielsweise die Aussage, dass Bielefeld eine Stadt ist. Für Properties existieren Ontologien – es gibt in der Wikipedia jedoch keine strikten Regeln zur Verwendung von bestimmten Properties bzw. Ontologien, hier existiert also eine gewisse Varianz an Bezeichnungen. Der ‚Value' kann entweder selbst wieder als eine Ressource (wenn zu dem entsprechenden Value eine Wikipedia-Seite existiert) oder aber in einem bestimmten Datentyp angegeben sein, wie z.B. der String *"Stadt"@de*. Dabei ist einem String jeweils die entsprechende Sprache hinzugefügt (im Beispiel durch *‚@de')* und anderen Datentypen jeweils die Bezeichnung des Datentyps (wie z.B. "33501"^^<http://www.w3.org/2001/XMLSchema#**integer**> für einen Integer).

Da die Wikipedia keine auf Geodaten beschränkte Datenquelle ist, müssen zunächst die Entitäten extrahiert werden, die für diesen Anwendungszweck relevant sind (also im Wesentlichen Gemeinden, Stadtteile, Städte, Orte, Bundesländer usw.). Aufgrund der Varianz in den Bezeichnungen für die Properties ist es jedoch nicht ganz einfach, beispielsweise alle Orte in Deutschland zu finden. Während Bielefeld, wie oben gezeigt, über die Property *<http://de.dbpedia.org/property/**art***> als Stadt erkannt werden kann, muss diese Property für andere Orte nicht unbedingt existieren; so wird die ‚Art' des Nieheimer Ortsteils Erwitzen beispielsweise über die Property <http://de.dbpedia.org/property/**typ**> angegeben:

<http://de.dbpedia.org/resource/Erwitzen><http://de.dbpedia.org/property/typ> "Ortsteil"@de

Für den Ort Hinterzarten ist gar keine Property *<http://de.dbpedia.org/property/**art***> vorhanden; <http://de.dbpedia.org/property/**typ**> existiert, jedoch mit dem Value *"g"@de*, dessen Bedeutung nicht unmittelbar klar ist.

In einigen Fälle ist auch eine Property mit mehreren Values zu einer Ressource vorhanden:

<http://de.dbpedia.org/resource/Holzwickede>

***<http://de.dbpedia.org/property/typ>** "StadtGemeinde"@de*

und auch

<http://de.dbpedia.org/resource/Holzwickede>

***<http://de.dbpedia.org/property/typ>** "g"@de .*

Außerdem gibt es Fälle, in denen weder ‚Typ' noch ‚Art'-Properties angegeben sind, beispielsweise im Fall von Kölkebeck (einem Ortsteil von Halle Westf.). Hier wird auf folgende Art beschrieben, dass es sich bei Kölkebeck um einen Ortsteil von Halle handelt:

<http://de.dbpedia.org/resource/Kölkebeck>

*<http://de.dbpedia.org/property/**ortsteil**> "Kölkebeck"@de*

<http://de.dbpedia.org/resource/Kölkebeck>

*<http://de.dbpedia.org/property/**gemeindeart**> "Stadt"@de*

<http://de.dbpedia.org/resource/Kölkebeck>

*<http://de.dbpedia.org/property/**gemeindename**> "Halle"@de*

Auch das Extrahieren von Relationen und Hierarchien (im Sinne von z.B. ‚alle Städte in Deutschland') ist damit nicht unmittelbar und einfach möglich, obwohl die Information im Prinzip vorhanden ist.

Zusammenfassung Quellen

Zusammenfassend ist festzuhalten: die beiden erstgenannten Quellen sind stärker strukturiert – Informationen und Aggregationen lassen sich leichter extrahieren und in das gewünschte Format übertragen. Dabei ist jedoch ein deutlicher Nachteil von openGeoDB die Einschränkung auf nur wenige Länder. Die Wikipedia (und DBPedia) enthalten von einer offenen Community zusammengetragene Daten, die infolgedessen nicht streng strukturiert sind, was das Extrahieren von Informationen und Aggregationen erschwert, jedoch den Vorteil einer sehr großen Datenmenge bietet.

Unterschieden werden muss hier der Einsatz der Quellen zur Erzeugung des Basisdatensatzes (also der Lookup-Tabellen für Geoentitäten und der Tabellen, die die Relationen zwischen den Geoentitäten beinhalten) und der Einsatz im Verfahren zur Zuordnung und zur Ermittlung weiterer Attribute. Während für ersteres eine klare Struktur sehr wichtig ist (und damit vor allem GeoNames und openGeoDB in Frage kommen), ist für letzteres eine größere Datenmenge (mit beispielsweise vielen verschiedenen Bezeichnern) sehr hilfreich.

Zu beachten ist jedoch, dass bei der Nutzung von verschiedenen Quellen jeweils auch ein Matching zwischen den genutzten Quellen vorhanden sein oder erzeugt werden muss. In den heruntergeladenen Wikipedia- (bzw. DBPedia-) Daten sind zwar ‚Geonames_Links' enthalten, in dieser Datei sind jedoch nur sehr wenige Deutschland zugeordnete DBPedia-Ressourcen tatsächlich enthalten, so dass diese Daten im Prinzip wertlos sind.

Auf der Seite http://wiki.dbpedia.org/Downloads2015-10 ist eine Liste mit Dateien vorhanden, die Links zu anderen Datenquellen enthalten sollen, Zugriff auf die Datei namens ‚Links to GeoNames' liefert jedoch einen File-not-found-Error (05.10.2016). Sofern die Datenquellen selbst also keine (oder nur unzureichende) Links zu anderen Datenquellen bereit stellen, müssten im Fall der Nutzung von mehreren Datenquellen diese Links zunächst erzeugt werden.

Abschätzung der Abdeckung in den Datenquellen

Um abschätzen zu können, wie wertvoll die Quellen für die Zuordnung der Adressdaten in WoS und Scopus zu Geoentitäten und Geokoordinaten sind, ist nicht entscheidend, wie viele Geoentitäten mit zugehörigen Koordinaten in den Quellen insgesamt vorhanden sind, sondern vielmehr, wie viele der in WoS und Scopus genannten Geoentitäten enthalten sind und ob sich die dazu in WoS und Scopus enthaltenen Namensvarianten und Postleitzahlen in den Quellen finden lassen. Deshalb wurde getestet, wie viele Werte der Attribute Stadt und Postleitzahl in WoS und Scopus sich mit geringer Vorprozessierung in den Datenquellen finden lassen.

Da die Geokodierung letztendlich für die Bibliometriedatenbanken (BDB) des Kompetenzzentrums Bibliometrie bereit gestellt werden soll, wurde hier die Vollständigkeit in Bezug auf die enthaltenen Entitäten bzw. Namensvarianten anhand der im Herbst 2016 aktuellen Bibliometriedatenbanken getestet (WOS_B_2016 und SCOPUS_B_2015). Dort sind die Informationen zu Städten und Postleitzahlen in den Spalten CITY und POSTALCODE der Tabelle INSTITUTIONS enthalten.

Matching zur Abschätzung der Abdeckung

Das Matching für die City-Werte wurde case-insensitive durchgeführt. Es wurde jeweils (sofern unten nichts weiteres beschrieben ist) nach genauen Treffern gesucht, d.h. ohne Front- oder Endtrunkierung und weitere Modifizierungen. Diese Auswertung soll nur einen groben Anhaltspunkt dafür bieten, welchen Wert die jeweilige Quelle für dieses Projekt haben könnte und ist nicht etwa bereits das Verfahren, das später zur Zuordnung angewendet wird. Daher wurden die Daten hier nur grob vor- bzw. aufbereitet, eine aufwändigere Prozessierung erfolgt im Verfahren selbst.

GeoNames

Für Postleitzahlen wurde die gesondert für Postleitzahlen bereit gestellte Tabelle genutzt (mit Einschränkung auf countrycode=‘DE‘). Die Postleitzahlen in der WoS-BDB liegen zum Teil mit Länderkenner vor der Postleitzahl vor (z.B. ‚D-...‘), die Postleitzahlen in GeoNames sind als fünfstellige Ziffernfolge gegeben. Daher wurde für das Matching die PLZ aus GeoNames mit den aus den Postleitzahlen der BDB extrahierten fünfstelligen Ziffernfolgen (*regexp_substr(postalcode,‘[0-9]{5,5}‘)*) verglichen. Für Scopus wurden die Postleitzahlen aus dem Citystring extrahiert (die Vorgehensweise wird unten näher beschrieben).

Für die Städtenamen wurde die Haupttabelle (‚MAIN‘) für Deutschland sowie die Tabelle ‚ALTERNATE_NAMES‘ genutzt.

OpenGeoDB

Für das PLZ- wie auch das City-Matching wurden die Werte (TEXT_VAL) aus der Tabelle TEXTDATA verwendet. Die ID für den Typ-Namen ‚Postleitzahl‘ ist 500300000, daher wurden für die Postleitzahl alle Werte aus Zeilen mit dieser Typ-ID verwendet. Für das City-Matching wurden die Typen 500100000 (‚Name‘) und 500100002 (‚Sortiername‘) einbezogen.

OSM

Für die OSM-Daten wurden sowohl in das PLZ- als auch in das City-Matching die Values zu den Tags aller Nodes einbezogen (unter Verwendung aller Keys), da sich in beiden Fällen die Bezeichner für Keys zu den Attributen nicht einfach bestimmen lassen.

Wikipedia/DBPedia

Hier wurden für das City-Matching Labels zu allen Wikipedia-Pages einbezogen. Für das Matching der Postleitzahlen wurden folgende Properties ausgewählt:

Property	**Anzahl dist. Ressourcen**	**Anzahl dist. Values**
<http://de.dbpedia.org/property/plz>	29.575	14.815
<http://de.dbpedia.org/property/postleitzahl>	54.556	21.534
<http://de.dbpedia.org/property/postalCode>	28	11

Tabelle 15: DBPedia-Properties zur Postleitzahl

Abdeckungsabschätzung WoS

Im WoS liegen die Postleitzahlen und Städtenamen bereits getrennt vor (Felder POSTALCODE und CITY in der Bibliometriedatenbank) – hier ist eine weitere Vorbereitung (Trennen von Postleitzahlen und Städtenamen) nicht erforderlich. Die City-Werte werden case-insensitive betrachtet.

distinkte CITY-Werte: 15.940 (case-insensitive)

distinke POSTALCODE-Werte: 23.524

Abdeckung	**GeoNames**	**openGeoDB**	**OSM**	**Wikipedia/ DBPedia**
PLZ	6.871 (29,21%)	5.348 (22,73%)	10.025 (42,62%)	6.107 (25,96%)
CITY	MAIN: 4.722 (29,62%)	4.795 (30,08%)	5.555 (34,85%)	6.310 (39,59%)
	ALT. NAMES: 4.139 (25,97%)			

Tabelle 16: Abdeckung WoS

Abdeckungsabschätzung Scopus

In Scopus sind zwar in einigen Fällen auch separate Postleitzahlen verfügbar, jedoch ist in der City-Spalte häufig Postleitzahl und Ortsname angegeben, so dass hier zunächst die Postleitzahl vom Ortsnamen getrennt werden muss, um die gewünschte Abschätzung durchführen zu können. Dabei wurden hier nur fünfstellige Postleitzahlen betrachtet und mit *regexp_replace* extrahiert: *regexp_replace(city,'.*D?E?-? ?([0-9]{5,5}).*','\1').* Analog wurde für den Ortsnamen die Postleitzahl aus dem City-String entfernt: *regexp_replace(city,'D?E?-? ?([0-9]{5,5})','')* (mit anschließender Weiterbearbeitung wie z.B. Löschung von Leerzeichen an Anfang und Ende des Strings).

Das Matching mit den verschiedenen Quellen wurde mit den so bearbeiteten PLZ und City-Werten durchgeführt.

distinkte CITY-Werte: 49.229 (case-insensitive, nach Extraktion)

distinke POSTALCODE-Werte: 15.434 (nach Extraktion)

Abdeckung	**GeoNames**	**openGeoDB**	**OSM**	**Wikipedia/DBPedia**
PLZ	4.684 (30,35%)	5.892 (38,18%)	10.424 (67,54%)	6.560 (42,50%)
CITY	MAIN: 4.292 (8,72%)	5.978 (12,14%)	7.739 (15,72%)	8.576 (17,42%)
	ALT. NAMES: 4.197 (8,53%)			

Tabelle 17: Abdeckung Scopus

Genauigkeit der Geokoordinaten

Die verwendeten Datenquellen sollten eine ausreichend große Genauigkeit in der Angabe der Geokoordinaten – also eine ausreichende Anzahl an Nachkommastellen in der Dezimaldarstellung – aufweisen. Die folgenden Tabellen zeigt den Zusammenhang von der Anzahl der Nachkommastellen und der Genauigkeit[26] von Breitengrad und Längengrad:

Breitengrad:

Anzahl Nachkommastellen	Genauigkeit in Metern
1	11.112
2	1.111
3	111
4	11,1
5	1,11
6	0,111

Tabelle 18: Anzahl Nachkommastellen und Genauigkeit (Breitengrad)

Längengrad (unter der Berücksichtigung des Breitengrades für Deutschland):

Anzahl Nachkommastellen	Genauigkeit in Metern
1	7111,68
2	711,04
3	71,04
4	7,104
5	0,7104
6	0,07104

Tabelle 19: Anzahl Nachkommastellen und Genauigkeit (Längengrad)

Für Auswertungen auf Stadt- bzw. Gemeindeebene liefern also 3-4 Nachkommastellen für den Breitengrad und auch für den Längengrad eine ausreichende Genauigkeit.

Die folgende Tabelle zeigt die Anzahl der Nachkommastellen für Geokoordinaten in den verschiedenen Datenquellen am Beispiel Breitengrad:

26 http://wiki.openstreetmap.org/wiki/DE:Genauigkeit_von_Koordinaten

Anzahl Nachkommastellen (Breitengrad)	**GeoNames**	**openGeoDB**	**OSM**	**Wikipedia/ DBPedia**
0	617	535		-
1	7.400	5.302	-	29.374
2	8.651	5.806	-	16.580
3	2.041	220	-	13.780
4	20.341	8.311	-	34.033
5	147.345	12.126	-	56.662
6	-	1.109	-	151.821
>6	-	27.438	9.113.976	247.246
Anzahl gesamt	**186.395**	**60.847**	**9.113.976**	**549.496**
> 2 Nachkommastellen (in %)	91,06%	80,87%	100%	91,64%

Tabelle 20: Anzahl Geo-Einheiten mit Anzahl Nachkommastellen in den verschiedenen Quellen

Hier wurden alle Geokoordinaten einbezogen (Geokoordinaten zu allen Entitäten und auch keine weitere Einschränkung auf Deutschland). Im Fall von DBPedia wurden die separat in der Datei ‚COORDINATES' erfassten und bereitgestellten Koordinaten betrachtet – weitere sind in den Infobox-Informationen zu finden. Zu berücksichtigen ist dabei, dass für die sinnvolle Angabe von Nachkommastellen eine entsprechende Messgenauigkeit Voraussetzung ist. Welche Messgenauigkeiten in den jeweiligen Quellen vorliegen, ist nicht klar.

Außerdem ist bei vielen Geo-Einheiten eine metergenaue Angabe nicht nötig und sinnvoll (beispielsweise im Fall von Bundesländern, Flüssen, großen Städten).

Die große Anzahl an Geo-Einheiten in OSM ist dadurch bedingt, dass hier Geo-Einheiten sehr viel kleinteiliger erfasst werden (z.B. auch Bäume, Parkbänke usw.).

Insgesamt ist erkennbar, dass in allen Datenquellen überwiegend eine ausreichende Genauigkeit gegeben ist.

Zusammenfassung Datenquellen

Insgesamt lässt sich also sowohl für das WoS als auch für Scopus feststellen, dass der Anteil der Namensvarianten von Städten, der mit einer nur geringen Vorbehandlung und einem exaktem Matching in den Geodatenquellen gefunden werden kann, eher gering ist. In Bezug auf die absoluten Zahlen können mit Ausnahme der Quelle GeoNames in allen Quellen mehr Namensvarianten aus Scopus als aus dem WoS gefunden werden – was aber schon allein darin begründet sein kann, dass in Scopus mehr Namensvarianten existieren.Denkbar ist aber auch, dass weniger Treffer aus dem WoS eine Folge der Vorstandardisierung von Adressen im WoS (Umwandlung von Umlauten und Sonderzeichen, die in den Geodatenquellen aber nicht umgewandelt sind) sein könnten. Relativ gesehen ist der Anteil der gefundenen Namensvarianten im WoS viel höher als in Scopus.

Bei den Postleitzahlen können (wegen geringerer Varianz) wie erwartet größere Anteile an der Gesamtmenge der in den Datenbanken WoS und Scopus enthaltenen Werte gefunden werden als im Fall der Städtenamen.

Eine ausreichende Genauigkeit der Geokoordinaten ist für die überwiegende Anzahl von Geo-Einheiten in allen Quellen gegeben.

Betrachtet man die Quellen im Vergleich – wobei hier wie oben bemerkt nur eine grobe Abschätzung stattfinden soll, ein exakter Vergleich würde aufwändigere Methoden erforderlich machen – , so wird deutlich, dass die beiden Quellen OSM und Wikipedia/DBPedia wesentlich mehr Daten enthalten als die besser strukturierten Quellen GeoNames und openGeoDB. Es erscheint also nicht sinnvoll, diese Quellen aufgrund nicht ausreichend strikter Strukturen auszuschließen.

Für das im Folgenden dargestellte Verfahren bedeutet dies: aus den gut strukturierten Quellen openGeoDB und/oder GeoNames werden die Basistabellen gespeist, während Informationen aus den beiden freier strukturierten Quellen Ergänzungen liefern bzw. zur Zuordnung genutzt werden.

5. Verfahren

Erstellen der Basistabellen aus den Datenquellen

Zum Erstellen der Basistabellen wurden insbesondere die Geodatenquellen GeoNames und openGeoDB genutzt. Nach Möglichkeit wurden die so erhaltenen Informationen durch weitere Daten aus OSM und Wikipedia/DBPedia ergänzt. Als Grundlage wurde dabei GeoNames genutzt, damit das Verfahren anschließend beliebig auf weitere Länder ausgeweitet werden kann.

Auswahl der aufzunehmenden Geo-Einheiten

In GeoNames ist eine Lookup-Tabelle ‚Admin1_Codes' enthalten, die (eingeschränkt auf Präfix DE) die deutschen Bundesländer als erste administrative Ebene des Landes enthält. Die hier vergebenen Codes sind DE01-DE16. Zu beachten ist, dass die Bezeichner 01 bis 16 im Gegensatz zu den weiteren Hierarchieebenen in GeoNames NICHT mit den entsprechenden Stellen (als hier ersten beiden Ziffern) des amtlichen Gemeindeschlüssels übereinstimmen.

GeoNameID	Code	Name in GeoNames
2953481	DE.01	Baden-Württemberg
2951839	DE.02	Bavaria
2944387	DE.03	Bremen
2911297	DE.04	Hamburg
2905330	DE.05	Hesse
2862926	DE.06	Lower Saxony
2861876	DE.07	North Rhine-Westphalia
2847618	DE.08	Rheinland-Pfalz
2842635	DE.09	Saarland
2838632	DE.10	Schleswig-Holstein
2945356	DE.11	Brandenburg
2872567	DE.12	Mecklenburg-Vorpommern
2842566	DE.13	Saxony
2842565	DE.14	Saxony-Anhalt
2822542	DE.15	Thuringia
2950157	DE.16	Berlin

Tabelle 21: Admin1 Codes in GeoNames

Eine weitere Tabelle ‚Admin2_Codes' enthält 19 Regionen unterhalb der Ebene der Bundesländer (Regierungsbezirke). Dabei sind nur die Bundesländer DE.01, DE.02, DE.05 und DE.07 in dieser Weise weiter unterteilt.

Diese Geo-Einheiten wurden zunächst in die Basistabellen aufgenommen. Anschließend wurden über die MAIN-Tabelle weitere administrative Einheiten niedrigerer Ebenen erfasst.

Hierarchiebeziehungen zu den jeweils höheren Ebenen wurden ebenfalls in die Basistabellen aufgenommen. Dabei wurde für jede aufgenommene Geo-Einheit eine neue/eigene ID (unabhängig von der GeoNames-ID) vergeben – so können auch unabhängig von GeoNames weitere Geo-Einheiten aus anderen Quellen aufgenommen werden. Die GeoNames-ID sowie die GeoNames-Admin-Codes wurden jedoch in der Tabelle GEO_U_IDENTIFIER erfasst, so dass eine Verlinkung der Daten in den Basistabellen mit den GeoNames-Daten erhalten bleibt.

Identifier

In die Tabelle GEO_U_IDENTIFIER sollten sowohl allgemeine Identifier wie der amtliche Gemeindeschlüssel (AGS) als auch Identifier der genutzten Quellen (Wikipedia-Page-ID, LOC ID aus openGeoDB, GeoNames-ID und OpenStreetMap-ID) aufgenommen werden.

Allgemeine Identifier

Der amtliche Gemeindeschlüssel (für Gemeinden) und die entsprechenden Teilschlüssel für Bundesländer und Kreise können aus GeoNames bezogen werden. Es besteht die Möglichkeit, in diese Tabelle weitere Identifier für Geo-Einheiten aufzunehmen (wie beispielsweise NUTS-Codes), sofern diese in den hier genutzten (oder weiteren externen) Quellen enthalten sind.

Identifier aus den Quellen GeoNames, openGeoDB, BDPedia, OSM

Soweit möglich wurden Identifier aus den Geodatenquellen erfasst (GeoNames-ID, openGeoDB-ID, URI aus DBPedia, Page-ID aus Wikipedia, OpenStreetMap-ID) – zum einen, um verschiedene Quellen im Verfahren nutzen zu können, zum anderen, um die aus der Geokodierung resultierenden Daten mit Daten aus anderen Datenquellen verlinken zu können.

Zu jeder Geo-Unit wurde zunächst die GeoNames-ID erfasst, da die Daten aus GeoNames den Grundbestand der Basisdaten stellen und die GeoNames-ID damit für nahezu alle aufgenommenen Geo-Einheiten zur Verfügung steht.

Bestehende Verknüpfungen:

Zum Teil sind in den GeoDatenquellen selbst schon Verlinkungen zu anderen Datenbeständen vorhanden, die hier genutzt werden können.

In OSM existieren Tags, die eine Verknüpfung zu openGeoBD (Tag: openGeoDB:loc_id) und zu Wikipedia (Tag: wikipedia und wikipedia:de) enthalten. Für openGeoDB sind Location IDs enthalten, während zum Tag ‚wikipedia‘ Values der Form ‚DE:<label>‘ enthalten sind. Hier ist also nicht direkt ein Identifier wie die Page_id oder die URI angegeben, sondern es muss über ‚Label‘ gematcht werden, um entsprechende URI und Page-Ids zu erhalten. Die in OSM vorhandenen Links auf Gemeindeebene sind nicht immer eindeutig, aber Stichproben haben gezeigt, dass diese (intellektuell gesetzten) Verlinkungen auch dann wertvoll sind, wenn sie nicht eindeutig sind. In diesen Fällen werden beispielsweise verschiedene Ortsteile mit einer Gemeinde verlinkt. Hier kann

es sich um Historisierungsfälle (wie z.B. Eingemeindungen) oder Hierarchien handeln. Eine sinnvolle Beziehung zwischen den Entitäten ist aber trotzdem gegeben.

Links zu GeoNames konnten nicht gefunden werden.

Links in OSM...	...zu openGeoDB	...zu wikipedia/DBpedia
	11.682	wikipedia 20.429 wikipedia:de 142

Tabelle 22: Bestehende Verlinkungen von OSM zu openGeoDB und Wikipedia (insgesamt, nicht auf Gemeindeebene eingeschränkt)

In GeoNames sind in der Tabelle ‚Alternate Names' Links zu Wikipedia-Ressourcen enthalten (Typ='link'), die sich leicht in DBPedia-Ressourcen umformen lassen (UTL_URL.UNESCAPE, Ersetzen von ‚wikipedia' durch ‚dbpedia' und ‚/wiki/' durch ‚/resource/').

In DBPedia sind sehr wenig Links zu GeoNames in einer gesonderten Tabelle angegeben (802 insgesamt, nur sehr wenige davon zu deutschen Geo-Entitäten).

In openGeoDB konnten keine Links zu den jeweils anderen Quellen identifiziert werden.

Zusätzlich erstellte Verlinkungen

Da der amtliche Gemeindeschlüssel (kurz AGS, wird im Folgenden noch näher vorgestellt) ein Identifier auf Gemeindeebene ist, können Datensätze in den verschiedenen Quellen auf Gemeindeebene über diesen Identifier verlinkt werden. Voraussetzung dafür ist zum einen die Existenz des AGS in den zu verlinkenden Quellen und zum anderen ein ‚Typ' der Entität, über den diese als Gemeinde identifiziert werden kann (da DBPedia verschiedene Typen von Entitäten – und insbesondere nicht nur Geo-Einheiten – enthält).

Typ der Entität: Die in GeoNames gefundenen Gemeinden sind bereits mit einem entsprechenden Typ in die Tabelle GEO_UNIT aufgenommen worden und können entsprechend einfach identifiziert werden. In openGeoDB lassen sich Gemeinden über text_type=400300000 und (text_val='Gemeinde' oder text_val='Stadt') identifizieren. Anschließend können die als Gemeinden identifizierten Entitäten über den AGS gematcht werden.

In DBPedia lassen sich Gemeinden nicht auf einfache Weise identifizieren. Hier wird nur über den AGS gematcht – dabei müssen wegen des Datentypen-Problems (AGS sind als Integer erfasst) im Fall von siebenstelligen AGS jeweils führende Nullen ergänzt werden. In allen Fällen der zusätzlich erstellten Verlinkungen werden Treffer/Verlinkungen nur dann verwertet, wenn sie eindeutig sind (nur Entitäten, die auf genau einen AGS abgebildet werden und ein AGS nur auf genau eine Entität abgebildet werden kann), um fehlerhafte Verlinkungen zu vermeiden.

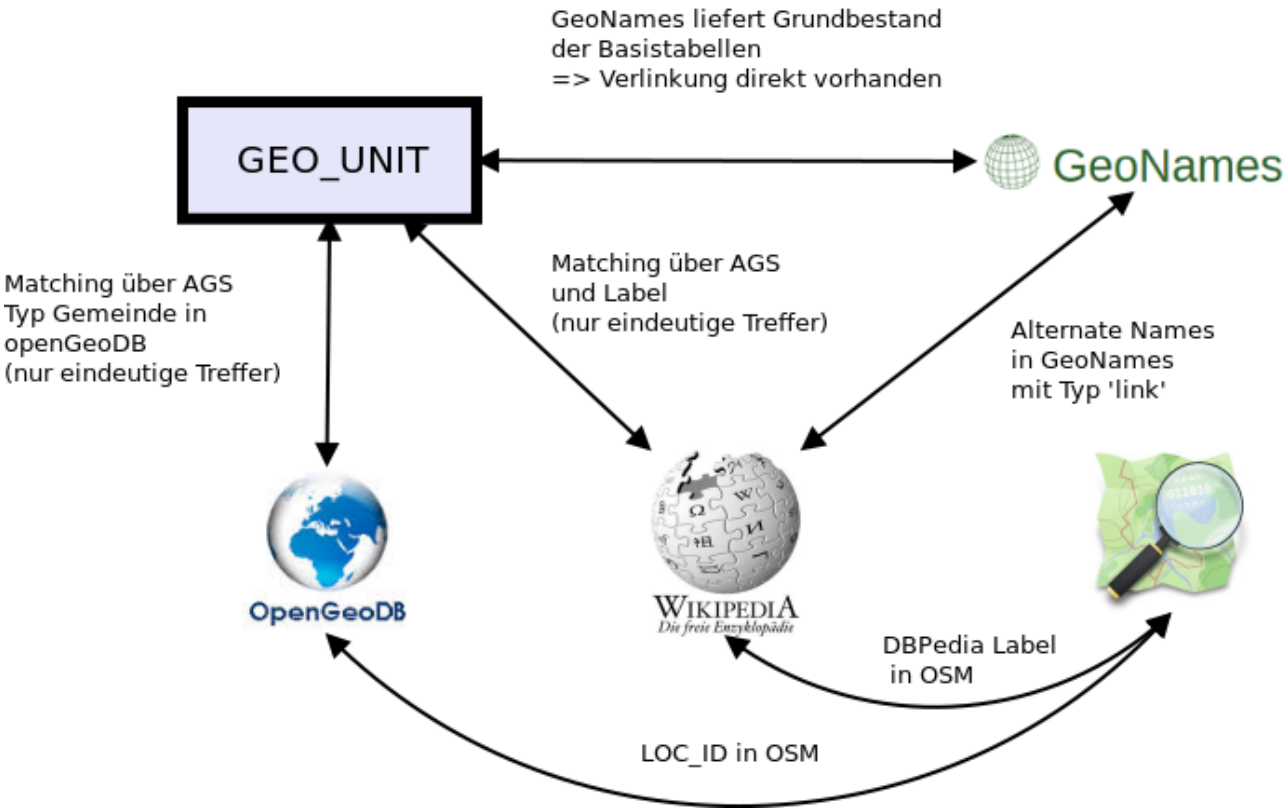


Abbildung 4: Identifier Matching

Auf diese Weise gefundene Verlinkungen wurden in die Tabelle GEO_U_IDENTIFIER aufgenommen (Links zu GEO_UNIT).

Durch Kombinationen von erstellten und in den Daten gegebenen Verlinkungen können weitere Identifier zu Geo-Units in den Basistabellen gefunden werden (beispielsweise ermöglicht eine Kombination der erstellten Verlinkungen zwischen DBPedia und GEO_UNIT und die in OSM gegebene Verklinkung zu DBPedia auch eine Verlinkung von GEO_UNIT und OSM).

Insgesamt sind in der Tabelle GEO_U_IDENTIFIER folgende Typen von Identifiern vorhanden:

Identifier-Typ	Anzahl verlinkter Geo-Units
GeoNameID	11.993
AGS, 2stellig (Ebene: Bundesland)	16
AGS, 3stellig (Ebene: Regierungsbezirk)	22
AGS, 5stellig (Ebene: Kreis)	416
AGS, vollständig (Ebene: Gemeinde)	11.499
DBPedia Ressource (URI)	11.109
Wikipedia Page ID	11.109
openGeoDB ID	9.690
OSM ID	8.102
GeoNames ADMIN1 Codes	19
GeoNames ADMIN2 Codes	22

Tabelle 23: Identifier-Typen mit Häufigkeit

Amtlicher Gemeindeschlüssel (AGS)[27]

Für die Zuordnung von Adressen zu Geo-Einheiten muss – analog zur Definition der Hauptinstitutionen in der Institutionenkodierung – eine Aggregationsebene gefunden werden, auf der die Zuordnung erfolgen soll. Die Gemeinde-Ebene scheint hier eine passende Aggregationsebene zu sein.

Diese lässt sich jedoch weder durch Orts- bzw. Städtenamen (gleiche Namen für verschiedene Städte, z.B. Frankfurt, Neustadt oder Neuenkirchen) noch durch Postleitzahlen (Postleitzahlengebiete können sich über mehrere Gemeinden erstrecken) allein definieren.

Kombinationen aus Postleitzahlen und Ortsnamen definieren ebenfalls nicht die Gemeinde-Ebene, sondern Teilbereiche. Diese Aggregationsebene ist zu kleinteilig, definiert also auch nicht in sinnvoller Weise eine Gemeinde (mehrere Postleitzahlen für die Gemeinde Bielefeld, wobei aber nicht ‚33615 Bielefeld' und ‚33619 Bielefeld' als unterschiedliche Klassifikationsziele betrachtet werden, sondern Zuordnungen zur gesamten Gemeinde Bielefeld erfolgen sollen).

Eine Definition der Gemeinde-Ebene liefert der 8-stellige amtliche Gemeindeschlüssel (AGS), wobei die ersten beiden Ziffern das Bundesland, die dritte bis fünfte Ziffer den Kreis bzw. die kreisfreie Stadt und die letzten drei Ziffern die Gemeinde innerhalb des Kreises bezeichnen.

27 https://de.wikipedia.org/wiki/Amtlicher_Gemeindeschl%C3%BCssel

Beispiel:

03 2 54 021 = Hildesheim

- *03 Niedersachsen*
- *2 ehemaliger Regierungsbezirk Hannover*
- *54 Landkreis Hildesheim*
- *021 Stadt Hildesheim*

Zusammenhang zwischen Postleitzahl, Ortsnamen und AGS

Zuordnungen von Postleitzahlen und Ortsnamen zu amtlichen Gemeindeschlüsseln müssen nicht eindeutig sein, Kombinationen von Postleitzahlen und Ortsnamen liefern dagegen eine eindeutige Zuordnung zu einem AGS (eine Kombination aus Postleitzahl und Ortsnamen kann genau einem AGS zugeordnet werden).

Aufgrund von unterschiedlichen Städten gleichen Namens kann ein Ortsname dagegen mehreren amtlichen Gemeindeschlüsseln zugeordnet werden (Beispiel: Neuenkirchen).

Ein Beispiel für eine Postleitzahl, die sich über mehrere Gemeinden erstreckt, ist die ‚01945‘:

PLZ	Ortsname	AGS
01945	Guteborn	12066120
01945	Grünewald	12066116
01945	Hermsdorf	12066124
01945	Ruhland	12066272
01945	Kroppen	12066168
01945	Lindenau	12066188
01945	Hohenbocka	12066132
01945	Tettau	12066316
01945	Schwarzbach	12066292

Tabelle 24: Verschiedene AGS zur Postleitzahl 01945

Postleitzahlen mit AGS-Zuordnung

In allen Quellen sind auch Postleitzahlen enthalten – jedoch sind diese hier nur dann verwertbar, wenn Zuordnungen zum entsprechenden amtlichen Gemeindeschlüssel (AGS) vorliegen. Postleitzahlen und deren Zuordnung zum AGS müssen aus den Quellen in unterschiedlicher Weise extrahiert werden:

In GeoNames ist eine separate Tabelle mit Postleitzahlen, weiteren Attributen und zugehörigem fünfstelligen (also nicht dem vollständigen achtstelligen Gemeindeschlüssel, sondern dem fünfstelligen Kreis-Teilschlüssel) AGS vorhanden – das heißt, der oder die vollständige(n) AGS eines Postleitzahlengebietes lassen sich nicht ohne weiteres extrahieren. Daher wurde hier für die Fälle, in denen der Ortsname und der fünfstellige AGS übereinstimmen, der vollständige AGS aus der MAIN-Tabelle ergänzt. Um Fehler zu vermeiden, wurde diese Zuordnung nur für Ortsnamen durchgeführt, die in dem entsprechenden Kreis (definiert durch den fünfstelligen AGS) eindeutig sind.

In openGeoDB wurden die Location_Ids ausgewählt, für die sowohl der text_type 500300000 (PLZ) als auch der text_type 500600000 (AGS) existieren und die zugehörigen Values zu einer Liste aus PLZ und zugehörigen AGS zusammengefügt.

In DBPedia ist nicht unmittelbar klar, mit welchen Properties die Postleitzahlen und AGS erfasst sind. Zur Identifikation der passenden Properties wurden daher zunächst alle Properties zu Values, die auch als Postalcode in GeoNames vorkommen (diese müssen nicht unbedingt Postleitzahlen sein – es kann sich beispielsweise auch um Flächen, Einwohnerzahlen oder ganz anderes handeln, die hier keine Rolle spielen), mit ihren Häufigkeiten ermittelt:

Property	**Anzahl**
<http://de.dbpedia.org/property/postleitzahl>	22.123
<http://de.dbpedia.org/property/plz>	13.151
<http://de.dbpedia.org/property/adresseVerband>	4.537
<http://de.dbpedia.org/property/adresse>	3.684
<http://de.dbpedia.org/property/objektid>	2.622
<http://de.dbpedia.org/property/nummer>	2.154
<http://de.dbpedia.org/property/insee>	1.254
<http://de.dbpedia.org/property/cp>	1.232
<http://de.dbpedia.org/property/einwohner>	1.126
<http://de.dbpedia.org/property/ew>	978

Tabelle 25: Top10-Properties in DBPedia für Values, die als Postalcode in GeoNames vorkommen

Analog für den AGS:

Property	Anzahl
<http://de.dbpedia.org/property/gemeindeschlüssel>	2.721
<http://de.dbpedia.org/property/ew>	23
<http://de.dbpedia.org/property/gnd>	21
<http://de.dbpedia.org/property/längengrad>	12
<http://de.dbpedia.org/property/eintragung>	12

Tabelle 26: Properties in DBPedia für Values, die als AGS in GeoNames vorkommen

Also wurden die Properties <http://de.dbpedia.org/property/plz> und Properties LIKE ‚<http://de.dbpedia.org/property/postleitzahl%' für Postleitzahlen sowie <http://de.dbpedia.org/property/gemeindeschlüssel> für den AGS verwendet. Wie bei openGeoDB wurden nur Ids ausgewählt, für die Values für eine Property für PLZ (wie angegeben) und für die für den AGS ausgewählte Property vorliegen. Außerdem muss in DBPedia eine Einschränkung auf ‚deutsche' Entitäten erfolgen, da hier zwar die deutschsprachige DBPedia verwendet wird, diese aber natürlich auch Entitäten (und zugehörige Postleitzahlen) aus anderen Ländern enthält. Diese Einschränkung sollte möglichst nicht zu viel ausschließen und wurde daher auf Ids vorgenommen, die Values LIKE '%DEUTSCHLAND%' besitzen (case-insensitive). Problematisch ist dabei, dass sowohl Postleitzahl als auch AGS in der deutschen DBPedia mit dem Datentyp Integer erfasst werden. Führende Nullen fehlen also in beiden Fällen. Während sich das Problem beim AGS mit der Ergänzung einer führenden Null im Fall eines siebenstelligen AGS recht einfach beheben lässt, können vierstellige Postleitzahlen sowohl alte Postleitzahlen sein als auch solche, bei denen die führende Null fehlt.

In OSM erscheinen die Tag-Keys zu Postleitzahlen recht unübersichtlich (viele verschiedene Bezeichnungen) – außerdem ist unklar, ob eine Einschränkung auf Deutschland unproblematisch erfolgen kann. Ein möglicher Tag-Key für die Angabe von Postleitzahlen ist ‚openGeoDB:postal_codes' – es werden also Postleitzahlen aus openGeoDB in OSM verwendet.

Analog zu der bei DBPedia verwendeten Vorgehensweise können aber auch in OSM wichtige zu Postleitzahl- bzw. AGS gehörigen Tag-Keys identifiziert werden.

PLZ:

Tag-Key	Anzahl
addr:postcode	2.268.157
object:postcode	15.654
postal_code	12.892
openGeoDB:postal_codes	9.459
TMC:cid_58:tabcd_1:LocationCode	1.487
VRS:ref	1.411
TMC:cid_58:tabcd_1:PrevLocationCode	1.385
TMC:cid_58:tabcd_1:NextLocationCode	1.343
memorial:addr:postcode	1.084
openGeoDB:loc_id	977

Tabelle 27: Top10-Tag-Keys in OSM für Values, die als Postalcode in GeoNames vorkommen

AGS:

Tag-Key	Anzahl
openGeoDB:community_identification_number	9.481
de:amtlicher_gemeindeschluessel	758
community_identification_number	22
openGeoDB:old_community_identification_number	4

Tabelle 28: Top10-Tag-Keys in OSM für Values, die als AGS in GeoNames vorkommen

Die folgende Tabelle zeigt die Anzahl der jeweils aus den Quellen auf die beschriebene Weise extrahierten PLZ und PLZ-AGS-Kombinationen:

	openGeoDB	GeoNames	DBPedia	OSM
Anzahl extrahierter PLZ	8.296	6.691	6.611	6.540
– davon 5stellig	8.296	6.691	6.104[28]	6.532
Anzahl extrahierter PLZ – AGS – Kombinationen	14.940	10.485	11.844	11.154

Tabelle 29: Anzahl extrahierter Postleitzahlen mit zugehörigem AGS

28 Fast alle verbleibenden vierstellig, bis auf 9 Ausnahmen – Formfehler.

Dabei kommen 6.746 PLZ-AGS-Kombinationen in allen vier Quellen vor, 8.111 PLZ-AGS-Kombinationen kommen in allen außer OSM vor, 10.121 in openGeoDB und in GeoNames, DBPedia und GeoNames haben 8.132 PLZ-AGS-Kombinationen gemeinsam, DBPedia und openGeoDB teilen 9.397 PLZ-AGS-Kombinationen. Es stellt sich die Frage, ob die Kombinationen, die jeweils nicht in anderen Quellen vorkommen, deshalb nicht vorkommen, weil die PLZ in der Quelle überhaupt nicht existiert oder weil in den Quellen unterschiedliche Zuordnungen vorliegen. Da DBPedia nicht das Ziel verfolgt, eine vollständige Liste von PLZ- AGS-Kombinationen zu liefern, ist es nicht verwunderlich (und auch kein Fehler der Datenquelle), dass dort nicht zu jeder Postleitzahl alle AGS extrahiert werden können. Anders verhält es sich mit den Quellen GeoNames und openGeoDB, die den Anspruch haben, möglichst vollständige Geodaten (einschließlich der Postleitzahlen) strukturiert zu erfassen.

12 der in openGeoDB gefundenen Postleitzahlen sind nicht in der Postleitzahlentabelle von GeoNames enthalten, 27 der in GeoNames enthaltenen Postleitzahlen sind nicht in openGeoDB enthalten (in diesen Zahlen sind alle Postleitzahlen aus GeoNames enthalten – also auch die, die keinem AGS zugeordnet werden konnten).

Eine Einzelüberprüfung in der PLZ-Onlinesuche der Deutschen Post[29] (als Referenz) für diese Postleitzahlen ergibt:

	PLZ in openGeoDB, aber nicht in GeoNames	**PLZ in GeoNames, aber nicht in openGeoDB**
Anzahl insgesamt	**12**	**27**
– davon nicht gefunden	6	6
– davon ‚normale' PLZ	4	4
– davon Postfach-PLZ	2	17

Tabelle 30: PLZ Einzelrecherche in der PLZ-Onlinesuche der Deutschen Post

Bei den nicht gefundenen Postleitzahlen könnte es sich um nicht oder nicht mehr gültige Postleitzahlen handeln. Zum Teil werden sie in diversen Quellen verwendet und einem Ort zugewiesen, wie zum Beispiel die Postleitzahl 96529, die über den Online Service der Post nicht gefunden werden kann und nicht in GeoNames, jedoch in openGeoDB gelistet ist. Postleitzahlen zu Postfächern können auch in den zuzuordnenden Adressen enthalten sein und bieten damit eine wertvolle Information. Aufgrund der geringen Menge der über den Online Service der Post nicht gefundenen Postleitzahlen und der Möglichkeit, dass es sich dabei um veraltete Postleitzahlen handelt (die ebenfalls von Bedeutung für die Zuordnung – insbesondere alter Adressen – sind), wurden diese nicht ausgeschlossen. Insgesamt ist also die Übereinstimmung zwischen GeoNames und openGeoDB in Bezug auf die enthaltenen Postleitzahlen recht hoch.

29 https://www.postdirekt.de/plzserver/

Vergleicht man die PLZ-AGS Zuordnungen in GeoNames und openGeoDB für Postleitzahlen, die in beiden Datenquellen enthalten sind, so sind 10.121 Kombinationen in beiden Datenquellen enthalten, 361 sind in GeoNames enthalten, nicht aber in openGeoDB während 2.270 Kombinationen in openGeoDB enthalten sind, nicht aber in GeoNames.

Aus beiden Mengen (in einer Quelle, aber nicht in der anderen) wurden je 10 Zuordnungen zufällig gezogen und manuell überprüft (über das Statistikportal des statistischen Bundesamtes[30] lässt sich der Gemeindename zu einem AGS unabhängig von den hier verwendeten Quellen recherchieren – dieses dient als Referenz, der oben erwähnte Online Service der Post liefert die Gemeinde zur Postleitzahl).

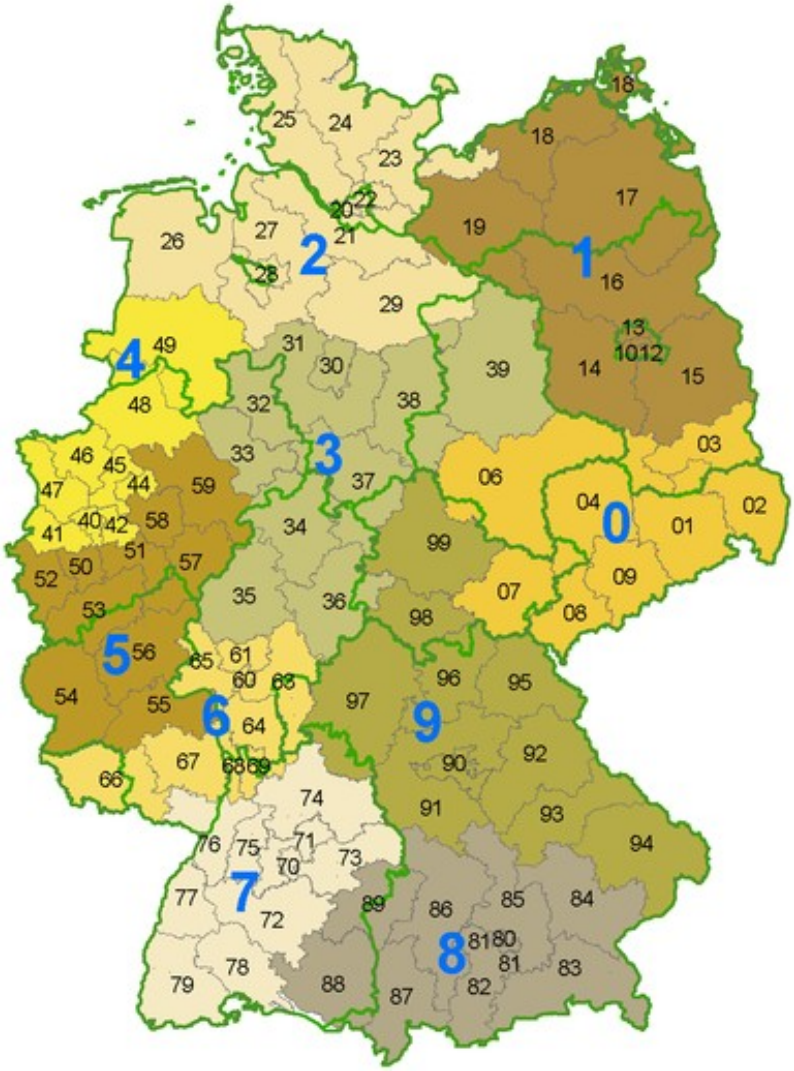


Abbildung 5: Postleitzahlengebiete in Deutschland (Quelle: Wikipedia)

30 http://www.statistik-portal.de/Statistik-Portal/gemeindeverz.asp?G=14625320

Von den 10 Zuordnungen, die in openGeoDB, aber nicht in GeoNames enthalten sind, konnte der gegebene AGS in 9 Fällen nicht im Gemeindeverzeichnis gefunden werden (ist also nicht bzw. nicht mehr gültig), in einem Fall war die zugeordnete PLZ nicht unter den im Online Service der Post zu dieser Gemeinde verfügbaren Postleitzahlen (diesen jedoch sehr ähnlich), dagegen konnte der größte Teil der 10 Zuordnungen, die in GeoNames enthalten sind – nicht aber in openGeoDB – als korrekte identifiziert werden. Nur in einem Fall konnte der AGS nicht gefunden werden.

Postleitzahlen wurden (mit ihrer Eigenschaft als ‚Postleitzahlengebiet') als Geo-Units aufgenommen und über die Verknüpfung mit dem AGS in der Tabelle GEO_RELATIONS mit TYPE ‚ist Postleitzahl von' bzw. ‚hat Postleitzahl' mit der entsprechenden Gemeinde verknüpft. Diese Zuordnungen müssen nicht eindeutig sein, da sich Postleitzahlengebiete über mehrere Gemeinden erstrecken und in einer Gemeinde mehrere Postleitzahlen vorkommen können. Ebenso können Postleitzahlen auf diese Weise mit anderen Geo-Units verlinkt werden (beispielsweise mit Kreisen oder Bundesländern).

Postleitzahlen (mit zugehörigen AGS) wurden wie folgt in die Basistabellen aufgenommen:

1. Aufnahme aller PLZ mit zugehörigen AGS aus GeoNames
2. Aufnahme aller PLZ mit zugehörigen AGS aus openGeoDB, die nicht bereits aufgenommen sind
3. Aufnahme aller PLZ mit zugehörigen AGS aus DEDBPedia (mit Einschränkung auf fünfstellige PLZ, siebenstellige AGS wurden um eine führende 0 ergänzt), sofern nicht bereits aufgenommen
4. Aufnahme aller PLZ mit zugehörigen AGS aus OSM (mit Einschränkung auf fünfstellige PLZ), sofern noch nicht bereits aufgenommen

Auf diese Weise wurden folgende Anzahlen von Postleitzahlen aus den jeweiligen Quellen entnommen:

Quelle	**Anzahl aufgenommener Postleitzahlen**
GeoNames	6.691
openGeoDB	1.608
DBPedia	132
OSM	7
Insgesamt	**8.438**

Tabelle 31: Anzahl aufgenommener Postleitzahlen nach Quellen

Namen

Die namentlichen Bezeichnungen für alle Geo-Units wurden zunächst aus GeoNames bezogen. Die Tabelle GEO_U_NAME enthält nur einen jeweils gültigen Namen in einem Zeitfenster. Alte Namen können erfasst werden, nicht aber mehrere Namensvarianten.

Weitere Namensvarianten wurden im Zuordnungsverfahren verwendet und in eine andere Tabelle ausgelagert. Die Tabelle GEO_U_NAME dagegen ist *einer* Namensvariante je Geo-Unit und Zeitfenster vorbehalten, die dann in Auswertungen als Bezeichnung genutzt werden kann.

Relationen

Die Hierarchiebeziehungen für die Geo-Units aus GeoNames wurden den Spalten Admin1-4 der Haupttabelle entnommen und in die gewünschte Struktur überführt. Es wurden dabei verschiedene Relationstypen verwendet (beispielsweise ‚ist Bundesland von' statt allgemein ‚ist Teil von').

Evaluation mit dem Gemeindeverzeichnis des Statistischen Bundesamtes[31]

GeoNames bietet also eine Grundlage zur Erstellung von Basistabellen. Diese Datenbasis kann für alle Länder genutzt werden. Je nach verfügbaren Datenquellen für das jeweilige Land ist es jedoch denkbar, weitere Datenquellen heranzuziehen, um die Vollständigkeit der GeoNames – Daten zu evaluieren oder diese zu ergänzen.

Für Deutschland wurde hier das Gemeindeverzeichnis des Statistischen Bundesamts herangezogen, um die Bezeichnungen für Bundesländer und Gemeinden zu überprüfen und ggf. zu korrigieren. Weiter können die Geokoordinaten und die Vollständigkeit in Bezug auf die Gemeindeschlüssel anhand dieser offiziellen Quelle überprüft werden. Postleitzahleninformationen lassen sich über diese Quelle nicht überprüfen, da jeweils nur eine einzige Postleitzahl je Gemeinde angegeben ist.

Vollständigkeit der erfassten Gemeinden (über AGS)

11.083 der 11.162 im Gemeindeverzeichnis des statistischen Bundesamtes (im Folgenden mit GV abgekürzt) sind in GeoNames (und damit in der Basistabelle GEO_U_IDENTIFIER mit zugehöriger Geo-Unit) erfasst; damit erreicht GeoNames eine Abdeckung von 99% und liefert an dieser Stelle eine verlässliche Datenbasis.

Allerdings sind 415 AGS in GeoNames enthalten, die lt. GV nicht (oder nicht mehr) gültig sind. In Stichproben zeigen sich dabei Gemeinden, die tatsächlich in GeoNames als ‚historisch' geflaggt sind (zum Beispiel GeoName-ID=6552243, lt. Wikipedia war die Gemeinde Siezbüttel bis zum 01.01.2013 eine eigenständige Gemeinde, jetzt ist sie Ortsteil von Schenefeld[32]). Für das Zuordnungsverfahren können solche historischen Einheiten auch von Interesse sein, deshalb wurden hier keine AGS aus den Basisdaten entfernt. Historische Einheiten wurden stattdessen mit ‚HISTORICAL=1' in GEO_UNIT geflaggt. So können historische Einheiten als solche identifiziert und ggf. ausgeschlossen werden. Eine umfassende Historisierung (die beispielsweise die Abbildung aller Eingemeindungsvorgänge mit Zeitstempel enthalten würde) wurde hier jedoch nicht angestrebt.

31 https://www.destatis.de/DE/ZahlenFakten/LaenderRegionen/Regionales/Gemeindeverzeichnis/Administrativ/Archiv/GVAuszugQ/AuszugGV3QAktuell.html

32 https://de.wikipedia.org/wiki/Siezb%C3%BCttel

Namen/Bezeichnungen für Gemeinden

Mit 78% ist der größte Teil der Gemeinden mit dem exakt gleichen Namen in GeoNames und GV erfasst. In 15% der Fälle ist im GV dem Gemeindenamen der Zusatz ‚Stadt' angehängt, in 3% ist ein anderer Zusatz gesetzt, in weiteren 3% ist der Zusatz nicht im GV vorhanden, sondern in GeoNames. In dem verbleibenden 1% sind ebenfalls nur geringe Abweichungen in den Bezeichnungen zu finden, wie beispielsweise in den folgenden Fällen:

- Utting a.Ammersee ↔ Utting am Ammersee
- Dratow-Schloen ↔ Schloen-Dratow
- Zell, Markt ↔ Zell im Fichtelgebirge, M

Offensichtlich hält sich GeoNames in Bezug auf die Gemeinden sehr nah an die offiziellen Bezeichnungen.

Namen/Bezeichnungen für Bundesländer

Wie bereits Tabelle 21 zeigt, ist die Situation in den Bezeichnungen für die Admin1-Codes eine andere. Hier werden zum Teil deutsche und zum Teil englische Namensvarianten verwendet und es gibt Abweichungen von den an offizieller Stelle verwendeten Namen. Daher wurden diese Bezeichnungen manuell korrigiert, d.h. durch einheitlich deutsche Bezeichnungen aus dem GV ersetzt.

Geo-Koordinaten auf Gemeinde-Ebene

Die beiden folgenden Tabellen dienen dem Vergleich der Genauigkeit des Breiten- bzw. Längengrades anhand der Anzahl der Nachkommastellen für die AGS, die sowohl in GeoNames als auch im Gemeindeverzeichnis des Statistischen Bundesamtes (GV) mit Geokoordinaten enthalten sind:

# Nachkommastellen	Anzahl Gemeinden GeoNames	Anzahl Gemeinden Statistisches Bundesamt
0	96	-
1	962	-
2	1.044	1
3	196	13
4	3.694	104
5	5.092	995
6	-	9.971

Tabelle 32: Vergleich Genauigkeit der Geokoordinaten, Breitengrad

# Nachkommastellen	Anzahl Gemeinden GeoNames	Anzahl Gemeinden Statistisches Bundesamt
0	84	1
1	984	1
2	1.053	13
3	179	105
4	6.016	988
5	2.768	9.976

Tabelle 33: Vergleich Genauigkeit der Geokoordinaten, Längengrad

Obwohl der überwiegende Teil der Geo-Koordinaten in GeoNames über ausreichende Genauigkeit verfügt und somit verwendet werden könnte, ist doch deutlich erkennbar, dass in den Daten des statistischen Bundesamtes insgesamt eine größere Genauigkeit gegeben ist (entsprechende Messgenauigkeit vorausgesetzt). Die Koordinaten aus GeoNames auf Gemeinde-Ebene könnten durch die Daten des statistischen Bundesamtes ersetzt werden – für eine Ausweitung auf andere Länder wäre aber evtl. eine solche zusätzliche und bessere Datenquelle nicht unbedingt vorhanden.

Entscheidend ist hier, inwiefern sich die durch die Koordinaten des Statistischen Bundesamtes und die durch die Koordinaten in GeoNames gegebenen Punkte unterscheiden, d.h. wie weit diese voneinander entfernt sind.

Die folgende Tabelle zeigt die Entfernungen von durch die Geokoordinaten des statistischen Bundesamtes vs. von GeoNames gegebenen Punkte zu einer Gemeinde:

Entfernung (Luftlinie) gerundet auf km	Anzahl Gemeinden
0	4.464
1	5.430
2	760
3	241
4	110
5	38
6	18
7	13
8	5
9	1
10	1
11	2
12	-
13	1

Tabelle 34: Entfernung der durch die Geokoordinaten des Stat. Bundesamt und GeoNames definierten Punkte auf Gemeindeebene

Die Entfernungen sind also ganz überwiegend im tolerierbaren Bereich. Beim statistischen Bundesamt sind jeweils Mittelpunktskoordinaten gegeben, bei GeoNames ist das nicht klar dokumentiert.

Vollständigkeit der erfassten PLZ

Die Deutsche Post bietet kostenpflichtige Datenpakete[33], u.a. zu Postleitzahlen an. Diese sollten nicht unmittelbar als Quelle genutzt werden, da sie erstens nicht frei (kostenpflichtig und Einschränkungen durch Lizenzbestimmungen) sind und zweitens auf Deutschland begrenzt, so dass eine Erweiterung auf andere Länder nicht möglich wäre. Sinnvoll ist dagegen die Nutzung zur Überprüfung der Vollständigkeit der Quellen in Bezug auf die Postleitzahlen (hier: alle Postleitzahlen deutschlandweit, nicht nur die im Web of Science vorhandenen – außerdem verschiedene Arten wie Postleitzahlen für Postfächer, Großkunden, Zustellung usw.). Die folgende Tabelle zeigt das Ergebnis einer solchen Prüfung.[34]

33 https://www.deutschepost.de/en/d/deutsche-post-direkt/datafactory.html
34 Genutzt wurde hier das Datenpaket ‚DATAFACTORY BASIC', Stand 2014.

Art	Anzahl PLZ DATA-FACTORY	...auch in GeoNames	...auch in openGeoDB	...auch in DEDBPedia	...auch in OSM	...auch in GEO_UNIT nach Verfahren
Zustellung (in % von Anzahl in DATA-FACTORY)	7.904	6.391 (80,86%)	7.881 (99,71%)	5.675 (71,80%)	6.210 (78,57%)	7.890 (99,82%)
Zustellung und Postfach	303	259	303	301	270	303
Postfach	16.510	13	26	116	4	142
Einzel-Großempfänger	2.328	3	7	2	2	10
Gruppen-Großempfänger	795	1	2	1	2	4
Aktions-PLZ	2.032	1	1			1

Tabelle 35: Vollständigkeit der PLZen in den Geodatenquellen

In Bezug auf Postfach und Großempfänger sind damit jeweils nahezu keine Postleitzahlen vorhanden. In DBPedia könnten weitere Postleitzahlen enthalten sein, wenn man keine Eingrenzung wie hier vornimmt (nur Resourcen, die gleichzeitig auch einen AGS haben). Bezüglich der Postleitzahlen für die Zustellung hat openGeoDB mit 99% eine sehr gute und die beste Abdeckung aller Quellen.

Nur jeweils sehr wenig Postleitzahlen sind in den Quellen enthalten, nicht aber in den Daten der Post (GeoNames: 23, openGeoDB: 76, DEDBPedia: 516 (nur fünfstellige: 9, siehe dazu oben beschriebenes Problem Datentyp), OSM: 52 (davon 44 fünfstellig – in einer Zufallsauswahl von 10 der 44 PLZ konnte keine über die PLZ-Suche der Post gefunden werden)).

Hier kann es sich um Fehler oder aber auch um Änderungen handeln (Stand der Quellen ist jeweils aktueller als bei der verwendeten DATAFACTORY-Version von 2014). Von den 23 PLZ in GeoNames konnte in der Online-PLZ-Suche der Post keine einzige gefunden werden – hier handelt es sich also um Postleitzahlen, die nicht (oder nicht mehr) gültig sind. Gleiches gilt für 10 aus den 76 openGeoDB-Fällen ausgewählten PLZ. Bei DBPedia sind 3 der 9 fünfstelligen Postleitzahlen aktuell gültig.

Es sind also Postleitzahlen enthalten, die weder 2014 gültig waren noch es aktuell sind. Die Anzahl dieser Fälle ist jedoch für alle betrachteten Quellen sehr gering und sie können daher vernachlässigt werden.

Nach Zusammenführung der Quellen nach dem oben beschriebenen Verfahren enthalten die Basistabellen 99,82% aller in der DATAFACTORY enthaltenen Zustellungs-Postleitzahlen. 88 der

aufgenommenen insgesamt 8.438 aufgenommenen Postleitzahlen sind nicht in der DATAFACTORY enthalten (23 aus GeoNames, 53 aus openGeoDB, 9 aus DBPedia und 3 aus OSM).

PLZ-AGS-Kombinationen

Auch hinsichtlich der PLZ-AGS-Kombinationen können die DATAFACTORY-Daten als Referenz für die Evaluation der Quellen herangezogen werden, da der AGS enthalten ist (‚PLZ_KGS‘). Relativ viele PLZ-AGS-Kombinationen aus der DATAFACTORY sind nicht in in den Quellen enthalten. Es handelt sich dabei überwiegend um Postfach-Postleitzahlen:

Art	**DATA-FACTORY (Referenz)**	**GeoNames**	**openGeoDB**	**DEDBPedia**	**OSM**	**GEO_UNIT nach Verfahren**
Zustellung	13.184	9.644 (73,15%)	11.311 (85,80%)	10.139 (76,90%)	8.735 (66,25%)	10.951 (83,06%)
Zustellung und Postfach	556	437 (78,60%)	522 (93,88%)	530 (95,32%)	475	480
Postfach	16.510	13	21	115	4	136
Einzel-Großempfänger	2.328	3	7	2	2	10
Gruppen-Großempfänger	795	1	2	1	1	3
Aktions-PLZ	2.032	1	1	-	-	1

Tabelle 36: Abdeckung PLZ-AGS-Kombinationen (Referenz DATAFACTORY)

Die Abdeckung der PLZ-AGS-Kombinationen liegt im Fall von GeoNames also für Zustellungs-Postleitzahlen bei ca. 73% (für Zustellung und Postfach bei 79%), für die anderen Arten liegen kaum Informationen vor. Bei openGeoDB ist die Abdeckung deutlich besser.

Bezüglich der PLZ-AGS-Kombinationen sind in GeoNames 386 Kombinationen enthalten, die nicht in der Referenz enthalten sind – davon sind aber 318 Gemeinden in GeoNames als ‚historisch‘ gekennzeichnet, es bleiben also nur noch 68 PLZ-AGS-Kombinationen, die Änderungen oder Fehler sein können.

In openGeoDB sind 3.076 PLZ-AGS enthalten, die nicht in der Referenz enthalten sind. Berücksichtigt man nur solche, die einen AGS haben, der in GeoNames enthalten ist und außerdem zu einer in GeoNames nicht als historisch gekennzeichneten Geo-Unit gehört, so bleiben 294.

In DEDBPedia sind (mit dem oben beschriebenen Vorgehen zur Extraktion von PLZ-AGS-Kombinationen) 1.057 PLZ-AGS-Kombinationen nicht in Referenz enthalten. Dabei sind jedoch viele vierstellige PLZ enthalten – diese sind problematisch, da nicht klar ist, ob sie eigentlich fünfstellige PLZ sind, die durch den fehlerhaft vergebenen Datentyp Integer für Postleitzahlen entstanden sind, oder aber, ob es sich um alte, tatsächlich vierstellige Postleitzahlen handelt. In beiden Fällen sind sie jedoch nicht in der Referenz enthalten. Werden nur fünfstellige PLZ einbezogen, so bleiben noch 34 PLZ-AGS-Kombinationen, die nicht in der Referenz enthalten sind.

In OSM sind 1.956 PLZ-AGS-Kombinationen nicht in der Referenz enthalten, davon gehören 244 zu AGS von Gemeinden, die in GEO_UNIT aufgenommen sind und 139 von diesen gehören zu Gemeinden, die nicht als historisch geflaggt sind.

Insgesamt sind in den Quellen also nur geringe Anteile möglicherweise problematischer/fehlerhafter PLZ-AGS-Kombinationen enthalten: Von insgesamt 12.038 in die Basistabellen aufgenommenen PLZ-AGS-Relationen sind 11.581 auch in der Referenz enthalten (es verbleiben also 457 als mögliche Fehler oder Änderungen – von diesen gehören 329 zu als historisch geflaggten Gemeinden).

Fazit

Insgesamt scheint GeoNames (jedenfalls für Deutschland) eine ausreichend verlässliche Quelle zum Füllen der Basistabellen zu sein. Diese Datenbasis kann aber sinnvoll ergänzt werden um Informationen aus den anderen Quellen (bspw. bietet openGeoDB in Bezug auf Postleitzahlen und PLZ-AGS-Kombinationen eine bessere Abdeckung). Die Quellen Wikipedia/DBPedia und OSM liefern dabei auch Informationen zu allen anderen Ländern, während openGeoDB derzeit noch nur für wenige Länder eingesetzt werden kann. Bei der Erweiterung des Verfahrens auf weitere oder auch alle Länder stellt das aber kein Problem dar – es könnten je nach Existenz regionale Quellen zusätzlich eingesetzt werden, durch die die von GeoNames gegebene Basis ggf. ergänzt werden könnte.

Zuordnung der Adressdatensätze aus WoS und Scopus zu Geoentitäten und -koordinaten

Bereits bei der groben Abschätzung zur Abdeckung hat sich gezeigt, dass für das Matching von Datenfeldern aus den Bibliometriedatenbanken mit Geodatenquellen Vorbereitungsschritte erforderlich sind – sowohl bei den Daten aus WoS und Scopus als auch bei den Daten aus den Geodatenquellen.

Vorbereitungsschritte

CITY und PLZ: Extraktion und Bereinigung von Postleitzahlen

Postleitzahlen liegen in unterschiedlichen Formaten vor:

- mit und ohne vorangestellten Länderkenner, der als ‚D' oder auch ‚DE' vorhanden sein kann (für ältere Adressen auch beispielsweise ‚O', ‚W' oder ‚DDR'),
- von der Postleitzahl getrennt durch ‚-', ‚:' oder Leerzeichen, zum Teil sind fehlerhaft Leerzeichen innerhalb der Postleitzahl enthalten (z.B. ‚76 128 Karlsruhe', ‚D-80,539 Munich', ‚DE-144 69 Potsdam'),
- es sind Scanfehler erkennbar, in denen beispielsweise ein ‚l' an Stelle einer 1 oder ein ‚G' an Stelle einer 6 erscheint (‚4532G Essen' (G statt 6), ‚D-l7489 Greifswald' (l statt 1)).

Postleitzahlen zu älteren Adressen sind im Matching zur Abschätzung der Abdeckung nicht erfasst, da sie nicht fünfstellig, sondern vierstellig sind.

Für Scopus-Adressen müssen die Postleitzahlen in den meisten Fällen zunächst aus dem CITY-String extrahiert werden. In der Abschätzung zur Abdeckung wurde eine vereinfachte Form dieser Extraktion angewendet, für das Verfahren selbst ist eine erweiterte Version notwendig, die die beschriebenen Besonderheiten mit erfassen und korrigieren kann.

Die CITY-Einträge wurden durch die Anwendung der Transformation aus der Institutionenkodierung[35] auch vorstandardisiert aufgenommen.

Erkennung von Länderfehlern

Die Extraktion der Postleitzahlen in Scopus eignet sich zudem zur Identifizierung von Fehlern in der Zuweisung des Countryattributs: nicht zu Deutschland gehörige Länderkenner in Postleitzahlen können Hinweise auf Adressen zu anderen Ländern mit deutschem Länderkenner in WoS oder Scopus geben. Beispiele sind ‚CH-8092 Zürich' und ‚DK-2100 Copenhagen Ø' als Werte zum City-Attribut in Adressen mit Countrycode=DEU.

Weitere Datenfelder

Im Abschnitt Statistik oben hat sich gezeigt, dass außer in den Feldern CITY und POSTALCODE bei Scopus auch in ORGANIZATION1 und ORGANIZATION2 Geoinformationen enthalten sein können, die auf Ebene der Gemeinde von Interesse sind. Aus diesem Grund wurden auch diese

35 Vgl. die Dokumentation zur Institutionenkodierung im Wiki des Kompetenzzentrums Bibliometrie (auf Anfrage erhältlich).

Felder vorbereitet, indem distinkte Werte im Original und in vorstandardisierter Form erfasst wurden (Nutzung der Transformation aus der Institutionenkodierung).

Zusammenfassung aller Namensvarianten

Alle relevanten Namensvarianten aus den verschiedenen Quellen wurden in einer Tabelle zusammengefasst. Dabei wurden – im Gegensatz zur Vorgehensweise beim Erstellen der Basistabellen – auch Namensvarianten von Teileinheiten der Gemeinden (Ortsteile) erfasst. Diese können zwar nicht direkt als identische Entität mit den Gemeinden verlinkt werden, da es sich nicht um die gleiche Entität handelt, jedoch können und sollen diese Bezeichnungen der entsprechenden Gemeinde im Verfahren zugeordnet werden, da es sich um Teile der jeweiligen Gemeinde handelt. Im einzelnen werden Namen aus den Quellen wie folgt extrahiert:

- **GEO_U_NAME:** alle Namen aus der Basistabelle.
- **GeoNames:** alle Namen aus der Tabelle ALTERNATE_NAMES. Hier sind diverse Namensvarianten erfasst (Abkürzungen, historische Namen, Namen in verschiedenen Sprachen usw.).
- **openGeoDB:** alle text_values zu den text_types 500100000 und 500100002 (Name und Sortiername) sowie die entsprechenden text_values von Entitäten, die über den text_type 400100000 (ist Teil von) als Teileinheiten identifiziert werden können.
- **DEDBPedia:** Labels von Seiten, die mit den Entitäten der Basistabellen verlinkt werden konnten, sowie die Labels von Seiten, für die Redirections zu diesen Seiten bestehen. Redirections enthalten wertvolle Hinweise auf Schreibvarianten von Namen, zum Beispiel sind
  - FRANKFURT/MAIN,
  - FRANKFURT A.M und
  - FRANKFURT A. M.

  Labels von Seiten, für die Redirections zu der Seite mit dem Label FRANKFURT (MAIN) existieren.
- **OSM:** In OSM müssen Namen über die tag_keys identifiziert werden. Hier gibt es wieder keine einfache Möglichkeit, alle tag_keys zu Namen/Bezeichnungen zu identifizieren. Es wurden daher alle tag_keys, die den String ‚name' enthalten und für Entitäten vorkommen, die in den Basistabellen enthalten sind, ausgewählt und daraus manuell einige ausgeschlossen.

  Die nachfolgende Tabelle 37 zeigt die häufigsten tag keys. Oft sind diese tag_keys von der Form name:<Sprache>, es sind aber auch Namen aus openGeoDB enthalten und auch andere wie z.B. ‚short_name' oder ‚alt_name'. Ausgeschlossen wurden beispielsweise Tag_keys, deren zugehörige Tag_Values ein Präfix, ein Suffix, Start- und Enddatumsangaben für die Gültigkeit von Namen oder eine URL enthalten.

	tag_key	Anzahl
1	name	8.652
2	openGeoDB:name	5.896
3	openGeoDB:sort_name	5.798
4	name:ru	1093
5	name:sr	312
6	name:nds	303
7	name:de	294
8	name:zh	213
9	name:fa	175
10	name:hsb	137

Tabelle 37: Top 10 der tag_keys like ‚%name%‘ für Entitäten aus den Basistabellen

Die is_in-Beziehung aus OSM kann nicht ohne weitere Bearbeitung genutzt werden, da hier die Beziehung als String und nicht als Verknüpfung von zwei OSM-IDs angegeben wird. Beispiel:

osmid=11420169
tag_key=‘is_in‘
tag_value=‘Leipheim,Günzburg,Schwaben,Bayern,Bundesrepublik Deutschland,Europe‘

Diese Beziehungen wurden hier daher vorerst außer Acht gelassen.

Die Anzahl der aus den jeweiligen Quellen extrahierten Namen, die nicht bereits in der Basistabelle GEO_U_NAME enthalten sind, zeigt die Tabelle 38. Namensvarianten können dabei natürlich in mehreren Quellen vorhanden sein.

Quelle	Anzahl extrahierter Namen, die nicht in GEO_U_NAME enthalten sind
GeoNames (Alternative Names)	5.627
GeoNames (Alternative Names über Hierarchy)	63
openGeoDB	5.019
OpenGeoDB (ist Teil von)	48.020
DEDBPedia (Labels)	2.328
DEDBPedia (Labels über Redirections)	9.860
OSM	4.102

Tabelle 38: Aus den Quellen extrahierte Namen, die nicht bereits in den Basistabellen enthalten sind

Außerdem wurde jede Namensvariante auch in vorstandardisierter Form aufgenommen. Für diese Vorprozessierung wurde die Transformationsprozedur aus der Institutionenkodierung (analog zum Vorgehen beim CITY-Feld, s.o. S.47) genutzt.

In der folgenden Tabelle werden Beispiele für Namensvarianten gegeben:

PK_GEO_UNIT	Namensvariante	Quelle
7745	Freiburg	OSM
	FREIBURG (BREISGAU)	DEDBPEDIA - redirections
	FREIBURG I. BR.	DEDBPEDIA - redirections
	Freiburg im Breisgau	OPENGEODB
	FRIBURG IM BRISGAU	DEDBPEDIA – redirections
	Günterstal	OPENGEODB - ist Teil von
5074	Aachen	OSM
	AIX-LA-CHAPELLE	DEDBPEDIA - redirections
	EILENDORF	OPENGEODB - ist Teil von
5180	Francfort	GEONAMES ALTERNATE_NAMES
	Francfort-sur-le-Main	GEONAMES ALTERNATE_NAMES
	Franckfort/Main	GEONAMES ALTERNATE_NAMES
	Frankfurt	OSM
	FRANKFURT A. M.	DEDBPEDIA – redirections
	FRANKFURT AM MAIN	DEDBPEDIA – labels
	FRANKFURT/MAIN	DEDBPEDIA – redirections
	FRANKFURT (MAIN)	DEDBPEDIA – redirections
	Frankfurte pie Mainas	GEONAMES ALTERNATE_NAMES
	Mainhattan	OSM
	Франкфурт на Майн	GEONAMES ALTERNATE_NAMES
	美因河畔法兰克福	GEONAMES ALTERNATE_NAMES
946	FFO	GEONAMES ALTERNATE_NAMES
	FRANKFURT	OPENGEODB
	FRANKFURT A. D. ODER	DEDBPEDIA – redirections
	Frankfurt (Oder)	OPENGEODB
11161	München	OSM
	Munich	GEONAMES ALTERNATE_NAMES

Tabelle 39: Beispiele für Namensvarianten aus den verschiedenen Quellen

Zusammenfassung aller PLZ

Die Postleitzahlen aus allen Quellen wurden bei Relation ‚hat Postleitzahl' in der Tabelle GEO_RELATION (die wie oben beschrieben gefüllt wurde) mit den Primary Keys (aus GEO_UNIT) der zugehörigen Gemeinden (identifiziert durch den AGS) in einer Tabelle abgelegt (diese Zuordnung muss nicht eindeutig sein!).

In den deutschen Adressen besteht dabei das Problem, dass in alten Adressen auch viele vierstellige/alte Postleitzahlen vorkommen. Um auch diese Adressen/Postleitzahlen sinnvoll zuordnen zu können (d.h. die Postleitzahleninformation tatsächlich nutzen zu können), wird eine Zuordnung auch der alte Postleitzahlen zu Gemeinden/AGS benötigt. Die Deutsche Post hat auf Anfrage eine Tabelle mit alten und zugehörigen neuen Postleitzahlen zur Verfügung gestellt. Damit können die alten Postleitzahlen über die zugehörigen neuen Postleitzahlen mit den entsprechenden Gemeinden bzw. PK_GEO_UNIT ebenfalls erfasst werden.

Diese beiden Tabellen (Zusammenfassung aller Namensinformationen und Zusammenfassung aller PLZ-Informationen) bilden die Basis für die Zuordnung.

Matching

Match-Typen

DieMit den wie oben beschrieben vorbereiteten Daten wird ein Matching der Postleitzahlen und Namensvarianten aus den Basistabellen und Datenquellen mit den Datenfeldern aus den Bibliometriedatenbanken durchgeführt.

Für PLZ- und CITY-Felder werden dabei drei Fälle unterschieden:

1. PLZ und CITY vorhanden
2. PLZ fehlt, CITY vorhanden
3. PLZ vorhanden, CITY fehlt

Diese Fälle beziehen sich auf die bereits vorbereiteten/bereinigten Werte. So kann beispielsweise eine PLZ-City-Kombination unter Fall 2 fallen, obwohl eine Postleitzahl gegeben ist, wenn sich diese nicht sinnvoll extrahieren ließ (beispielsweise eine nur dreistellige Ziffernfolge).

Dabei werden verschiedene Stringvergleichsverfahren angewendet (reguläre Ausdrücke, die bereits in die Vorbereitung eingegangene Transformation, Levenshteindistanz, Substrings) – im folgenden als Match-Typen bezeichnet. Tabelle 40 zeigt einen Auszug aus den 25 Match-Typen, die für die verschiedenen Fälle Anwendung finden:

Fall 1: PLZ und CITY vorhanden	Fall 2: PLZ fehlt, CITY vorhanden	Fall 3: PLZ vorhanden, CITY fehlt
1a: Match für PLZ und City		
1b: Match für PLZ und transformierte City		
1c: Match für PLZ LD=1[36] und City		
1d: Match für PLZ LD=1 und transformierte City		
1e: Match für PLZ und City LD=1		
1f: Match für PLZ und transformierte City LD=1		
1g: Match für PLZ und Substring City		
[...]	**2a:** Match für City	
	2b: Match für transformierte City	
	[...]	**3:** Match für PLZ
4a: Match für City ohne Berücksichtigung PLZ		
4b: Match für transformierte City ohne Berücksichtigung PLZ		
[...]		
5: Match für PLZ ohne Berücksichtigung City		

Tabelle 40: Match-Typen (Auszug)

Für die Datenfelder ORGANIZATION1 und ORGANIZATION2 sind dabei nur ein Teil der Match-Typen anwendbar, da davon auszugehen ist, dass hier keine Postleitzahleninformation enthalten ist. Da in diesen Feldern hauptsächlich Information enthalten ist, die keine Geoinformation ist, ist hier die Fehlerwahrscheinlichkeit sehr viel höher. Daher werden hier einige Match-Typen (beispielsweise Match-Typen, die die Levenshteindistanz verwenden) nicht angewendet.

Die Treffer werden jeweils mit zugehörigem Match-Typ erfasst. So entstehen für eine Adresse ggf. mehrere Zuordnungen zu verschiedenen Geo-Einheiten – mit identischen oder auch unterschiedlichen Match-Typen.

36 LD1=Levenshteindistanz von 1

Zusammenfassung der Match-Typen zu Gruppen (Punktesystem)

Die verschiedenen Match-Typen unterscheiden sich in der Zuverlässigkeit ihrer Treffer. Außerdem gibt es Treffer, die zwar im Prinzip korrekt sind, aber nachrangig erfasst werden sollten.

In Stichproben hat sich herausgestellt, dass Treffer über Ortsteile häufig zu unerwünschten Ergebnissen führen – beispielsweise ist Münster nicht nur eine Stadt in Westfalen, sondern auch Name eines Stadtteils (Weiler) von Gaildorf in Baden-Württemberg. Gaildorf tritt also als Treffer für den Match-Typ ‚Genauer Treffer für Ortsteil' für ‚Münster' auf. Der Treffer für die Stadt Münster soll hier größeren Wert haben als der Treffer für den Ortsteil Münster.

In einem weiteren Schritt werden daher die Match-Typen über ein Punktesystem in Gruppen zusammengefasst, wobei hier ein niedriger Punktwert eine hohe Zuverlässigkeit anzeigt. Auch hier können prinzipiell drei Fälle unterschieden werden: Treffer über City und PLZ beginnen mit 100, Treffer nur über die Postleitzahl mit 200 und Treffer über die City (ohne Postleitzahl) mit 300. Zu beachten ist dabei, dass sich diese Fälle von den bei den Match-Typen beschriebenen Fällen unterscheiden: während dort das Unterscheidungskriterium ist, ob die Felder gefüllt sind (z.B. PLZ not NULL) oder nicht, unterscheiden sich die Fälle hier darin, über welche Felder ein Match erzielt werden konnte. So fällt eine PLZ-City-Kombination, die für PLZ und City gültige Werte hat, bei den Match-Typen unter Fall 1. Sollte aber kein Match über PLZ und CITY möglich sein, sondern beispielsweise nur über City, so wird – je nach Match-Typ – eine Punktzahl über 300 vergeben. Da ein Match über Ortsteile besonders problematisch erscheint (jedoch in einigen Fällen erforderlich ist und daher nicht völlig ausgeschlossen werden kann), werden diese Fälle gesondert betrachtet – also nicht in die Gruppen mit Punktzahlen ab 100, 200 oder 300 aufgenommen, sondern mit Punktzahlen ab 1.000 besonders gekennzeichnet.

Prioritätsregeln

Anhand des Punktesystems können nun Prioritäten für die Zuordnungen festgelegt werden. Beispielsweise soll für eine PLZ-CITY-Kombination, für die ein genauer Treffer für sowohl PLZ als auch für City vorliegt, kein weiterer Treffer (über einen anderen Match-Typen) aufgenommen werden.

Bei der Erstellung der Prioritätsregeln sind u.a. folgende Punkte zu beachten:

- *Vorstandardisierung:* insbesondere im WoS (aber auch teilweise in Scopus) muss davon ausgegangen werden, dass eine Vorstandardisierung vorgenommen wurde. Beispielsweise tritt die Stadt Münster im WoS als MUNSTER auf.

  Wird die Priorität ‚Treffer City' vor ‚Treffer City transformiert' festgelegt und damit dann also für den Fall, dass ein genauer Treffer für City vorliegt, nicht mehr nach einem Treffer für die transformierten City-Werte gesucht, tritt folgender unerwünschter Effekt auf:

  MUNSTER → genauer Treffer für City: ‚Munster' (in der Lüneburger Heide[37])

  Da ein Treffer für City gefunden wurde, wird nicht weiter nach einem Treffer für die transformierte City gesucht (der wäre MUNSTER (transformiert: MUNSTER) → Münster

37 http://www.munster.de/index.php

(transformiert MUNSTER)) – obwohl hier die Vermutung nahe liegt, dass es sich um Münster, nicht um Munster handelt. Münster würden bei einer solchen Vorgehensweise keine Adressen aus dem WoS zugeordnet.

Hier handelt es sich um eine Besonderheit der Adressen im Kontext der bibliometrischen Datenbanken WoS und Scopus, die berücksichtigt werden muss.

- *Mehrfachzuordnungen*: Prinzipiell sollen Mehrfachzuordnungen zugelassen werden und diese entstehen nicht immer durch Zuordnungen identischer Match-Tpen. Ein streng hierarchisches System im Sinne von ‚Adressen werden über Match-Typ X zugeordnet, nur die verbleibenden Adressen werden weiter behandelt' ist daher nicht sinnvoll. Bestimmte Match-Typen sollen (einzeln oder als Gruppe von Match-Typen) weitere Zuordnungen über andere Match-Typen ausschließen, andere dagegen nicht.

Im Prinzip soll das Punktesystem als Orientierung dienen, also Match-Typen mit niedrigen Punktzahlen sollen Vorrang vor Match-Typen mit höheren Punktzahlen haben, wobei genaue Treffer (für sowohl City als auch transformierte City) weitere Zuordnungen ausschließen sollen – Treffer durch Match-Typen, die über Levenshteindistanz oder Substrings zuordnen, sollen dagegen weitere Treffer von identischem oder verschiedenem Typ zulassen.

Über die Anwendung der Prioritätsregeln werden die Matches aus dem vorangegangenen Schritt gefiltert, mit dem Ziel, ‚sichere' Zuordnungen vorrangig zu erfassen und korrekte, aber ggf. unerwünschte Zuordnungen bereits teilweise auszuschließen.

Eindeutigkeit

Auch nach der Anwendung der Prioritätsregeln verbleiben Mehrfachzuordnungen. Diese können aus verschiedenen Gründen entstehen, beispielsweise:

1. Zuordnungen einer PLZ-City-Kombination können zum einen dadurch entstehen, dass im jeweiligen Feld tatsächlich *mehrere Geo-Einheiten* genannt sind, beispielsweise CITY='Aachen und Bonn'. Hier ist die Mehrfachzuordnung gewünscht und richtig.
2. In anderen Fällen erfolgt beispielsweise durch die Postleitzahl eine andere Zuordnung als durch den Ortsnamen. Hier erfolgt die Mehrfachzuordnung aufgrund von *Datenfehlern* und sie kann entweder zugunsten einer Information aufgelöst werden oder bestehen bleiben.
3. Außerdem gibt es den Fall der *Unvollständigkeit/Mehrdeutigkeit* innerhalb eines Werts, beispielsweise wird ‚Frankfurt' sowohl Frankfurt am Main als auch Frankfurt an der Oder zugeordnet.
4. Es gibt *mehrere Zuordnungen, die unterschiedlich viel Information nutzen* (beispielsweise über Match-Typen mit Substrings) – nicht alle nutzen die volle enthaltene Information. Beispiel: CITY=‚Frankfurt Oder'.

   In den Namensvarianten tritt ‚FRANKFURT' sowohl als Namensvariante für Frankfurt am Main als auch für Frankfurt an der Oder auf und würde über Match-Typen, die Substrings

verwenden, damit beiden zugeordnet, obwohl die Angabe in CITY – durch den Zusatz ‚Oder' – eigentlich ausreichend Information enthält.

Im Fall 1 sind die Mehrfachzuordnungen gewünscht und sollen bestehen bleiben. In Fall 2 und 3 sollten die Mehrfachzuordnungen ebenfalls bestehen bleiben (und geflaggt werden), um die Unvollständigkeit/den Widerspruch in der Zuordnung zu kennzeichnen.

Für die unter Fall 4 fallenden Zuordnungen müssen in diesem Schritt Regeln angewendet werden, die sicher stellen, dass die enthaltene Information vollständig eingebracht wird, um – sofern möglich – mehr eindeutige Zuordnungen zu erhalten, ohne dabei jedoch erwünschte Mehrfachzuordnungen auszuschließen.

Niedrigste Aggregationsebene

Nach diesem Schritt entsteht eine Liste von Zuordnungen von Adressen zu Geo-Einheiten. Dabei sind ggf. Zuordnungen einer Adresse zu Geo-Einheiten mehrerer Aggregationsebenen enthalten. Hier wird nur die Zuordnung auf der jeweils niedrigsten Aggregationsebene beibehalten.

Beispiel: ‚BIELEFELD, NRW'

Hier ist eine Zuordnung zu Bielefeld und eine Zuordnung zum Bundesland NRW gegeben. Da in der Basistabelle die Beziehung ‚NRW ist Bundesland von BIELEFELD' erfasst ist, wird die Zuordnung zum Bundesland nicht benötigt und daher gelöscht – nur die Zuordnung auf Stadt-/Gemeindeebene bleibt bestehen.

Die für die jeweiligen Anwendungen gewünschten Aggregationen auf verschiedenen Ebenen können über die Basistabellen von Anwendern selbst erstellt werden.

Die folgende Abbildung 6 zeigt die wesentlichen Schritte des Zuordnungsverfahrens in einer Übersicht.

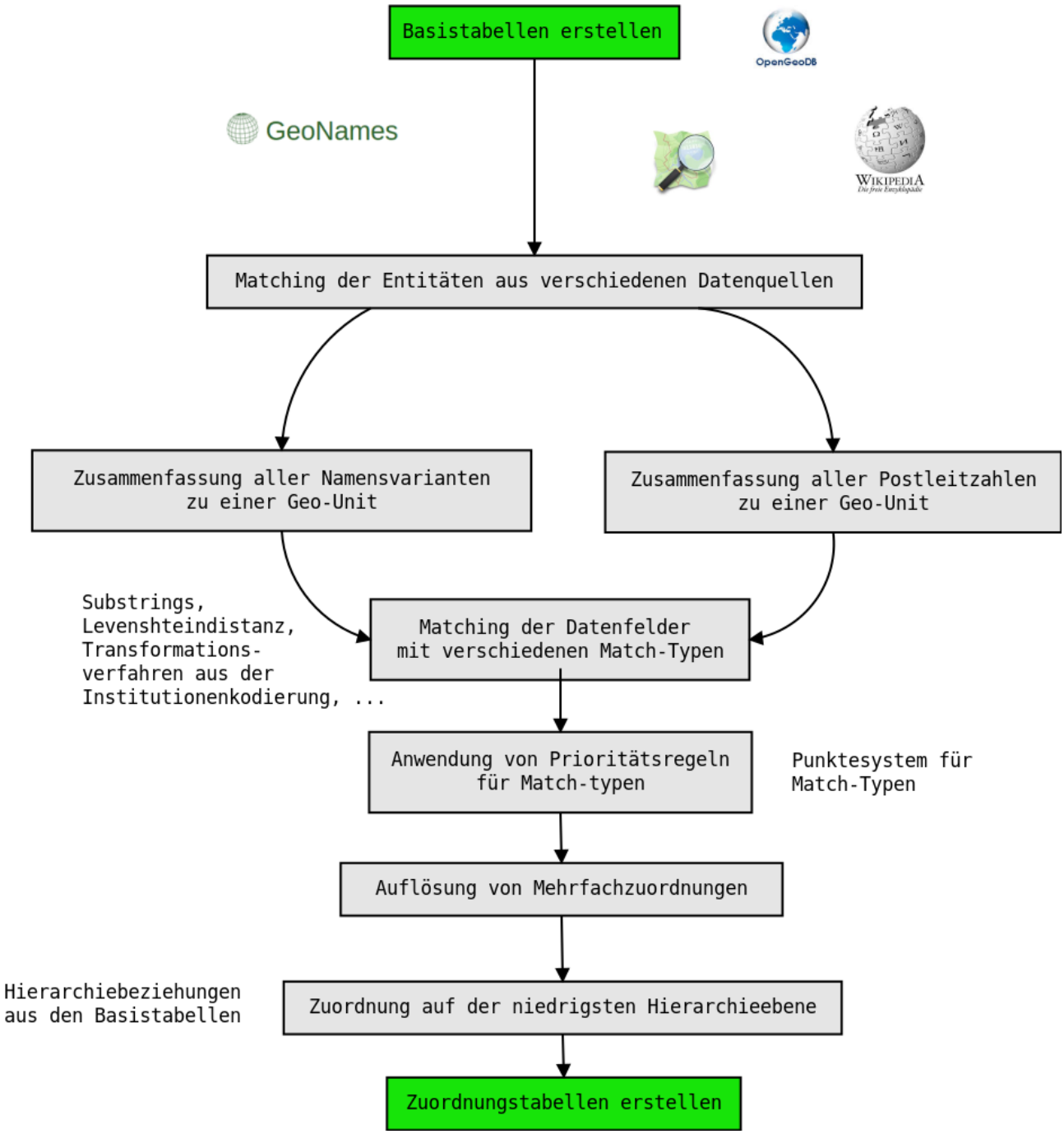


Abbildung 6: Verfahren (Übersicht)

Extraktion von weiteren Daten zu Geo-Einheiten aus Wikipedia

Über DBPedia können zu allen erfassten Geo-Units weitere Daten aus Wikipedia bezogen werden, sofern diese in DBPedia erfasst sind (insbesondere Auszüge der in der sog. Infobox enthaltenen Daten).

Abbildung 7: Infobox Wikipedia

Es müssen dabei jeweils die Properties identifiziert und zusammengefasst werden, die die gewünschten Daten enthalten.

Die folgende Tabelle 41 zeigt Properties für Entitäten, die mit Geo-Einheiten der Tabelle GEO_UNIT verlinkt sind mit ihren Auftrittshäufigkeiten. Hier wird erneut deutlich, dass in Wikipedia/DBPedia viele Daten vorhanden, diese jedoch nicht immer einfach zu extrahieren oder interpretieren sind. Beispielsweise ist nicht unmittelbar klar, welche Werte zur Property *<http://de.dbpedia.org/property/ergebnisalt>* zu erwarten sind. Andere Properties sind aussagekräftiger benannt und so könnten beispielsweise die GND, das Kfz-Kennzeichen, die Vorwahl, der Bürgermeister usw. extrahiert mit Geo-Units verlinkt werden. Außerdem bietet auch DBPedia Geo-Koordinaten und Zugehörigkeiten, z.B. zum Landkreis (wobei Geokoordinaten und Hierarchien bereits aus GeoNames bezogen und daher hier nicht mehr benötigt werden).

Zu beachten ist dabei, dass bei Daten, die sich über die Zeit ändern, nicht unmittelbar klar ist, welchen Stand diese zeigen – die der DBPedia zugrunde liegende Triplestruktur sieht keine Berücksichtigung von Zeitstempeln vor.

Hier wäre zu überlegen, welche Informationen tatsächlich Relevanz für bibliometrische Auswertungen haben können und aufgenommen werden sollen.

Property	Häufigkeit
<http://de.dbpedia.org/property/partei>	17.296
<http://de.dbpedia.org/property/plz>	11.830
<http://de.dbpedia.org/property/vorwahl>	11.448
<http://de.dbpedia.org/property/ergebnis>	11.370
<http://de.dbpedia.org/property/typ>	11.230
<http://de.dbpedia.org/property/fläche>	11.102
<http://de.dbpedia.org/property/längengrad>	11.099
<http://de.dbpedia.org/property/bundesland>	11.099
<http://de.dbpedia.org/property/breitengrad>	11.099
<http://de.dbpedia.org/property/lageplan>	11.094
<http://de.dbpedia.org/property/gemeindeschlüssel>	11.087
<http://de.dbpedia.org/property/höhe >	11.066
<http://de.dbpedia.org/property/bürgermeister>	11.056
<http://de.dbpedia.org/property/wappen>	10.799
<http://de.dbpedia.org/property/website>	10.435
<http://de.dbpedia.org/property/gnd>	10.100
<http://de.dbpedia.org/property/adresseVerband>	10.068
<http://de.dbpedia.org/property/landkreis>	9.473
<http://de.dbpedia.org/property/ergebnisalt>	8.604
<http://de.dbpedia.org/property/kfz>	8.268

Tabelle 41: Top20 Properties für mit Geo-Units verknüpfte DBPedia-Entitäten

6. Tabellenstruktur Datenlieferungen

Die Ergebnisse der Geo-Kodierung werden in einem System von relationalen Tabellen abgelegt. Das Schema dieser Tabellen ist so konzipiert, dass eine unmittelbare Anschlussfähigkeit an die Datenbanken des Kompetenzzentrums Bibliometrie gewährleistet ist. Es folgt in der Grundstruktur dem Schema, das für die Ergebnisse der Institutionenkodierung verwendet wird. Damit könnten die durch das Verfahren erzielbaren Kodierungen in einer Form ausgeliefert bzw. in die Datenbanken des Kompetenzzentrums Bibliometrie eingespeist werden, die den Anwendern eine weitgehend analoge Nutzung von Institutionen- und Geokodierung ermöglicht.

Die Tabellen erlauben nicht nur eine Zuordnung der Adressen zu einer oder mehreren Geoentitäten (und für diese Geokoordinaten zu liefern), sondern machen auch Aggregationen möglich (Stadt – Bundesland – Land).

Die oben beschriebenen Basistabellen

- GEO_UNIT (und GEO_UNIT_TYPE) mit den Geokoordinaten und dem Type der Geo-Unit,
- GEO_U_NAME mit einem Namen je Geo-Unit,
- GEO_U_IDENTIFIER mit GEO_U_IDENTIFIER_TYPE (die u.a. den amtlichen Gemeindeschlüssel enthalten) sowie
- GEO_RELATION mit GEO_RELATION_TYPE (mit Relationen wie beispielsweise die zwischen Gemeinden, Kreisen und Bundesländern)

könnten gemeinsam mit einer Zuordnungstabelle ausgeliefert werden, die Zuordnungen von Adressen zu den Geo-Entitäten in GEO_UNIT enthält. Wie auch bei der Institutionenkodierung sind hierbei Mehrfachzuordnungen möglich. Eine Zuordnungstabelle könnte – analog zu den entsprechenden Tabellen der Institutionenkodierung – folgende Spalten zur Verknüpfung der Adressen in der Bibliometriedatenbank mit den hier erstellten Basistabellen enthalten:

KB_WOS/SCP_ADDR_GEO	
FK_INSTITUTIONS	Foreign key zu INSTITUTIONS.PK_INSTITUTION
ADDRESS_FULL	INSTITUTIONS.ADDRESS_FULL
FK_GEO_UNIT	Foreign key zu GEO_UNIT.PK_GEO_UNIT

Eine separate Aggregationstabelle wie z.B. KB_SECTORS ist nicht erforderlich, da hier – ohne Berücksichtigung von Strukturveränderungen über die Zeit – einfach über die Tabelle GEO_RELATION wie gewünscht aggregiert werden kann. Dabei kann die Aggregationsebene jeweils vom Anwender selbst gewählt werden.

7. Fazit und Ausblick

Mit dem Projekt wurde ein Verfahren zur Geokodierung von Autorenadressen entwickelt und getestet, das Anwendungen ohne zusätzliche Lizenzkosten ermöglicht (auf der Basis von open access Quellen). Die Methode weist im Grundsatz eine Parallelität zu dem bereits etablierten Verfahren der Institutionenkodierung auf (Zuordnung von Adressen zu Geo- wie Org-Units).

Über eine statistische Auswertung zu Art und Vollständigkeit der in den Adressdatensätzen von Web of Science (WoS) und Scopus enthaltenen (und extrahierbaren) Geoinformationen konnte gezeigt werden, welche Datenfelder in diesen Datenbanken für eine Geokodierung sinnvoll zu nutzen sind.

Es wurden geeignete Quellen für den Bezug von Informationen zu geografischen Daten wie Ortsnamen, Geokoordinaten, Postleitzahlen und Hierarchiebeziehungen zwischen geografischen Einheiten recherchiert und in Bezug auf ihren Nutzen im spezifischen Kontext der Autorenadressen evaluiert. Darüber hinaus wurden Möglichkeiten für den Bezug weiterer Daten zu geografischen Einheiten aus diesen freien Quellen getestet.

Mit den so gewonnenen Erkenntnissen und Daten wurden Grundzüge einer Methode für die Geokodierung entwickelt sowie eine geeignete Tabellenstruktur entworfen, mit der die durch dieses Verfahren erzielbaren Kodierungen zukünftig ausgeliefert bzw. in die Datenbanken des Kompetenzzentrums Bibliometrie eingespeist werden können.

Datengrundlage

In Bezug auf die Gewinnung der notwendigen Basisdaten hat sich gezeigt, dass Daten aus GeoNames hierfür eine gute Grundlage bieten, die durch weitere (länderspezifische) Daten (z.B. die des Statistischen Bundesamtes) noch angereichert bzw. verbessert werden können. Aber auch ohne weitere Verbesserung kann das Verfahren bereits ohne Einschränkung angewendet werden (im Prinzip auch für andere Länder).

Bezüglich der PLZen in Deutschland ist ein Schwachpunkt, dass in den hier genutzten freien Quellen nahezu keine Postfach- und Großempfänger-PLZen existieren, die aber in den bibliometrischen Datenbanken durchaus auftauchen. Hier bleibt daher für die Zuordnung wesentliche Information noch ungenutzt, solange keine geeignete Quelle erschlossen werden kann.

Stand und Weiterentwicklung des Verfahrens für die deutsche Geokodierung

Das vorgeschlagene Verfahren wurde für Adressen aus dem WoS und Teilschritte auch für Adressen aus Scopus getestet. Die Informationen zu Gemeinden und Postleitzahlen wurden durch zwei weitere zuverlässige Quellen (vom Statistischen Bundesamt und der Deutschen Post) auf Vollständigkeit und Korrektheit überprüft – das Füllen der Basistabellen ist weitgehend vollständig und korrekt erfolgt. Wichtig für das Verfahren ist eine Verlinkung von möglichst vielen Entitäten der verschiedenen Quellen mit den Geo-Units. Auch das konnte mit den angewendeten Verfahren erreicht werden (wobei der Grad der Vollständigkeit der Verlinkungen für die Quellen unterschiedlich hoch ist). Damit konnte eine Zusammenfassung zahlreicher Namensvarianten zu den gegebenen Geo-Units erfolgen und eine Anwendung der verschiedenen Match-Typen zeigt,

dass sich mit diesen Daten zu Namensvarianten insbesondere in Kombination mit dem Transformationsverfahren aus der Institutionenkodierung viele der in den bibliometrischen Datenbanken verfügbaren Informationen auf Geo-Units abbilden lassen (bei 45.942 distinkten PLZ-City-Kombinationen aus dem WoS kann bisher für 43.538 Kombinationen mindestens eine Zuordnung zu einer Geo-Unit erfolgen). Unter den verbleibenden Kombinationen befinden sich nach grober Sichtung viele Einträge, die keine Geoinformation auf Gemeindeebene enthalten (z.B. Straßennamen) sowie Geoinformationen zu ausländischen Geo-Units.

Hier besteht Potential zur Aufdeckung von Fehlern in der Länderzuordnung – gerade auch dann, wenn die Geokodierung auf mehrere Länder angewendet werden kann.

Für die Zuordnung wurden Prioritätsregeln entwickelt und angewendet – damit kann eine Reduktion der Mehrfachzuordnungen erfolgen und es werden unerwünschte Zuordnungen ausgeschlossen. An dieser Stelle bietet eine noch weitergehende Analyse der bisherigen Ergebnisse weiteres Verbesserungspotential – wie auch eine Erweiterung des Verfahrens zur Auflösung von Mehrfachzuordnungen.

Die Bestimmung einer ‚Precision' im üblichen Sinn ist hier schwierig – in einigen Fällen ist tatsächlich unklar, welche Zuordnung ‚die richtige(n)' ist/sind (Beispiel: ‚Frankfurt'). Bisher wurden die geografischen Zuordnungen zu City-PLZ- Kombinationen und die geografischen Zuordnungen zu anderen Feldern (Organization1 und Organization2) einzeln betrachtet – in einer Weiterentwicklung könnten diese Informationen (evtl. in Verbindung mit weiteren Informationen aus den Feldern Organization1 und Organization2) besser verknüpft werden, um ggf. bestehende Mehrdeutigkeiten aufzulösen.

Eine weitere Option wäre, das Zuordnungsverfahren nicht nur zur Zuordnung, sondern auch zum Flaggen von Fehlertypen zu nutzen (Fehler in der Länderzuordnung von Adressen, Mehrfachzuordnung aufgrund von Datenfehlern und -widersprüchen oder auch unvollständigen Adressen). Ähnlich zum Verfahren des FIZ zur Dublettenerkennung könnte ein Ergebnis einer Geokodierung auch Gewichtungen von Zuordnungen enthalten.

Möglichkeiten und Probleme der Erweiterung des Verfahrens auf weitere Länder

Für die Erweiterung des Verfahrens muss beachtet werden, dass länderspezifische Unterschiede und Besonderheiten hinsichtlich der Struktur und der historischen Gegebenheiten der Länder bestehen (auch: Struktur von Postleitzahlen und Verfügbarkeit von passenden Aggregationsebenen und zugehörigen Identifiern). Für Deutschland sind beispielsweise alte vierstellige Postleitzahlen zu berücksichtigen. Insgesamt muss ein allgemeineres Verfahren mit den länderspezifischen Formaten von Postleitzahlen und auch verschiedenen Identifiern für Gemeinden umgehen können. Es muss zusätzliche Informationen aus nur regional verfügbaren Quellen zulassen, sie dürfen aber nicht notwendige Voraussetzung sein. Ob die Daten für andere Länder in den Quellen in ähnlicher Qualität und Vollständigkeit verfügbar sind, bedarf noch der Überprüfung.

In Bezug auf Wikipedia ist jeweils die landesspezifische Wikipedia/DEPedia (zum Beispiel die DEDBPedia für Deutschland) eine bessere Quelle als die englische Wikipedia. Wenn für alle Länder nur die letztere verwendet wird, geht Information verloren. Werden jedoch landesspezifische

Versionen benutzt, liegen auch Properties in der jeweiligen Sprache vor und die Datenmenge und der dafür erforderliche Speicherplatz steigen stark an. Eventuell kommen hier Untersuchungen zur ausschließlichen Nutzung der englischen Wikipedia in Betracht.

Mögliche Struktur von Datenlieferungen

Die durch das Verfahren erzielbaren Geo-Kodierungen könnten zukünftig mit der vorgestellten Tabellenstruktur ausgeliefert bzw. in die Datenbanken des Kompetenzzentrums Bibliometrie eingespeist werden. Die Nutzung könnte dann weitgehend analog zu den Ergebnissen der Institutionenkodierung erfolgen.

Literatur

Abramo, G., D'Angelo, C. A., & Di Costa, F. (2015). A new approach to measure the scientific strengths of territories. *Journal of the Association for Information Science and Technology*, *66*(6), 1167–1177. DOI: 10.1002/asi.23257

Bornmann, L., & Leydesdorff, L. (2011). Which cities produce more excellent papers than can be expected? A new mapping approach, using Google Maps, based on statistical significance testing. *Journal of the American Society for Information Science and Technology*, *62*(10), 1954–1962. DOI: 10.1002/asi.21611

Frenken, K., Hardeman, S., & Hoekman, J. (2009). Spatial scientometrics: Towards a cumulative research program. *Journal of Informetrics*, *3*(3), 222–232. DOI: 10.1016/j.joi.2009.03.005

Gao, S. (2015). Towards a frontier of spatial scientometric studies. *ACM SIGWEB Newsletter*, (Spring), 1–9. DOI: 10.1145/2749279.2749284

Rimmert, C. (2012). How Geocoding Tools Can Help Cleaning Data. In É. Archambault, Y. Gingras, & V. Larivière (Hrsg.), *Proceedings of 17th International Conference on Science and Technology Indicators* (Bd. 2, S. 881–883). Montréal: Science-Metrix and OST.

Rimmert, C. (2013). Evaluation of Yahoo! Placefinder on institutional addresses. In S. Hinze & A. Lottmann (Hrsg.), *Translational twists and turns: Science as a socio-economic endeavor. Proceedings of STI 2013 Berlin* (S. 566–570). Berlin: iFQ/ENID.

Xuemei, W., Mingguo, M., Xin, L., & Zhiqiang, Z. (2014). Applications and researches of geographic information system technologies in bibliometrics. *Earth Science Informatics*, *7*(3), 147–152. DOI: 10.1007/s12145-013-0132-4

Winterhager, M., Schwechheimer, H., & Rimmert, C. (2014). Institutionenkodierung als Grundlage für bibliometrische Indikatoren. *Bibliometrie - Praxis und Forschung*, *3*(14), 1-22. https://pub.uni-bielefeld.de/publication/2703536