erschienen in: Tagungsband FOSSGIS 2015, Münster, 2015, 18-24

Offene Geodaten: Lage von Orten im Vergleich

Thomas Brinkhoff

Einführung

Neben amtlichen und kommerziellen Geodaten stellen seit einigen Jahren auch insbesondere durch *Crowdsourcing* erhobene nutzergenerierte Inhalte (user-generated content, UGC) eine wichtige Grundlage für raumbezogene Anwendungen dar. So ist OpenStreetMap (OSM) sicherlich derzeit das bekannteste Projekt für gemeinschaftliche Kartierungen (collaborative mapping). Sowohl durch solche Aktivitäten als auch durch die zunehmende Bereitstellung von amtlichen Daten unter Nutzung von freien Lizenzformen nimmt der Umfang von *offenen Geodaten* (open geospatial data) stetig zu [1]. Mit dieser Zunahme steigt aber auch die Heterogenität der Datenquellen und damit auch der Datenqualität. So kann amtlichen Daten (meist) eine hohe Lagequalität unterstellt werden, während diese bei anderen Herkunftsquellen stark variieren kann. Andererseits bietet die zunehmende Zahl von Datenquellen auch die Chance, die Daten untereinander zu vergleichen und Daten(quellen) hoher Qualität zu bestimmen und auszuzeichnen.

Eine andere wichtige Entwicklung, die man im Geoinformationsbereich beobachten kann, ist die zunehmende Verfügbarkeit einerseits funktional weit entwickelter und andererseits auch einfach zu nutzender Software unter freien Lizenzmodellen. Dies umfasst (u.a.) Software-Pakete und -Bibliotheken für Geodatenbanksysteme, Geodienste, Web Mapping, Geoinformationssysteme (GIS) und zur Programmierung raumbezogener Anwendungen [2]. Diese Entwicklung schlägt sich auch in den Lehrinhalten von Hochschulstudiengängen nieder. So wird im Bachelor-Studiengang „Geoinformatik“ an der Jade Hochschule in Oldenburg in den Modulen „Geoinformationssysteme I und II“ QGIS als eines von zwei Hauptsystemen gelehrt. Die Einführung in Geodatenbanken erfolgt mit PostgreSQL/PostGIS und die in Geodienste mit GeoServer. Die nachfolgenden Ausführungen beruhen u.a. auf einer Abschlussaufgabe für das Modul „GIS-Programmierung“. In diesem Modul wird in Entwicklung raumbezogener Anwendungen mittels Software-Bibliotheken hauptsächlich unter Nutzung von GeoTools und JTS (Java Topology Suite) eingeführt.

Aus den bisherigen Ausführungen leitet sich die Aufgabe ab, die nachfolgend betrachtet werden soll: eine Qualitätsuntersuchung von offenen Geodaten. Die zu betrachtenden Daten sind auf einen wichtigen Spezialfall beschränkt: die Lage von Orten (wie Städte, Gemeinden und Siedlungen) inklusive administrativer Gebiete (Bezirke, Gemeinden usw.). Solche Daten sind eine wichtige Grundlage für die Erstellung von Karten und für diverse Geoanwendungen. Die Untersuchung wird durch ein mit GeoTools/JTS entwickeltes Werkzeug unterstützt.

Offene Daten zur Lage von Orten

Die Lage von Orten kann entweder über eine Punktkoordinate oder über Flächen (i.d.R. als Multipolygon) vorliegen. Wichtige Quellen für Lagedaten sind u.a. die GeoNames Geographical Database, OpenStreetMap und Wikipedia; sie weisen folgende Vorteile auf:

- weltweite Abdeckung
- zentrales Portal (OSM z.B. über http://download.geofabrik.de/)
- einheitliches Datenmodell
- (relativ) einfache Zugriffsmöglichkeiten auf Teile der Daten

GeoNames unterliegt der „Creative Commons Attribution 3.0 License“. Auf die Daten kann über Download und Dienste zugegriffen werden. Wichtige Attribute sind:

- geonameid: globale ID
- name: Name des Ortes
- alternatenames: Liste mit alternativen Namen und Schreibweisen
- latitude: geografische Breite in WGS‘84
- longitude: geografische Länge in WGS’84
- feature class: Grundtyp; „P“ für Orte (places), „A“ für administrative Einheiten
- feature code: Untertyp; z.B. „PPL“ für bewohnte Orte (populated place)
- country code: Staat mit ISO-3166 2-Code
- admin1 code, …, admin4 code: Administrative Ebene 1/2/3/4
- population: Einwohnerzahl (meist nicht belegt; ohne Referenzdatum)
- modification date: Datum der letzten Änderung

Als Defizite können genannt werden, dass keine Qualitätsindikatoren für die Lageinformation und keine nationalen Schlüssel in den Daten enthalten sind.

OSM-Daten werden (seit 2012) unter der „Open Database License“ (ODbL) veröffentlicht. Die Daten stehen (oftmals staatenweise) zum Download zur Verfügung. Administrative Einheiten liegen i.d.R. als flächenhafte Objekte vor, die ein „tag“-Element besitzen, dessen Schlüsselattribut „k“ den Wert „admin_level“ aufweist (nachfolgend werden solche Elemente verkürzt als „xxx“-Tag bezeichnet). Das Wertattribut „v“ des „admin_level“-Tags gibt die administrative Ebene an. Orte können als Punkt- oder Flächengeometrien vorliegen; sie sind über „place“-Tags kategorisiert. Oftmals sind Geoobjekte sowohl Ort als auch administratives Gebiet. Über „name“-Tags sind Namen und ggf. Sprachvarianten der Namen angegeben. Teilweise liegt (i.d.R. ohne Referenzdatum) die Einwohnerzahl über das „population“-Tag vor. Gliederungsinformationen können über „is_in“-Tags definiert sein. In Abhängigkeit vom Staat geben Tags (in Deutschland z.B. mit Schlüssel „de:amtlicher_gemeindeschluessel“) ggf. Auskunft über nationale Schlüssel.

Alle Elemente enthalten die Attribute „version“, „changeset“, „uid“, „user“ und „timestamp“ mit Informationen über die Herkunft und Aktualität. Importe sind ggf. über „source“-Tags gekennzeichnet. Quantitative Qualitätsindikatoren bezüglich Lage und Grenzverlauf existieren nicht. Auch variiert die tatsächliche Belegung der oben genannten Attribute. Durch geeignete Programme müssen die gewünschten Teildaten aus den vollständigen OSM-Dateien extrahiert werden, was aufgrund des recht komplexen Modells für Multipolygone, teilweise fehlender Geometrieteile und großen Dateien eine nicht immer triviale Aufgabe darstellt.

Wikipedia stellt zu vielen Artikeln eine (oder selten auch mehrere) WGS’84-Koordinaten zur Verfügung. Eine Möglichkeit des Zugriffs ist der kategorieweise KML-Export über den Tool Server:

http://toolserver.org/~para/cgi-bin/kmlexport?project=<*language*>&article=<*category*>

Optional kann zudem über den Parameter *l* die maximale Zahl von Rekursionen bezüglich der Kategorien angegeben werden. Das Resultat ist eine KML-Datei mit „Placemark“-Elementen, in denen eingebettet das „name“-Element den Artikelnamen, das „Point“-Element die Koordinaten und das „description“-Element einen Link auf den Wikipedia-Artikel enthält. Eine andere Möglichkeit stellt die Auswertung von Wikipedia-Dumps dar, was allerdings aufwändiger ist und – außer einer ID – i.d.R. keinen inhaltlichen Mehrwert liefert.

Eine administrative Zuordnung der Ortsangaben kann – wenn eine entsprechende Kategorisierung der Artikel vorliegt – durch den Aufruf des KML-Exports erzielt werden, z.B.

http://toolserver.org/~para/cgi-bin/kmlexport?project=de&article=Kategorie:Gemeinde in Niedersachsen

Amtliche offene Geodaten haben im Gegensatz zu den vorgenannten Datensätzen weder eine weltweite Abdeckung noch ein zentrales Zugangsportal. Sie liegen in keinem einheitlichen Datenmodell vor. Die Einfachheit des Zugriffs auf die Daten variiert; oftmals nutzen die Daten eine nationale Projektion. Positive Aspekte sind oftmals hohe Lagegenauigkeiten, die Nutzung nationaler Schlüssel und die vollständige Angabe von administrativen Zugehörigkeiten.

Vollständigkeit und Eindeutigkeiten in GeoNames und OSM

Für die folgenden Untersuchungen werden vier Staaten betrachtet: Argentinien, China, Deutschland und Neuseeland.

Tabelle 1 zeigt für diese vier Staaten wesentliche Eigenschaften für GeoNames (Stand 5.2.2015). Die Belegung der administrativen Zugehörigkeiten ist teilweise sehr gering. Damit lassen sich insbesondere in China (33%), aber auch in Deutschland (12,6%) viele Orte nicht eindeutig über Angabe des Namens und der administrativen Zugehörigkeit geokodieren (d.h. mehrere Orte mit unterschiedlicher Lage werden bestimmt). Die Spalte „Letzte Änderung“ gibt den durchschnittlichen Zeitraum seit der letzten Modifikation in Jahren an.

Land	Features gesamt	Feature Class "P" und "A"	Adm1 belegt	Adm2 belegt	Adm3 belegt	nicht eindeutig	Letzte Änderung
Argentinien	47.150	7.458	99,96%	16,73%	0,08%	0,46%	5,70
China	447.884	414.124	99,93%	0,12%	0,70%	32,95%	3,21
Deutschland	182.058	91.450	99,99%	39,78%	53,96%	12,60%	3,21
Neuseeland	67.444	3.395	89,66%	77,64%	0,06%	2,24%	4,29

Tabelle 1: Eigenschaften von Orten und administrativen Einheiten in GeoNames

Tabelle 2 zeigt ähnliche Eigenschaften für OpenStreetMap-Daten (Stand 5.2.2015), wobei Deutschland sich auf Niedersachsen beschränkt. Ausgewählt wurden „node“-Elemente, die ein „place“-Tag beinhalteten. Die Spalte „Adm1 belegt“ bezieht sich auf die jeweils relevanten „is_in“-Tags. Auch hier ist nicht immer eine eindeutige Geokodierung möglich (China: 29%). Nationale Schlüssel sind kaum vorhanden.

Land	Anzahl Places	Adm1 belegt	nicht eindeutig	nationaler Schlüssel	letzte Änderung
Argentinien	11.344	57,11%	7,47%	0,00%	1,50
China	59.059	2,48%	29,08%	0,00%	1,67
D: Niedersachsen	9.238	59,68%	2,18%	0,06%	3,05
Neuseeland	2.572	24,46%	3,54%	0,00%	2,97

Tabelle 2: Eigenschaften von punktförmigen Orten (places) in OSM

Tabelle 3 zeigt die Eigenschaften von administrativen Einheiten in den OSM-Daten. Die Verfügbarkeit von Daten schwankt international sehr stark. Dies gilt auch für das Vorhandensein

von nationalen Schlüsseln. Die Spalte „Features mit Geometrie“ gibt die Anzahl der Geoobjekte an, bei denen sich die Geometrie tatsächlich konstruieren lässt.

Land	Typ	tatsächl. Anzahl	Features	Features mit Geometrie	nationaler Schlüssel
Argentinien	Bezirke	513	290	290	0
China	Präfekturen	341	330	321	0
China	Bezirke	2.853	2.427	2.395	0
China	Gemeinden	43.575	615	608	0
D: Niedersachsen	Gemeinden	988	988	988	988
Neuseeland	Bezirke	67	61	59	0

Tabelle 3: Eigenschaften von administrativen Gebieten in OSM

Werkzeug zur Untersuchung von Wikipedia-Koordinaten

Wie zuvor erwähnt, enthalten die Wikipedia-Daten die administrative Zugehörigkeit nur indirekt. Nationale Schlüssel lassen sich über den KML-Export nicht gewinnen. Damit sind für einen Lagevergleich mit amtlichen Geodaten folgende Funktionalitäten zweckmäßig:

- Laden der amtlichen Geodaten über Shapefiles.
- Einlesen der Wikipedia-Daten über den o.g. KML-Export:
  - Sprache und Kategorie können frei eingeben werden.
- Verwaltungsgebiete geometrisch bestimmen:
  - Es werden zwei Layer ausgewählt: Layer 1 mit Punktkoordinaten und Layer 2 mit Flächengeometrien. Bei Layer 2 wird zudem ein Namensattribut interaktiv bestimmt. Layer 1 wird um ein String-Attribut „admin“ ergänzt. Der Wert dieses Attributs wird aus dem Namensattribut des Features bestimmt, dessen Fläche den Punkt enthält bzw. schneidet.
- Qualitätsanalyse der Wikipedia-Georeferenzierung:
  - Der Benutzer wählt dazu das Wikipedia-Layer und das Layer mit den Referenzdaten. Folgende Fälle sind zu beachten: Das Wikipedia-Layer enthält das Attribut „name“ sowie ein oder kein „admin“-Attribut (Namenskonvention).
  - Das Referenz-Layer enthält immer ein Schlüsselattribut, ein Namensattribut sowie ein oder kein Attribut für das Verwaltungsgebiet. Schlüssel-, Namens- und optionales Attribut mit Angabe des Verwaltungsgebietes werden interaktiv ausgewählt (keine Namenskonvention).
  - Im ersten Schritt werden die zugehörigen Objektpaare aus Wikipedia- und Referenz-Layer bestimmt, indem eindeutige Paare aufgrund Gleichheit bezüglich Namen und (wenn gegeben) Verwaltungsgebiet bestimmt werden. Falls für den Namen eines Wikipedia-Objektes, das eine Ergänzung in Klammern besitzt (z.B. „Holzhausen (Wildeshausen)“) kein Partner im Referenz-Layer mit gleichem Namen gefunden wird, ist der Test auch ohne die Klammerergänzung (also z.B. mit „Holzhausen“) durchzuführen.
  - Es wird der horizontale und vertikale Abstand zwischen den minimalen und maximalen x- und y-Koordinaten im Referenz-Layer berechnet und davon der Mittelwert als Referenzabstand *r* genommen.

- Für die Objektpaare wird der jeweilige Abstand *d* bestimmt. Falls *d*=0 ist die Abstandsklasse 0; falls 0 < *d* < r/100000 ist die Abstandsklasse 1; falls *r*/100000 < *d* < *r*/10000 ist die Abstandsklasse 2; falls *r*/10000 < *d* < *r*/1000 ist die Abstandsklasse 3; falls *r*/1000 < *d* < *r*/100 ist die Abstandsklasse 4; falls *r*/100 < *d* < *r*/10 ist die Abstandsklasse 5 und falls *d* > *r*/10 die Abstandsklasse 6.
- Die Wikipedia-Objekte aus den Objektpaaren werden um die Attribute Abstand, Abstandsklasse und Kennnummer des Referenzpartners ergänzt und als neues Layer der Karte hinzugefügt. Die Gestaltung der Objekte erfolgt gemäß der Abstandsklasse.

Im Rahmen einer Abschlussaufgabe im Wintersemester 2013/14 für das Modul „GIS-Programmierung“ (5. Semester Bachelor-Studiengang Geoinformatik) sollten die Studierenden auf Basis von GeoTools und JTS eine entsprechende Anwendung entwickeln. Zu Untersuchungszwecken wurden folgende amtlichen offenen Geodaten bereitgestellt (jeweils mit nationalem Schlüssel):

- Gemeinden und Kreise in Deutschland (12.744 bzw. 439 Flächenobjekte)
- Urban Areas und Rural Centers in Neuseeland (276 Flächen)
- Localidades (3.560 Punktobjekte) und Departamentos (526 Flächen) in Argentinien

Die nachfolgenden Abbildungen zeigen exemplarisch eine so entwickelte Anwendung:

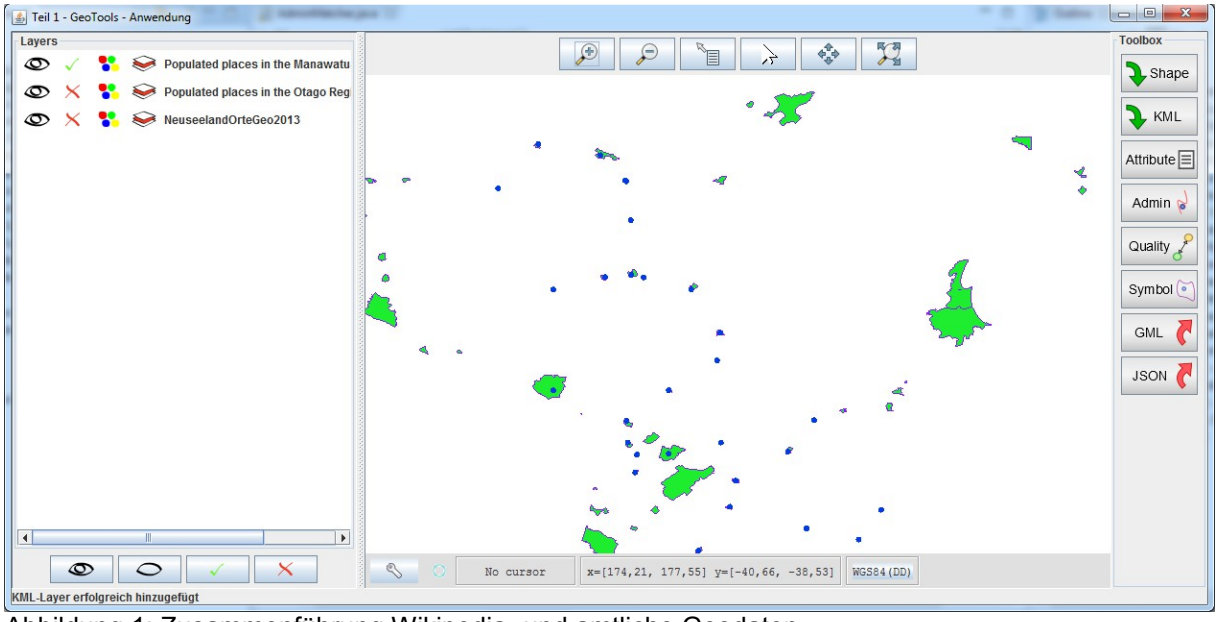

Abbildung 1: Zusammenführung Wikipedia- und amtliche Geodaten

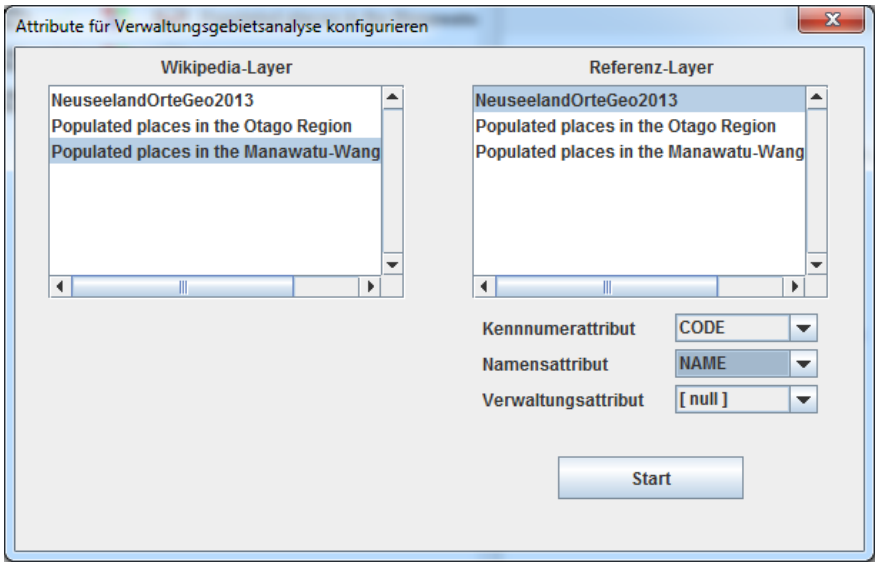

Abbildung 2: Konfiguration eines Vergleichs

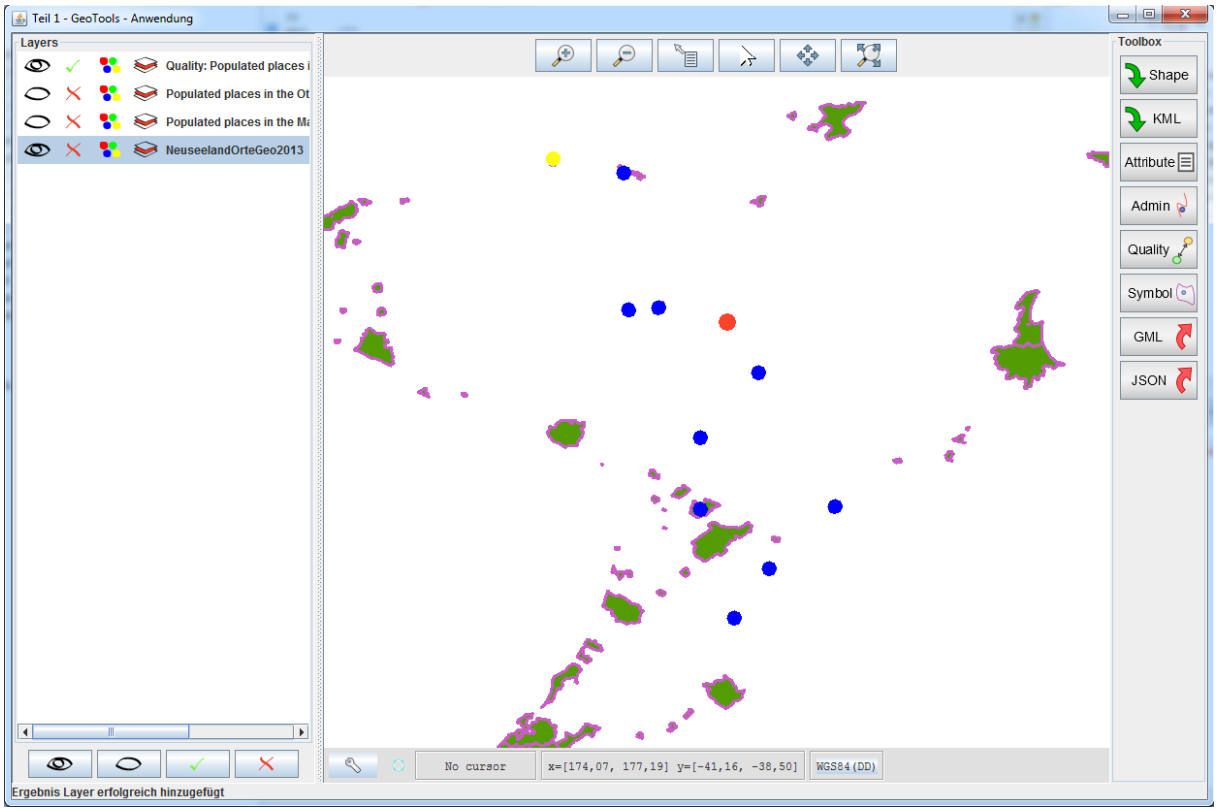

Abbildung 3: Resultat des Vergleichs (blau: Abstandsklasse 0, gelb: Abstandsklasse 1, rot: Abstandsklasse 3)

Fazit

Die Verfügbarkeit und Qualität von offenen Geodaten für Orte und administrative Gebiete ist recht heterogen. Durch die Aufnahme offener amtlicher Geodaten in die Datenbestände von gemeinschaftlichen Kartierungen konnte teilweise eine deutliche Steigerung der Qualität er-

reicht werden (z.B. durch den Import der BKG-Daten in OpenStreetMap). Weltweit ist die Situation noch unbefriedigend. Die Verknüpfung von GeoNames- und Wikipedia-Daten beispielsweise mit OSM-Daten [3] birgt aufgrund der beobachteten Eigenschaften durchaus Risiken. Durch freie Software-Bibliotheken lassen sich aber mächtige Werkzeuge schaffen, um solche Geodaten trotzdem anwendungsorientiert nutzen zu können.

Kontakt zum Autor:

Prof. Dr. Thomas Brinkhoff
Jade Hochschule Wilhelmshaven/Oldenburg/Elsfleth
Institut für Angewandte Photogrammetrie und Geoinformatik
Ofener Str. 16/10
26122 Oldenburg
thomas.brinkhoff@jade-hs.de
http://iapg.jade-hs.de/personen/brinkhoff/

Literatur

[1] *Heipke, Christian*: „Crowdsourcing geospatial data“, ISPRS Journal of Photogrammetry and Remote Sensing 65(6), 2010, 550-557.

[2] *Open Source Geospatial Foundation (OSGeo)*: OSGeo Projects, http://www.osgeo.org/

[3] *Stadler, C., Lehmann, J., Höffner, K., Auer, S.*: „Linkedgeodata: A core for a web of spatial open data. Semantic Web“, 3(4), 2012, 333-354.