Comparing OSM Area-Boundary Data to DBpedia

Doris Silbernagl
Department of Computer Science, University of Innsbruck, Austria
doris.silbernagl@uibk.ac.at

Nikolaus Krismer
Department of Computer Science, University of Innsbruck, Austria
nikolaus.krismer@uibk.ac.at

Günther Specht
Department of Computer Science, University of Innsbruck, Austria
guenther.specht@uibk.ac.at

ABSTRACT

OpenStreetMap (OSM) is a well known and widely used data source for geographic data. This kind of data can also be found in Wikipedia in the form of geographic locations, such as cities or countries. Next to the geographic coordinates, also statistical data about the area of these elements can be present. Since it is possible to extract these data from OpenStreetMap as well, it is sensible to examine the quality of the OSM information about those specific boundary elements and compare them to an also crowd-sourced source like Wikipedia. Hence, in this paper OSM data of different countries are used to calculate the area of valid boundary (multi-) polygons and are then compared to the respective DBpedia (a large-scale knowledge base extract from Wikipedia) entries.

CCS Concepts

•**Information systems** → **Geographic information systems;** *Multimedia databases;* •**Human-centered computing** → *Wikis;*

Keywords

OpenStreetMap, Data Validation and Verification, DBpedia

1. INTRODUCTION

The OpenStreetMap (OSM) project[1] has the aim of providing a free, digital map of the world. Everybody is able to contribute content to the project. The geographic elements that represent the real world are stored in OSM as nodes, ways and relations [17]. Each of the elements can be enriched with semantic data via tags. Those tags allow an identification of the type of object and the search for specific data. For the investigation in this paper, the boundary tags are of interest, specifically administrative boundaries which define cities, districts, states or similar. In order to draw conclusions about the quality of OSM data, the boundary-tagged relations using their geometries are investigated. Calculating the area of these objects and comparing them to other ground truth data sources, such as DBpedia [13] or open government data, is a way of verification for this OSM data.

This paper is organized as follows: beginning with Section 2 an overview of similar work in this topic is given. In Section 3 the methodology for data retrieval for OSM (Subsection 3.1) and from DBpedia (Subsection 3.2), as well as the workflow from data generation to storage performed by the tool are shortly described. The subsequent Section 4 presents the results of the comparison of the two data sources and evaluates how well the data matches. The paper closes with a short conclusion and outlook to future work (see Section 5).

2. RELATED WORK

The freedom of global participation is a widely discussed topic regarding the quality of such volunteered geographic information (VGI) [10, 12], like the one in OSM. Many aspects have to be considered when performing analyses and making assertions about the data and its quality. There exist scientific investigations about the definition of quality parameters for VGI data. In the very beginnings of research in this topic, Veregin [23] started to develop some basic definitions of quality parameters for geospatial data: accuracy (spatial, temporal and thematic), precision (spatial, temporal and thematic resolution), consistency (logical) and completeness. Ciepłuch et al. [9] listed as possible quality indicators the length of particular features, the density of data points within grid squares, as well as the contributor's profile/characteristics and history. Mooney et al. [15] investigated shape similarities of polygons representing lakes in OSM with commercial, publicly available government-generated spatial data. Empirical studies about the data quality of OSM were also executed by Neis et al. [16] and Zielstra et al. [26]. In their studies they compared the street network and points of interest of Germany from OpenStreetMap to the commercial dataset of Teleatlas. Another comparative work about commercial geographic data versus OSM data was done by Haklay et al. [12]. They examined, among other things, the positional accuracy of roads in OSM and compared the data to the official governmental dataset Ordnance Survey of the UK.

Taking this idea of comparison to other publicly available data sources, this paper presents the investigation of OSM

[1] www.openstreetmap.org

OpenSym '16, August 17 - 19, 2016, Berlin, Germany

ISBN 978-1-4503-4451-7/16/08...$15.00

DOI: http://dx.doi.org/10.1145/2957792.2957806

data compared to another non-commercial, crowd-sourced data source, namely DBpedia. This source is a large-scale, multilingual knowledge base extracted from Wikipedia [13]. Data quality observations were already executed [8] in this area, hence it can be used as ground truth data [24].

3. METHODOLOGY

In order to execute the desired comparison of OSM data to the DBpedia information, the data needs to be retrieved from both sources. For OSM a preprocessing step is required that allows the further usage of the data for the purpose of area calculation. On the other side, for DBpedia a SPARQL query is defined to retrieve the desired information. Both steps are described in the following subsections. Figure 1 depicts the overall workflow of the steps done.

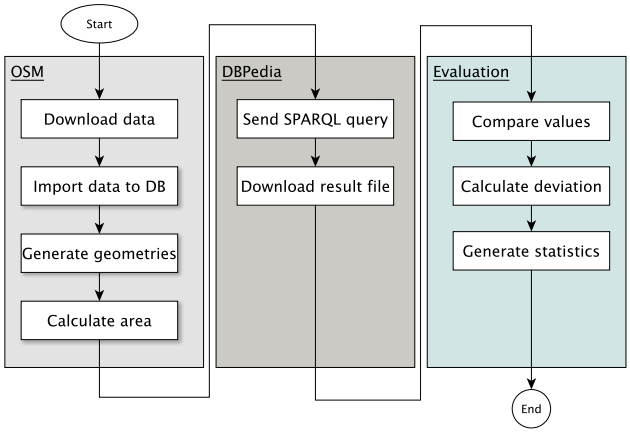

Figure 1: Workflow steps from data retrieval to evaluation

3.1 Data retrieval for OSM

For OSM, the data of various countries are taken from Geofabrik [2] (an organization that offers prefabricated OSM files for regions and countries) and imported into a PostgreSQL/PostGIS [5, 4] database with the "osmosis" tool (version 0.44.1) [18] in the "pg_snapshot_simple" schema. PostGIS, which is the geo-information extension for the PostgreSQL database, delivers many geographic specific functions, such as area calculation. The chosen import tool with its schema is advantageous when looking for a clear, separate storage between ways (linestrings), nodes (points) and relations (e.g. polygons, street network), but also for different applications to the database, such as applying change sets, comparison of different dump files, etc. Having stored the desired data in the database, it is necessary for the comparison step to first retrieve all boundary relations in the dataset. For this, a SQL query is used that requests the OSM relations with the respective tagging: "$type = boundary$", "$boundary = administrative$", "$wikipedia$". The first tag specifies a relation to be a boundary, the second one reduces the boundaries to administrative ones, being countries, cities, districts or similar. The last one selects only relations that have a Wikipedia link. This is necessary for finding the correct counterpart for reconciliation in DBpedia. In total, currently 0,02% of all stored elements in OSM (3.6 billion elements, of those only 1.3 billion are tagged, statistics from 2016-04-07) include this Wikipedia tag, according to the taginfo website [22], a statistical site that analyzes OSM tags. Nevertheless, this number should be still high enough to find entries in DBpedia. This can be clearly observed in the evaluation section 4 of this paper.

In the next step, after the retrieval of the relevant relations, the area of these remaining entries is of interest. To calculate the area of these boundaries in order to compare it to DBpedia, the geometry of the boundary objects needs to be present. Extracting this information in the PostgreSQL/PostGIS database with the database schema described above, is only possible for linestring (ways) and point (nodes) by using just simple SQL queries (the geometry is explicitly present). However, for more complex relations like (multi-)polygons it has to be computed. This and the fact that the process of requesting complex elements directly in the database is quite cumbersome and computationally intensive led to the development of a self-implemented tool presented in Silbernagl et al. [19]. Its algorithm allows the generation and storage of (multi-)polygon geometries in the database itself, as well as in JTS (Java Topology Suite), a Java library for GIS objects [20]. This makes a further processing of the data possible, especially for (intrinsic) OSM data analyses. Mathematical calculations, such as area computation, can be executed either with PostGIS [4] functions in the database itself or in Java with JTS functions.

3.2 Data retrieval from DBpedia

Having executed the tool and the area calculation process, the next step is to find and retrieve the respective data from DBpedia. For this, a SPARQL query is used and sent to the DBpedia SPARQL API. The query includes the value of the Wikipedia tag from OSM to specify the RDFS label (with language tag) used during data retrieval. Depending on the type of element, i.e. if it is a city or district, the request for the area varies. Therefore, optional fields are comprised in the SPARQL query that ask for different possible representations of the area property. These values are very inconsistently used in DBpedia and consequently the most encountered ones during evaluation are used. They vary from $populatedPlace : areaTotal$ to $[property] : area*$ (where $*$ marks different additional values like $areaTotalKm$), to language specific labels like, e.g. in German "fläche".

The reason for requesting DBpedia in this way is quite obvious: using a dump of DBpedia implies too much overhead for these types of requests and parsing static n-triple files for such specific information is not of high performance as well. As in Section 4 can be seen, the number of SPARQL queries for the selected data sets is rather low, so the straightforward way of getting the desired data is to use the SPARQL API of DBpedia.

Having explained what data is used and how the area values are gained, the following Section 4 shows the evaluation of the comparison of the OSM data to DBpedia.

4. EVALUATION

In order to have some amount of data, different datasets are used for evaluation. The countries Austria, UK and the US Northeast region are selected. From the literature it is known that the OSM data of these countries are of high quality [11, 7, 21, 12, 25] and hence are suitable for this experiment.

Table 1 includes the facts about the database sizes of the

Table 1: Dataset facts

Dataset	Size (GB)	Relations	Boundaries	Geometries	DBpedia entries
Austria	16	84747	1978	1679	1669
UK	33	200258	725	699	686
US–NE	25	45351	4201	4144	4133

OSM data, number of relations in OSM and how many of those are boundary-tagged ones. For these, the geometries are produced and stored in the database (number shown in column four of Table 1). It should be noted that the number of boundary tagged relations and the ones where geometries are effectively generated and saved is different. This is because occasionally polygons are either invalid or the relation data is incomplete. The last case arises due to the fact that the countries are cut out of the world map. The reference to a member of a relation may still be included in the subset, but the actual element is not.

For all geometries the area values are calculated within the database. The respective entries are then queried with SPARQL from DBpedia. The number of DBpedia entries polled is shown in the last column of Table 1. The total number of boundaries that are requested from DBpedia varies from the number of stored boundaries in the database. This is due to false tagging in OSM, i.e. some relations include the same value in their Wikipedia tag. In this case the first entry found is taken for matching. This may lead to wrong values in the comparison. However, for the datasets used only a maximum of 13 entries are affected and thus have a negligible impact on the final outcome.

For the evaluation the values of OSM to DBpedia are matched and the percentage of congruence is calculated. As unit square kilometers are used, thus adjustments needed to be made for deviating values that are present in square meters or miles.

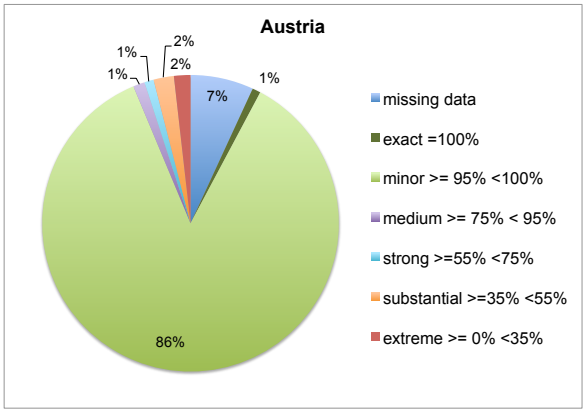

Figure 2: Statistics of the Austria dataset compared to DBpedia.

Figures 2, 3 and 4 show the result of the comparison to DBpedia for each dataset. In Figure 2 it can be seen, that for Austria about 1% of all found data matches exactly, more than 86% only differ 5% or less. Extreme deviation can only be observed for about 2% of all entries (30 out of 1669). Fortunately, for the majority of the requests (93%) data could

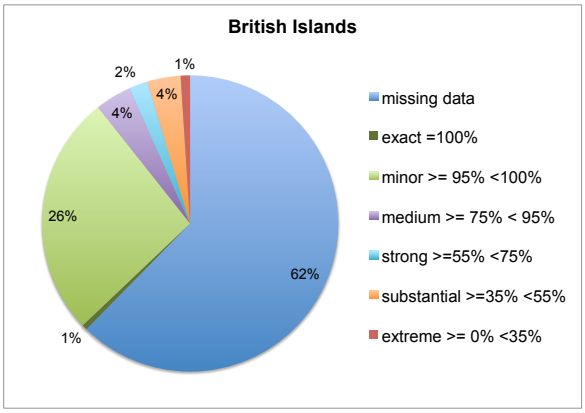

Figure 3: Statistics of the UK (British Islands) dataset compared to DBpedia.

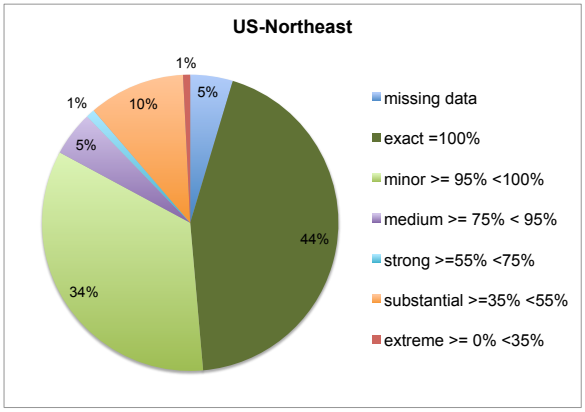

Figure 4: Statistics of the US Northeast dataset compared to DBpedia.

be found in DBpedia, only 7% of all polled boundaries delivered no result. Checking these entries manually at DBpedia led to the insight that these values are simply missing.

For the UK, pictured in Figure 3, it can be noticed that although the SPARQL query is very flexible and checks various labels for area specifications, still many data is missing: 62% of the requested information could not be found in DBpedia. As the number is quite high, manual checks for entries have been executed in order to investigate the problem. It is the case that for those checked entries, actually no area value is present at all. However, for entries that have some values specified only a few show a 100% accuracy (4 out of a total of 686 boundaries), 26% possess a discrepancy of 5% or less and for the other categories a small amount (<12%) reveal a higher divergence from the DBpedia data.

In Figure 4 it is visible that for the US Northeast region the comparison delivered quite well-distributed findings. Only 5% of the data is missing, 44% matches exactly. Also, it can be seen that for 34% the deviation to the DBpedia data is 5% or less. The medium, strong and substantial discrepancies sum up to 17% in total. With only less than 1% of extreme differences (13 out of 4133) to the ground truth dataset, the final results can be declared to be quite well.

5. CONCLUSION

The idea of comparing one crowd-sourced dataset to another seems to be promising for the selected objects. OpenStreetMap includes boundary-tagged elements for different regions of the world, allowing to perform statistics about the data. As the evaluation showed, DBpedia seems to lack these information for some parts of the world. However, in total the discrepancy of the polygon boundaries for Austria to the values stored in DBpedia were minor. This may be because the Austrian government makes these statistical data publicly available via open government data platforms [3]. Also, the regionally available geographic data set is of high quality [1]. For the UK it is more difficult to find a valid statement about the quality, as many data was missing in DBpedia. It is hard to tell why such little data for this kind of information is available in the UK. For OSM the data in the UK was investigated by Haklay et al. [12]. The rather high matching score in the US may be caused by the import of governmental data from the TIGER bureau [6, 14]. The effects of this import into OSM were examined by Zielstra et al. [25].

For future work it is imaginable to perform more evaluations for various other countries of the world. Presumably, DBpedia is, similar to OSM, regionally dependent. People contribute information and publicly or governmental data is not available everywhere. It would be interesting to investigate on similar phenomenons by examining different geoinformation like accuracy of geo-location data, especially for POIs, or the size of other elements such as forests, parks, etc.

6. REFERENCES

[1] basemap.at-Verwaltungsgrundkarte Österreich. http://www.basemap.at. (2016-04-08).

[2] Geofabrik GmbH. http://www.geofabrik.de. (2015-10-28).

[3] Offene Daten Österreich. https://www.data.gv.at. (2016-04-07).

[4] PostGIS. http://www.postgis.net. (2015-10-28).

[5] PostgreSQL. http://www.postgresql.org. (2015-10-28).

[6] TIGER/Line - Geography - U.S. Census Bureau. https://www.census.gov/geo/maps-data/data/tiger-line.html. (2016-04-07).

[7] A. L. Ali and F. Schmid. Data quality assurance for volunteered geographic information. In *Geographic Information Science*, pages 126–141. Springer, 2014.

[8] D. Anthony, S. W. Smith, and T. Williamson. The quality of open source production: Zealots and good Samaritans in the case of Wikipedia. *Rationality and Society*, 2007.

[9] B. Ciepłuch, P. Mooney, and A. Winstanley. Building generic quality indicators for OpenStreetMap. 2011.

[10] A. K. Cooper, S. Coetzee, I. Kaczmarek, D. G. Kourie, A. Iwaniak, and T. Kubik. Challenges for quality in volunteered geographical information. 2011.

[11] S. Gröchenig, R. Brunauer, and K. Rehrl. Digging into the history of VGI data-sets: results from a worldwide study on OpenStreetMap mapping activity. *Journal of Location Based Services*, 8(3):198–210, 2014.

[12] M. Haklay et al. How good is volunteered geographical information? A comparative study of OpenStreetMap and Ordnance Survey datasets. *Environment and planning. B, Planning & design*, 37(4):682, 2010.

[13] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, et al. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web*, 6(2):167–195, 2015.

[14] R. W. Marx. The TIGER system: automating the geographic structure of the United States census. *Government Publications Review*, 13(2):181–201, 1986.

[15] P. Mooney, P. Corcoran, and A. C. Winstanley. Towards quality metrics for OpenStreetMap. In *Proceedings of the 18th SIGSPATIAL international conference on advances in geographic information systems*, pages 514–517. ACM, 2010.

[16] P. Neis, D. Zielstra, A. Zipf, and A. Strunck. Empirische Untersuchungen zur Datenqualität von OpenStreetMap-Erfahrungen aus zwei Jahren Betrieb mehrerer OSM-Online-Dienste. *Angewandte Geoinformatik 2010*, 2010.

[17] OSM Foundation. OpenStreetMap Wiki. http://wiki.openstreetmap.org/. (2015-12-09).

[18] OSM Foundation. Wiki OpenStreetMap Osmosis. http://wiki.openstreetmap.org/wiki/Osmosis. (2016-04-05).

[19] D. Silbernagl, N. Krismer, and G. Specht. osmpg2java - Konvertierung von OSM Datenbankelementen zu JTS Objekten. *Angewandte Geoinformatik AGIT*, 2016.

[20] V. Solutions. JTS topology suite. *Developer's Guide*, 795, 2003.

[21] R. Steinmann, R. Brunauer, S. Gröchenig, and K. Rehrl. *Wie aktiv sind freiwillige Mapper?: ein Vergleich der OpenStreetMap-Aktivitäten in den Jahren 2005-2012 am Beispiel der DACH-Region.* Wichmann Verlag, 2013.

[22] J. Topf. http://taginfo.openstreetmap.org. (2016-04-08).

[23] H. Veregin. Data quality parameters. *Geographical information systems*, 1:177–189, 1999.

[24] A. Zaveri, D. Kontokostas, M. A. Sherif, L. Bühmann, M. Morsey, S. Auer, and J. Lehmann. User-driven Quality Evaluation of DBPedia. In *Proceedings of the 9th International Conference on Semantic Systems*, pages 97–104. ACM, 2013.

[25] D. Zielstra, H. H. Hochmair, and P. Neis. Assessing the effect of data imports on the completeness of OpenStreetMap–a United States case study. *Transactions in GIS*, 17(3):315–334, 2013.

[26] D. Zielstra and A. Zipf. A comparative study of proprietary geodata and volunteered geographic information for Germany. In *13th AGILE international conference on geographic information science*, volume 2010, 2010.