An Approach to gecoding based on volunteered Spatial Data

Christof AMELUNXEN
Department of Geography, Chair of Geoinformatics, University of Heidelberg

Abstract. The automated process of assigning geographic coordinates to textual descriptions of a place, generally referred to as *geocoding*, plays an important role in various fields of geographic information technologies, ranging from the analysis of healths records [28] or crime incidents [17] to location based services like route planning applications [20]. However, since the collection and maintenance of appropriate spatial data is the traditional domain of official surveying offices and commercial companies, there are only very few publicly available geocoding services which can be used free of charge, and those which exist are usually limited to a specific country or even smaller units. Furthermore, no freely available geocoding service offering house number level precision has ever been implemented based on volunteered geographic data. The goal of the work summarized in this paper (originating from the author's MSc thesis [1]) was thus to explore the suitabilty of freely available volunteered geographic information[1] for the purpose of geocoding and the findings of the conducted research are able to serve as a proof of concept for the usage of volunteered spatial data as a reference dataset for geocoding services. The geocoder itself has been integrated into the OpenRouteService [20-22] project, providing a framework compliant to the OpenGIS Location Service (OpenLS) [23] specifications.

Keywords. Geocoding, Volunteered Geographic Information, OpenStreetMap, OpenRouteService, OpenGIS Location Service (OpenLS)

Introduction

Until recently, the generation, maintenance and distribution of geographic information had been, with only very few exceptions, solely the domain of either official land surveying offices or commercial companies. This was presumably mainly due to the immense costs related to the actual surveying and maintenance and the lack of possibilities to effectively share and distribute the collected spatial data. However, this has recently changed, for the two following reasons[2]:

1. The dramatically reduced costs along with the enhanced usability of modern satellite navigation handheld devices have enabled a mass of people to collect geographic data with ease of use and in precision levels which had formerly been simply beyond reach for private persons.
2. The progress of the internet from a formerly "read-only media" to the "web 2.0" participatory approach has made collaborative efforts to generate and share content of various kinds very common.

Among a broad list of projects dealing with user generated geographic information, OpenStreetMap is one of the most promising. Its primarily goal is to generate a free map of the world [24] through volunteered effort. Nevertheless, although the generation of maps still is the focus of the project, the collected spatial data is made publicly available and may be used for other purposes as well. OpenRouteService[3] e.g. is an example of a project which has successfully implemented a routing service based on OpenStreetMap data.

However, the focus of the work presented in this paper was to explore the suitability of OpenStreetMap data for the purpose of geocoding, simplified as the conversion of textual address information into point coordinates and vice versa[4].

If a working geocoding service could successfully be build based on OpenStreetMap data, this would be a substantial advance in the improvement and progression of a wide range of projects, based in the field of volunteered geographic information.

A major objective of the work was further to evaluate the possibilities to compensate incomplete data (namely house number positions) by probability based approaches to locate house number positions, supported by official house numbering guidelines for the study area of Northrhine-Westfalia, Germany.

[1] Goodchild [13] proposed the term *Volunteered Geographic Information (VGI)* for geographic information generated by collaborative volunteered effort

[2] based on suggestions by Goodchild [14]

[3] http://www.openrouteservice.org/

[4] The definition and usage of the term geocoding varies in scientific literature. Some authors limit the scope of input data to postal addresses [2-3, 5] whereas others widen the scope to include named places [7] or even arbitrary textual representations of a place [12, 25]

1. Approach

The *OpenStreetMap* project has been selected as the data source for this research as it provides an impressively extensive database originating from collaborative volunteered effort and the exponential growth of the project data since its start in 2004 is very promising. The first task was thus to analyze the data provided by the project and to develop an appropriate process to transform the data in a format usable for geocoding purposes. The next task has then been the actual design and implementation of the geocoding application.

At first, the general suitability of the OpenStreetMap data for geocoding purposes has been evaluated with respect to its data model, relational integrity and completeness. Based on this analysis the proposed data model for the geocoder's reference dataset was designed and an appropriate data transformation and integration processes have be developed following concepts presented by Han and Kamber [16] and Rahm and Do [26].

This has been followed by the definition and analysis of uses cases to be provided by the geocoding service. The actual processing of the geocoding use cases has then been designed following standard geocoding practices as described by Goldberg [12], Davis et al. [7], Borkar et al. [4] and Christen and Churches [6].

The treatment of incomplete house number data received special attention. In order to compensate missing house number data in OpenStreetMap, different probability based approaches had been developed in order to effectively approximate house number locations.

This included the analysis of house numbering systems in general [9-10] and research concerning habits and regulations for house number assignment within the study area[5] [29] in order to construct and evaluate hypotheses for the approximation of house number locations. This part was based on the work of Goldberg [12], Ratcliffe [27] and Bakshi et al. [2].

The quality of the geocoder, implemented according to the concepts and guidelines developed before, was finally measured using the standard key figures *match rate* and *positional accuracy* as described by Cayo and Talbot [5] and additionally by comparing the positional accuracy measured to a commercial geocoding service provided by Google™.

2. Results

The match rate, defined as the percentage of geocoding requests returning a correct match, has been found to be 96% on municipal level requests (sample size *n=333*), 83% on street level requests (*n=1000*) and 5% (*n=1000*) on house number level requests for randomly chosen addresses within the study area.

When considering a match rate of 85% to be the minimum acceptable rate necessary to reliably detect spatial patterns in address datasets (as proposed by Ratcliffe [27]) it has to be concluded that the achieved match rate on street and house number level is not yet sufficient for detailed spatial analysis purposes.

The average positional error for house number level requests (determined by comparing the results to the real positions of the buildings as provided by the surveying office for the study area) was measured differentiating the availability of housenumber positions in the OpenStreetMap data (see table 1).

Table 1. Geocoding accuracy of ORS geocoder depending on housenumber data availability in OSM

Location Method	Sample Size	Mean Positional Error
Exact housenumber match	13933	11m
Interpolation between known housenumber positions	890	31m
No housenumber data available	255073	142m

These figures have to be considered not suitable for fine-scale spatial analyses of address datasets unless housenumber data is available. Zandbergen [31] e.g. showed that even a medium error of 41 meters with a 90th percentile of just 100 meters can significantly bias analysis results on the example analysis of traffic-related air pollution on school children (using a sample of 104,865 addresses).

The average positional accuracy achieved when interpolation between two known housenumber positions was possible is nevertheless significantly better than the medium error of 41m measured by Zandbergen [31] for 104,865 sample addresses located in Orange County, Floria, geocoded using official street centerline and parcel data of the Property Appraisers Office of Orange County.

The measured medium positional error of merely 11m for exact house number matches can be considered to be an extraordinary accuracy. Literature research revealed no case study presenting a geocoding service providing accuracy figures even close [5, 8, 11, 15, 18, 19, 27, 30].

A comparison with the accuracy provided by the geocoding service offered by Google™[6] showed that whenever house number data was available, the positional error was significantly lower than Google's (see table 2 and figure 1) and about equal when interpolation between two known house numbers was possible. Yet for the

[5] The study area for this research has been the federal state Northrhine-Westfalia in Germany

[6] http://code.google.com/apis/maps/documentation/geocoding/

case when no house number data was available, the average positional accuracy proved significantly worse than the one provided by Google™.

Table 2. Comparing the positional accuracy of ORS geocoder and Google geocoder depending on housenumber availability in OSM

Location Method	Sample Size	Mean Error ORS	Mean Error Google
Exact housenumber match	13283	11m	32m
Interpolation between known housenumber positions	853	31m	32m
No housenumber data available	54889	142m	34m

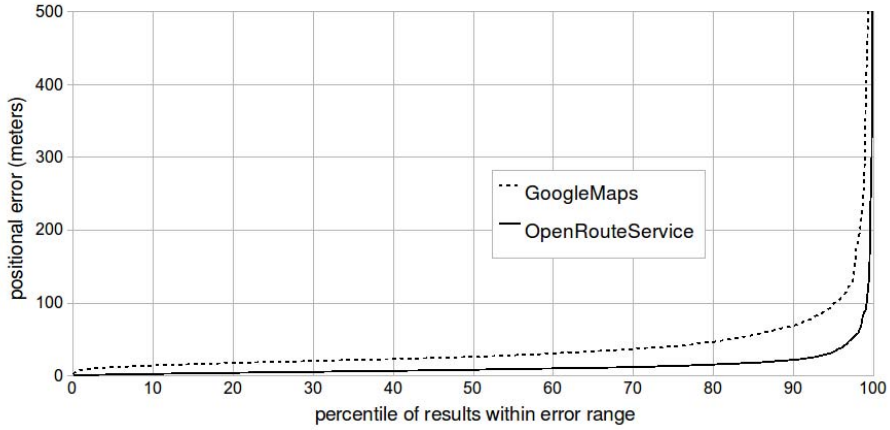

Figure 1. Positional errors of OpenRouteService and GoogleMaps geocoder when house number data available in OSM

It could further be demonstrated that it is indeed possible to effectively approximate house number locations using probability based approaches based on hypotheses according to official house numbering guidelines. For a sample set of cities within the study area, an improvement of up to 52% in terms of positional accuracy could be achieved [1]. It was nevertheless as well found that the effectiveness of these approaches, although showing a significant overall average improvement, depends heavily on the suitability of the study area. It was further found that these improvements are still not sufficient to generate accuracy levels comparable to cases where actual house number data was available.

3. Conclusion

The findings of the conducted research are able to serve as a proof of concept for the usage of volunteered spatial data as a reference dataset for geocoding services. The inherent inconsistencies present in the OpenStreetMap data however required substantial concessions in terms of referential integrity and the positional accuracy to be expected strongly depends on the availability of house number data, although means to partially compensate incomplete data have been successfully developed.

The result of this work is already used in production as the geocoding engine for various research projects, of which *OpenRouteService*[7] and *OSM-3D*[8] presumably being the most prominent.

The recent development of the OpenStreetMap project is very promising, too. Within a mere four months during the implementational phase of the research presented, the amount of house number locations for the area of Germany, stored in the OpenStreetMap database, has almost doubled from approximately 172,000 house numbers at the end of December 2008 to more than 330,000 house numbers at the end of April 2009. At time of writing (February 2010) there are around 600,000 housenumbers in the database for the area of Germany and about 4.5 million for the area of Europe.

References

[1] Christof Amelunxen. An approach to geocoding based on volunteered spatial data. Master's thesis, Center for Geoinformatics (Z_GIS), University of Salzburg, 2009.

[2] Rahul Bakshi, Craig A. Knoblock, and Snehal Thakkar. Exploiting online sources to accurately geocode addresses. In Proceedings of the 12th annual ACM international workshop on Geographic information systems, pages 194–203, Washington DC, USA, 2004. ACM. ISBN 1-58113-979-9.

[7] http://www.openrouteservice.org/

[8] http://www.osm-3d.org/

[3] Franz-Josef Behr, Astrit Rimayanti, and Hui Li. Opengeocoding.org - a free, participatory, community oriented geocoding service. Technical report, Department of Geomatics, Computer Science and Mathematics, University of Applied Sciences Stuttgart, Stuttgart, Germany, 2008.
[4] Vinayak R. Borkar, Kaustubh Deshmukh, and Sunita Sarawagi. Automatic segmentation of text into structured records. In Proceedings of the SIGMOD Conference, Santa Barbara, California, 2001.
[5] Michael R Cayo and Thomas O. Talbot. Positional error in automated geocoding of residential addresses. International Journal of Health Geographics, 2:10, December 2003. doi: 10.1186 1476-072X-2-10.
[6] Peter Christen and Tim Churches. A probabilistic deduplication, record linkage and geocoding system. In Proceedings of the ARC Health Data Mining workshop, Adelaide, Australia, April 2005.
[7] Clodoveu Davis, Frederico Fonseca, and Karla A. V. Borges. A flexible addressing system for approximate geocoding. In Brazilian Symposium on GeoInformatics, 2003.
[8] SM Dearwent, RR Jacobs, and JB Halbert. Locational uncertainty in georeferencing public health datasets. Journal of Exposure Analysis and Environmental Epidemiology, 11:329–334, 2001.
[9] Catherine Farvacque-Vitkovic, Lucien Godin, Hugues Leroux, Florence Verdet, and Roberto Chavez. Street Addressing and the Management of Cities. The International Bank for Reconstruction and Development / The World Bank, Washington, DC, USA, 2005.
[10] Peter Fonda-Bonardi. House numbering systems in los angeles. In GIS/LIS ‘94 Proceedings, pages 322–331. Los Angeles County Urban Research Section, 1994.
[11] Daniel Goldberg, John Wilson, Craig Knoblock, Beate Ritz, and Myles Cockburn. An effective and efficient approach for manually improving geocoded data. International Journal of Health Geographics, 7(1):60, 2008.
[12] Daniel W. Goldberg. A Geocoding Best Practices Guide. University of Southern California, GIS Research Laboratory, 2008.
[13] Michael F. Goodchild. Citizens as sensors: the world of volunteered geography. GeoJournal, 69(4):211–221, 2007.
[14] Michael F. Goodchild. Citizens as sensors: Web 2.0 and the volunteering of geographic information. Geofocus, 7:8–10, 2007.
[15] Tony H. Grubesic and Alan T. Murray. Assessing positional uncertainty in geocoded data. In Proceedings of the 24th Urban Data Management Symposium, Chioggia, Italy, 2004.
[16] Jiawei Han and Micheline Kamber. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management. Diane Cerra, San Francisco, USA, 2nd edition, 2006. ISBN 1558604898.
[17] Keith Harris. Mapping Crime: Principle And Practice. Diane Pub Co, 1999.
[18] Nancy Krieger, Pamela Waterman, Kerry Lemieux, Sally Zierler, and Joseph W. Hogan. On the wrong side of the tracts? Evaluating the accuracy of geocoding in public health research. American Journal of Public Health, 91(7):1114–1116, 2001.
[19] Soumya Mazumdar, Gerard Rushton, Brian Smith, Dale Zimmerman, and Kelley Donham. Geocoding accuracy and the recovery of relationships between environmental exposures and health. International Journal of Health Geographics, 7(1):13, 2008.
[20] Pascal Neis. Location based services mit openstreetmap daten. Master’s thesis, Fachhochschule Mainz Fachbereich I, 2008.
[21] Pascal Neis and Alexander Zipf. Zur kopplung von opensource, openls und openstreetmaps in openrouteservice.org. In Proceedings of AGIT, Salzburg, Austria, 2008.
[22] Pascal Neis and Alexander Zipf. Openrouteservice.org is three times open: Combining opensource, openls and openstreetmaps. In Proceedings of the GISRUK 2008 conference, Manchester, April 2008. UNIGIS UK.
[23] OGC. Opengis location service (openls) implementation specification: Core services, Sep 2008.
[24] OpenStreetMap. The free wiki world map, 2009. URL http://www.openstreetmap.org/.
[25] Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tom De Groeve. Geographical information recognition and visualization in texts written in various languages. In SAC ’04: Proceedings of the 2004 ACM symposium on Applied computing, pages 1051–1058, New York, USA, 2004. ACM.
[26] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. IEEE Data Engineering Bulletin, 23:2000, 2000.
[27] Jerry H. Ratcliffe. On the accuracy of tiger-type geocoded address data in relation to cadastral and census areal units. Geographical Information Science, 15:473–485, 2001.
[28] Gerard Rushton, Marc P. Armstrong, Josephine Gittler, Barry R. Greene, Claire E. Pavlik, Michele M.West, and Dale L. Zimmerman. Geocoding in cancer research: A review. American Journal of Preventive Medicine, 30:16–24, 2006.
[29] Städtetag NRW. Richtlinien für die nummerierung von gebäuden oder bebauten grundstücken, 1979.
[30] Eric A. Whitsel, Kathryn M. Rose, Joy L. Wood, Amanda C. Henley, Duanping Liao, and Gerardo Heiss. Accuracy and repeatability of commercial geocoding. American Journal of Epidemiology, 160(10):1023–1029, 2004.
[31] Paul Zandbergen. Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. BMC Public Health, 7(1):37, 2007.