

On the suitability of Volunteered Geographic Information for the purpose of geocoding

Christof AMELUNXEN

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 health records or crime incidents to location based services like route planning applications. 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 had so far been implemented based on volunteered geographic data. The goal of the work summarized in this paper (originating from the author's MSc thesis) was thus to explore the suitability of volunteered geographic information for the purpose of geocoding.

1 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 (based on suggestions by GOODCHILD (2007b):

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.

The OpenStreetMap (OSM) 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. Its primary goal is to generate a free map of the world 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 well². OpenRouteService (ORS) (<http://www.openrouteservice.org/>) e.g. is an example of a project that has successfully implemented a routing service based on OpenStreetMap data.

The definition and usage of the term geocoding varies in scientific literature. Some authors limit the scope of input data to postal addresses (BAKSHI et. al. 2004, BEHR et. al. 2008, CAYO & TALBOT 2003) whereas others widen the scope to include named places (DAVIS et. al. 2003) or even arbitrary textual representations of a place (GOLDBERG 2008, POULIQUEN et. al. 2004). 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. 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 for incomplete data. Because house number data in OpenStreetMap is still rare and inhomogeneously distributed (some areas are almost completely mapped whereas others contain no house number data at all) one specific challenge of this work was to find out, whether the location of a house number along a given street may be effectively approximated by probability based approaches. In other words, the question to be answered was: “Is it possible to effectively approximate the position of a house number along a given street in the absence of real house number data?”. The term “effectively” in this case was meant in the sense of “better than simply returning the centerpoint of the street”.

2 Approach

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 was the actual design and implementation of the geocoding application.

The geocoding application has been integrated into the OpenRouteService (NEIS 2008, NEIS & ZIPF 2008a, NEIS & ZIPF 2008b) project, providing a framework compliant to the OpenGIS Location Service (OGC 2008) specification.

At first, the general suitability of the OpenStreetMap data for geocoding purposes was 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

² GOODCHILD (2007a) proposed the term “Volunteered Geographic Information (VGI)” for this type of geographic data, which is generated by collaborative volunteered effort

an appropriate data transformation and integration processes were developed following concepts presented by HAN & KAMBER (2006) and RAHM & DO (2000).

This has been followed by the definition and analysis of the use cases to be provided by the geocoding service. The actual processing of the geocoding use cases was then designed following standard geocoding practices as described by GOLDBERG (2008), DAVIS et. al. (2003), BORKAR et. al. (2001) and CHRISTEN & CHURCHES (2005).

The treatment of incomplete house number data received special attention. In order to compensate for missing house number data in OpenStreetMap, probability based approaches were developed in order to effectively approximate house number locations. The approximation is based on the following parameters:

1. Average distance between two house numbers
2. Average offset of the first house from the beginning of the street
3. Direction of the street

This approach is limited to the sequential alternating house numbering system that is used in most parts of Europe (FARVACQUE-VITKOVIC et. al. 2005). Given these parameters are known, the position of a house number may be approximated as shown in figure 1.

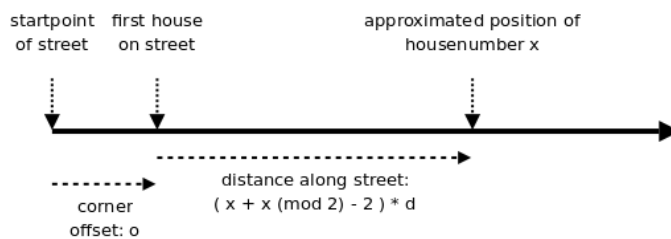


Fig. 1: Calculating approximate house number positions along a street.

The first two of these parameters were determined by spatial analyses of known reference datasets including house numbers. In order to guess the direction of a street, official guidelines on house numbering available for the study area - the federal state of Northrhine-Westfalia, Germany - were consulted (STÄDTETAG NRW 1979), resulting in the following hypotheses:

1. If only one of the end points of a street is connected to another street, the street starts at this point (dead-end street approach).
2. If a street proceeds away from the city center in a radial direction, the street starts at the end point closest to the city center (city center approach).
3. If a street is at one end connected to a street of a higher rank than at the other end point, the street starts at the end point connected to the higher level street (street rank approach).

3 Results

The quality of the geocoder, implemented according to the concepts and guidelines developed before, was measured using the standard key figures match rate and positional accuracy as described by CAYO & TALBOT (2003) and additionally by comparing the positional accuracy measured to a commercial geocoding service provided by Google™. The match rate, defined as the percentage of requests returning a correct match, was 96% at the municipal level requests (sample size $n = 333$), 83% at the street level requests ($n = 1000$), and 5% ($n = 1000$) at the 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 (2003), it has to be concluded that the achieved match rate at the 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 house number positions in the OpenStreetMap data (see table 1).

Table 1: Geocoding accuracy of ORS geocoder depending on the house number data availability in OSM.

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

These figures must be considered as non-suitable for fine-scale spatial analyses of address datasets unless house number data is available. ZANDBERGEN (2007) e.g. demonstrates that even a medium error of 41 meters with a 90th percentile of just 100 meters can significantly bias analysis results as shown on the example analysis of traffic-related air pollution affecting school children (using a sample of 104,865 addresses). The average positional accuracy achieved when interpolation between two known house number positions was possible, is nevertheless significantly better than the medium error of 41m measured by Zandbergen for 104,865 sample addresses located in Orange County, Florida; these addresses were geocoded using official street centerline and parcel data of the Property Appraisers Office of the Orange County.

The measured medium positional error of merely 11m for exact house number matches can be considered as an extraordinary accuracy. Literature research revealed no case study presenting a geocoding service providing accuracy figures even close (Cayo & Talbot 2003, Dearwent et. al. 2001, Goldberg et. al. 2008, Grubestic & Murray 2004, Krieger et. al. 2001, Mazumdar et. al. 2008, Ratcliffe 2001, Whitsel et. al. 2004).

A comparison with the accuracy provided by the geocoding service offered by Google™ (see: <http://code.google.com/apis/maps/documentation/geocoding/>) shows that whenever

house number data was available, the positional error was significantly lower than Google's (see table 2 and figure 2) and about equal when interpolation between two known house numbers was possible. Yet for the 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 and Google geocoder depending on house number availability in OSM.

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

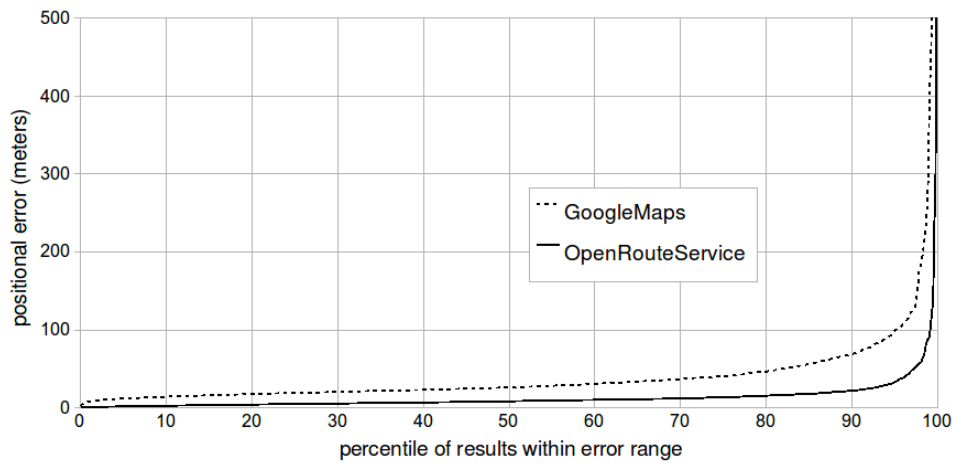


Fig. 2: Positional errors of OpenRouteService and GoogleMaps geocoder when house number data is available in OSM.

It was further found that it is indeed possible to effectively approximate house number locations by two different probability based approaches to guess street directions (see table 3). It was also found that the effectiveness of these approaches, although showing a significant overall average improvement, depends heavily on the study area. It was further found that these improvements are still insufficient to generate accuracy levels comparable to cases where actual house number data is available.

Table 3: Effectiveness of probability based house number approximations

Approach to guess street direction	Sample Size	Improvement of positional accuracy
Dead-end street approach	23656	29%
City center approach	10423	28%
Street rank approach	17242	0%

4 Conclusion and Outlook

The research presented - - can serve as a proof of concept for the use of volunteered spatial data as a reference dataset for geocoding services. It could be shown that it is indeed possible to build a working geocoder based on volunteered geographic information. The inherent inconsistencies presented in the OpenStreetMap data however required substantial concessions in terms of referential integrity. Furthermore, the positional accuracy to be expected strongly depends on the availability of house number data, although means to partially compensate incomplete data have successfully been developed.

The result of this work is already used operationally as the geocoding engine for various research projects, of which OpenRouteService (<http://www.openrouteservice>) and OSM-3D (<http://www.osm-3d.org/>) presumably are the most prominent.

The recent development of the OpenStreetMap project is very promising, too. The amount of house number locations stored in the OpenStreetMap database for the area of Germany has almost doubled during the implementation phase of the research presented. Starting with 172,000 house numbers at the end of December 2008 the amount increased to more than 330,000 house numbers at the end of April 2009. At the time of writing this paper (February 2010) there were around 600,000 house numbers in the database for the area of Germany and about 4.5 million for the scope of Europe.

References

- BAKSHI, R., KNOBLOK, C. A. & THAKKAR, S. (2004), Exploiting online sources to accurately geocode addresses. Proc. 12th annual ACM international workshop on Geographic information systems, pages 194–203, Washington DC, USA.
- BEHR, F.-J., RIMAYANTI, A. & LI, H. (2008), 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.
- BORKAR, V. R., DESHMUKH, K. & SARAWAGI, S. (2001), Automatic segmentation of text into structured records. Proc. SIGMOD Conference, Santa Barbara, California.

- CAYO, M. R. & TALBOT, T. O. (2003), Positional error in automated geocoding of residential addresses. *International Journal of Health Geographics*, 2:10, December 2003.
- CHRISTEN, P. & CHURCHES, T. (2005), A probabilistic deduplication, record linkage and geocoding system. Proc. ARC Health Data Mining workshop, Adelaide, Australia, April 2005.
- DAVIS, C., FONSECA, F. & BORGES, K. (2003), A flexible addressing system for approximate geocoding. In *Brazilian Symposium on GeoInformatics*, 2003.
- DEARWENT, S. M., JACOBS, R. R. & HALBERT, J. B. (2001), Locational uncertainty in georeferencing public health datasets. *Journal of Exposure Analysis and Environmental Epidemiology*, 11:329–334, 2001.
- FARVACQUE-VITKOVIC, C., GODIN, L., LEROUX, H., VERDET, F. & CHAVEZ, R. (2005), Street Addressing and the Management of Cities. The International Bank for Reconstruction and Development / The World Bank, Washington, DC, USA, 2005.
- GOLDBERG, D., WILSON, J., KNOBLOCK, C., RITZ, B. & COCKBURN, M. (2008), An effective and efficient approach for manually improving geocoded data. *International Journal of Health Geographics*, 7(1):60, 2008.
- GOLDBERG, D. (2008), A Geocoding Best Practices Guide. University of Southern California, GIS Research Laboratory.
- GOODCHILD, M. F. (2007a), Citizens as sensors: the world of volunteered geography. *GeoJournal*, 69(4):211–221, 2007.
- GOODCHILD, M. F. (2007b), Citizens as sensors: Web 2.0 and the volunteering of geographic information. *Geofocus*, 7:8–10, 2007.
- GRUBESIC, T. H. & MURRAY, A. T. (2004), Assessing positional uncertainty in geocoded data. In *Proceedings of the 24th Urban Data Management Symposium*, Chioggia, Italy, 2004.
- HAN, J. & KAMBER, M. (2006), *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management. Diane Cerra, San Francisco, USA, 2nd edition.
- HARRIS, K. (1999), *Mapping Crime: Principle and Practice*. Diane Pub Co, 1999.
- KRIEGER, N., WATERMAN, P., LEMIEUX, K., ZIERLER, S. & HOGAN, J. (2001), 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.
- MAZUMDAR, S., RUSHTON, G., SMITH, B., ZIMMERMAN, D. & DONHAM, K. (2008), Geocoding accuracy and the recovery of relationships between environmental exposures and health. *International Journal of Health Geographics*, 7(1):13, 2008.
- NEIS, P. (2008), Location based services mit openstreetmap daten. Master's thesis, Fachhochschule Mainz Fachbereich I.
- NEIS, P. & ZIPF, A. (2008a), Zur kopplung von opensource, opens und openstreetmaps in openrouteservice.org. Proc. AGIT, Salzburg, Austria, 2008.
- NEIS, P. & ZIPF, A. (2008b), Openrouteservice.org is three times open: Combining opensource, opens and openstreetmaps. Proc. GISRUK 2008 conference, Manchester, April 2008. UNIGIS UK.

- OGC (2008), OpenGIS Location Service (OpenLS) implementation specification: Core Services, Sep 2008.
- POULIQUEN, B., STEINBERGER, R. IGNAT, C. & DE GROEVE, T. (2004), Geographical information recognition and visualization in texts written in various languages. In SAC '04: Proc. 2004 ACM symposium on Applied computing, pages 1051–1058, New York, USA, 2004.
- RAHM, E. & DO, H. H. (2000), Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23:2000, 2000.
- RATCLIFFE, J. H. (2001), On the accuracy of tiger-type geocoded address data in relation to cadastral and census areal units. *Geographical Information Science*, 15:473–485, 2001.
- RUSTHON, G., ARMSTRONG, M. P., GITTLER, J., GREENE, B. R., PAVLIK, C. E., WEST, M. M. & ZIMMERMAN, D. L. (2006), Geocoding in cancer research: A review. *American Journal of Preventive Medicine*, 30:16–24, 2006.
- STÄDTETAG NRW (1979), Richtlinien für die Nummerierung von Gebäuden oder bebauten Grundstücken.
- WHITSEL, E. A., ROSE, K. M., WOOD, J. L., HENLEY, A. C. LIAO, D. & HEISS, G. (2004), Accuracy and repeatability of commercial geocoding. *American Journal of Epidemiology*, 160(10):1023–1029, 2004.
- ZANDBERGEN, P. (2007), Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. *BMC Public Health*, 7(1):37, 2007.