Quality Assessment of Volunteered Geographic Information Using Open Web Map Services Within OpenAddresses

Hans-Jörg STARK

The GI_Forum Program Committee accepted this paper as reviewed full paper.

Abstract

Geocoded address data serve as reference datasets for a broad range of applications. OpenAddresses (OA) is a volunteered geographic information (VGI) project integrating address data collected by volunteers into a central database and offering access to this database free of charge. However, the value of the data depends strongly on its quality. The ISO/TC 211 19100 series of standards provide a framework to assure and document the quality of geo-spatial information, acting as a toolset for the assessment and documentation of gathered data. Open Web Mapping Services (OWMS), such as Bing Maps and others are open and freely accessible services that provide maps along with interfaces to customise and use this data infrastructure. This paper examines and outlines how these OWMS can be integrated as reference dataset in the quality management of OA records.

This paper mainly summarises the findings of the master thesis 'Quality assurance of crowdsourced geocoded address-data within OpenAddresses. Concepts and implementation.' (STARK 2010) and introduces further approaches in the quality management applied by volunteers.

1 Introduction

Geocoded address data are of high value (HANCOCK 2010) as reference datasets for a broad range of applications such as delivery services, emergency services, business mapping, etc. However, its value depends heavily on its quality: it must provide quality in terms of positional accuracy, correct spelling and currency. If quality of the reference dataset is poor the resulting geocoding results will implicitly be equally poor (RATCLIFFE 2001, RATCLIFFE 2004, ZANDBERGEN 2007). In the European countries, especially German speaking ones, high quality geodata are available through either public or commercial organisations (AUER & ZIPF 2009) but their cost is high. This situation led to the conception and implementation of the Open Geo-data OpenAddresses (OA) project in 2007 (STARK 2009), the aim of which is to collect geocoded addresses as volunteered geographic information (VGI) in a central database and provide this dataset to all at no charge.

As useful as the integration of volunteers into information collection may be, the quality of the gathered information remains a valid concern (GOODCHILD 2008). According to AGICHTEIN et al. (2008: 183) 'The quality of user-generated content varies drastically from

Car, A., Griesebner, G. & Strobl, J. (Eds.) (2011): Geospatial Crossroads @ GI_Forum '11.

excellent to abuse and spam.' The acceptance of (spatial) data in general by the user community depends heavily on the data's quality. Thus research in the field of quality assurance of VGI is necessary.

The ISO/TC 211 19100 family standards provide a framework to assure and document the quality of geo-spatial information. These standards serve as a framework in conceptualising, assessing and documenting the quality of spatial data. They are used as reference in the conception of quality assurance of OA.

For further investigation users of OA are not only integrated into the data collection process but also into the quality management aspect. A specific user interface along with a wide range of filters allows for the dedicated quality management.

1.1 Approach of Quality Assessment of OpenAddresses

To assess the quality of OA a reference dataset or service must cover the complete area of investigation. Originally, OA was focussed solely on Swiss address data. However, since OA has received more and more international contributions, in addition to being openly the reference resource should also provide international data. Therefore, Open Web Map Services (OWMS) (JAIN 2007) such as Google Maps, Bing Maps and Yahoo! Maps are used as the reference data-set. Hence their suitability for the task of quality assessment for OA is investigated. The challenge in this context is that the dataset to be assessed claims to have higher accuracy than the reference dataset which it is compared to.

Two basic steps are necessary to perform the quality assessment of OA with OWMS: Firstly the three introduced OWMS must themselves be assessed individually. Secondly it must be determined how the results of the OWMS assessment can be used to appraise each address collected in the OA project.

1.2 Volunteered Geographic Information

The general concept of volunteer-contributed geographic information is well documented (FISCHER 2008, FLANAGIN & METZGER 2008, COLEMAN, GEORGIADOU et al. 2009, ELWOOD 2009). The basic concept of VGI takes advantage of modern technical infrastructure such as handheld Global Positioning System (GPS) receivers, the internet, and Web 2.0 applications incorporating asynchronous JavaScript and XML (AJAX) software to provide highly interactive web-based applications. This development has greatly reduced former distinctions between professional and amateur contributions (WALSH 2008).

1.3 Quality Assessment and Characteristics of geocoded Address Data

The term ‘quality’ expresses various unquantifiable characteristics, and no consensus can be found among experts on a single definition. In the context of spatial data, the term fitness for use (JAKOBSSON & TSOULOS 2007) is used quite often. It means that, used in different contexts, the same product may conform to one context's quality requirements but not to another’s. GOODCHILD defines spatial data quality as ' […the] measure of the difference between the data and the reality that they represent, and becomes poorer as the data and the corresponding reality diverge' (GOODCHILD 2006: 13).

OORT (2006) and FISHER et al. (2006) list a number of various aspects that express spatial data quality such as lineage, accuracy, completeness, logical consistency, semantic accuracy, currency, usage, purpose, constraints, variation in quality, meta-quality and resolution. Due to the characteristics of OA as a dynamic project only a few of the above mentioned aspects can be considered in the assessment process. Hence the focus is on accuracy in terms of attribute and spatial accuracy. Attribute correctness mainly consists of completeness of information and correct spelling while spatial accuracy is defined as the deviation or error distance between the true location – in this case the location provided by an OWMS geocoder – and the user entered position.

Figure 1 illustrates how buildings are located along a street in the sample of Gellertstrasse in Basel. Some buildings are close to the street, others are farther away etc. Such characteristics have a direct impact on the quality of street geocoding results. Implicitly the introduced error distances can vary greatly for street-based (linear) geocoding algorithms that are used within OWMS.

Fig. 1: Map excerpt from parcel map of City of Basel
(source: http://www.stadtplan.bs.ch/geoviewer [viewed January 29 2011])

Additionally, the issue of malicious data entry must be addressed. There is a potential within any VGI project that data is intentionally falsified as an act of vandalism. This could mean that address values are incorrect or that addresses are positioned incorrectly. The presented approach evaluates whether and how, with the use of OWMS, such malicious data can be detected or at least indicated in OA.

From the ISO/TC 211 19100 family standards ISO/TC 211:19113 (2001) (Quality principles), ISO/TC 211:19114 (2001) (Quality evaluation procedures) and ISO/TC 211:19138 (2006) (data quality measures) are applied in the quality assessment process.

1.4 Open Web Map Services

All three of these APIs also provide well documented interfaces with comprehensive functionality offering a range of actions to be taken by the client among which is geocoding. Since all three OWMS use both different spatial datasets as reference data and different geocoding algorithms their geocoding results are not equal for a specific address. Figure 2

presents a number of sample addresses in Basel's Gellertstraße (cf. Figure 1), showing clearly the differences of the three OWMS geocoding results. Google Maps provides the best spatial accuracy, with data very close to the true building location (reference address). Bing Maps, for its part, uses an algorithm that arranges the locations of geocoded addresses closely along or even on the street axis while Yahoo! Maps uses an algorithm that applies uniform lateral offsets to its street-geocoded locations, depending on whether the street-number is odd or even.

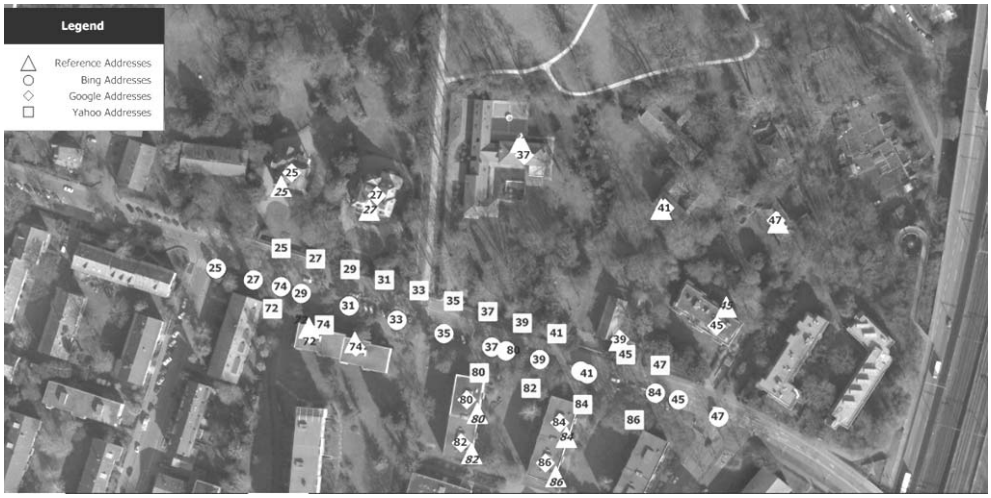

Fig. 2: Map excerpt showing OWMS derived locations versus true locations of addresses at Gellertstraße in Basel

2 Quality Assessment of Open Web Map Services

2.1 Attribute Accuracy

The three OWMS are quality assessed using 93,623 reference addresses of the Canton of Solothurn. Each of these addresses is geocoded by all three OWMS and investigated on its attribute completeness and its error distance. None of the three OWMS geocoders achieved 100% attribute completeness. While Google Maps approaches 97%, the rates of the other two are circa 95%. For all three OWMS datasets, average error distances as positional accuracy are high.

2.2 Spatial Accuracy

For each OWMS geocoder, further constraints are applied to yield the best possible error distance not biased by either bad geocoding quality or bad thematic accuracy. The constraints are mainly: Values of street name & house number & zip code & city name must match, and the Geocoding Quality level must be on the address level.

Table 1: Results of OWMS quality evaluation: Applied quality methods

OWMS	Error distance [m]	Number of Records	Percentage of all Records
Bing Maps	43.5	47,786	51.0%
Google Maps	16.5	54,281	58.0%
Yahoo! Maps	27.2	41,524	44.4%

These constraints lowered the error distances significantly (cf. Table 1).

So far deviations have been analysed only as Euclidean distances, which convey no directional information. Following ZIMMERMAN et al. (2007), differences in x and y error distance directions for each address are analysed. A first visual analysis involves drawing scatterplots. Figures 3 to 5 show scatter plots of deviations split into x- and y-directions.

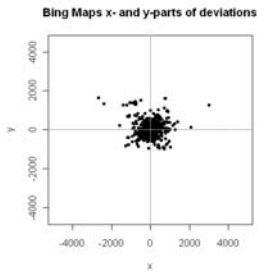

Fig. 3: Scatter plot of deviations for Bing Maps

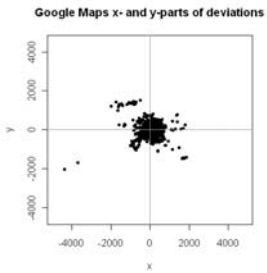

Fig. 4: Scatter plot of deviations for Google Maps

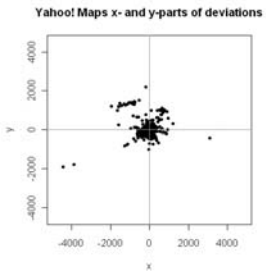

Fig. 5: Scatter plot of deviations for Yahoo! Maps

All three scatter plots show distributions around the origin or intersection of the two axes. They also show a group of outliers that were later detected as erroneous geocoding results.

One clear characteristic common to all three OWMS is that relatively small numbers of outliers increase the range of error distances significantly. ISO/TC 211:19138 (2006, p. 42) suggests the application of a threshold value e_{max} to determine the mean value of positional uncertainties excluding outliers. Deviations are calculated as follows:

$$e'_i = \begin{cases} e_i, if\ e_i \le e_{\max} \\ e_i, if\ e_i \ge e_{\max} \end{cases} \quad (1)$$

with

$$e_i = \sqrt{(x_{mi} - x_{ti})^2 + (y_{mi} - y_{ti})^2} \quad (2)$$

and

$$\bar{e}_{excluding\ outliers} = \frac{1}{N_R}\sum_{i=1}^{N} e'_i \qquad (3)$$

x_{mi} and y_{mi} are coordinates of the OWMS returned location. x_{ti} and y_{ti} represent the coordinates of the true position. N_R is the remaining number of errors. In other words, e_i is the error distance or deviation; e'_i is an accepted deviation if its value is below the outlier threshold; and $\bar{e}_{excluding\ outliers}$ is the average error distance based on all e'_i.

Because the range of deviations can vary greatly, setting a precise definition for e_{max} is difficult. The approach to determining e_{max} has to involve analysing x- and y- components of deviations. To exclude gross errors, only addresses whose x- and y- parts of the deviation are within 95% of the total number of values are considered for further analysis.

The definition e_{max} is derived from the computed values of the 95% Quantile in x- and y-direction for each OWMS. To evaluate reasonable estimators for threshold values for the quality assessment of positional accuracy in OA, the maximum distance of the 95% Quantile in x- and y-directions defines the threshold to determine outliers (cf. Table 2).

Table 2: Threshold values for positional accuracy in meters

	Bing Maps	**Google Maps**	**Yahoo! Maps**
Threshold Quantile 95%	67.08	15.36	42.62
Threshold Outlier	111.76	40.81	68.41

It must be emphasised at this point that these figures apply primarily to Switzerland. In other countries data quality of OWMS may vary and thus threshold values should be assessed accordingly. Hence in OA currently arbitrary threshold values are applied for international use that indicate colour-wise the range of deviations in 20m intervals.

3 Quality Assessment of OpenAddresses

3.1 Approach

Unlike the assessment of OWMS the quality assessment of OA is dynamic, i.e., a new address that is entered or an existing one that is altered shall be assessed immediately. The basic idea is to send the user entered address parameter values to the three OWMS and evaluate the returned OWMS information. If the spelling of the user entered address values match with those of the OWMS returned values it can be assumed that the address was entered correctly. A binary approach is applied for attribute accuracy.

In terms of positional accuracy the user entered position is compared to the OWMS returned positions for the specific address. The computed error distance – user entered position versus OWMS position – is compared to the corresponding threshold values (cf. Table 4) for each OWMS.

3.2 Proof of concept

In order to test whether the OWMS quality assessment was successful and serves as a reference for the quality assessment of OA data a set of test-addresses was used. These test-addresses were classified into three categories: the first category contained addresses with correct locations, the second category contained addresses with small positional, while the third category contained addresses with gross positional errors. For all addresses the address parameter values were entered without errors.

The goal of the test was to evaluate whether a) correct addresses were indicated as correct, b) addresses with gross positional errors (= malicious edits) could be detected, and c) whether this OWMS based approach is able to detect addresses with small positional errors.

3.3 Results

User entered address parameter values are considered correct if at least one of the three OWMS returns a true match for these values. This leads to the result that statements on the correctness of attribute values of addresses are reliable in around 77%. In 23% an additional manual check of the entered values must – erroneously – be conducted. Since this is a Type I error (false positives) it causes only unnecessary effort without compromising quality.

Positional accuracy is more difficult to assess because error distances between true location and OWMS interpolated location can vary a lot. The assessment process uses both of the earlier introduced threshold values: The threshold of Quantile 95% value to check if the user entered address location seems correct and the threshold outlier value to check for gross errors. If only one of the threshold values for each OWMS is considered the results are not satisfactory. Thus a combination of constraints leads to a more robust classification.

The first constraint is used for the classification of a user entered position as correct. It says that for all three OWMS the error distance must be smaller or equal the Quantile 95% threshold value and that for the corresponding address all OWMS must return a geocoding level that indicates address accuracy. This constraint is rather strict and it leads to a quota of error type I (false positives) of 51.2%. This constraint may cause unnecessary additional work but it also eliminates the danger that addresses with gross errors are erroneously classified as correct (error type II).

The second constraint deals with outliers. It says that if for any of the three OWMS derived error distances of an address the value is above the outlier threshold the address is classified as outlier and needs further investigation. This constraint detects outliers well with a rate of 92.7%. Thus chances for an error of type II are minimized to 7.3%. With this constraint chances for an error of type I are 34.3%. Both percentage-values for error types I and II can be considered as acceptable for the remaining risk.

Small positional errors could not be detected with this approach. There must be further research to find alternative approaches to handle addresses with small positional errors.

In order to post-process the entered or altered addresses a web-based user interface was designed that lists the latest addresses along with the values of their quality assessment as explained (cf. Figure 6).

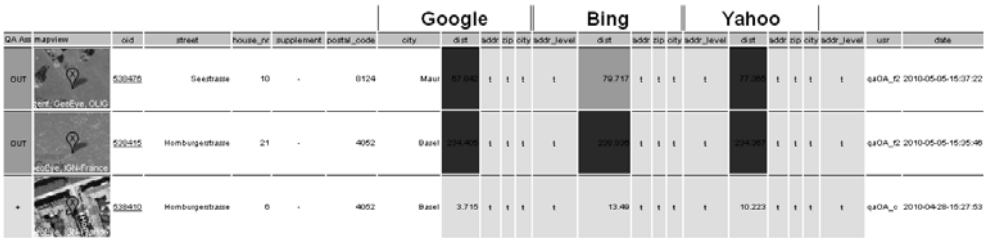

Fig. 6: An overview of OpenAddresses quality assessment

A successful binary comparison of an address parameter value to the OWMS is indicated by a 't'-value while 'f' indicates an unsuccessful comparison result.

The first column indicates the classification according to the presented constraints. Additionally a small static Google Map with a marker that indicates the position of the address helps to visually get an impression of whether the address might be correct or it needs further investigation.

The presented work confirms that a less accurate reference dataset can help in assessing a better dataset in terms of being an indicator especially for gross errors.

4 Implementation of Quality Management in OpenAddresses

Currently the findings of the presented research are implemented in the OpenAddresses project in a slightly modified way. Whenever an address is digitized it is compared to all three OWMS and the comparison results along with the deviation are stored in the database. The originally strict comparison of address parameter values has been slightly relaxed (no distinction between upper and lower case spelling). The user interface has also been adapted slightly omitting the small Google Maps picture and applying a simpler colour code on deviation classification and providing a link for each address that launches OA in a new browser window so that the location along with the address values can be checked online (cf. Fig 7).

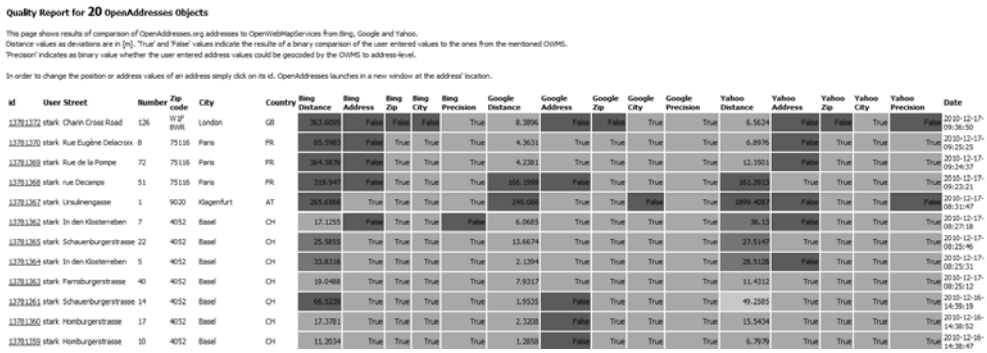

Fig. 7: Current interface of OpenAddresses quality assessment report

The quality assessment report can be customised by several filter variables. It allows for the filtering according to user, date, deviation and address parameter values and the limitation of the number of records in the report.

A further extension is the establishment of regional quality managers. People may apply as regional quality managers providing the geometry of the area in which they want to act as quality managers. This option allows them to customise the quality assessment report by applying a spatial filter – which is the quality manager's area.

Additionally an option of automatic notification exists: if a data donator wants its addresses to be tagged with his contact information he is informed by mail about any change on one of his donated data records. This implies both address deletion and manipulation. If it is a manipulation that improves data quality the contributor will be happy to learn of this improvement. If it is an act of vandalism he may re-establish the address' original state.

5 Conclusion and Outlook

The presented work confirms that a less accurate reference dataset can help in assessing a better dataset. Although small positional errors may not be detected gross errors or malicious edits are identified.

In the future the address parameter value comparison could be changed from a relatively strict binary comparison to a more distinct fuzzy-kind of comparison. Algorithms from the area of text matching may be investigated and applied.

Another challenge to face is the heterogeneous syntax of addresses in an international context. The architecture of an address depends on national standards and is far from being globally homogeneous. This leads to complex comparison algorithms for the quality assessment of user entered data. The future will show whether the work in international context will improve this situation with (ISO/TC 211:19160 2010).

References

AGICHTEIN, E., CASTILLO, C. et al. (2008), Finding high-quality content in social media. Proceedings of the international conference on Web search and web data mining. Palo Alto.

AUER, M. & ZIPF, A. (2009), How do free and Open Geodata and Open Standards fit together? From Sceptisim versus high Potential to real Applications. The First Open Source GIS UK Conference, Nottingham.

COLEMAN, D. J., GEORGIADOU, Y. et al. (2009), Volunteered Geographic Information: The Nature and Motivation of Produsers. International Journal of Spatial Data Infrastructures Research, 4: 332-358.

ELWOOD, S. (2009), Geographic Information Science: new geovisualization technologies – emerging questions and linkages with GIScience research. Progress in Human Geography, 33 (2): 256-263.

FISCHER, F. (2008), Collaborative mapping. How Wikinomics is Manifested in the Geo-Information Economy. GEOInformatics, 11 (2): 28-31.

FISHER, P., COMBER, A. et al. (2006), Approaches to Uncertainty in Spatial Data. In: DEVILLERS, R. & JEANSOULIN, R. (Eds.), Fundamentals of spatial data quality. London: ISTE, pp. 43-59.

FLANAGIN, A. J. & METZGER, M. J. (2008), The credibility of volunteered geographic information. GeoJournal, 72: 137-148.

GOODCHILD, M. F. (2006), Foreword. Fundamentals of spatial data quality. In: DEVILLERS, R. & JEANSOULIN, R. (Eds.), Fundamentals of spatial data quality. London: ISTE, pp. 13-16.

GOODCHILD, M. F. (2008), Citizens as sensors. GIS Trends + Markets, 6: 27-29.

HANCOCK, C. (2010), Address management for emergency services. GEOconnexion International Magazine, 9 (2): 20-21.

ISO/TC 211:19113 (2001), Geographic Information – Quality Principles, International Organization for Standardization (ISO), pp. 1-32.

ISO/TC 211:19114 (2001), Geographic Information - Quality Evaluation Procedures, International Organization for Standardization (ISO), pp. 1-71.

ISO/TC 211:19138 (2006), Text for TS 19138 Geographic Information – Data quality measures, as sent to ISO for publication. http://www.isotc211.org/protdoc/211n2029/ (March 25 2010).

ISO/TC 211:19160 (2010), Draft Review summary of project 19160, Addressing. http://www.isotc211.org/address/docs/Review_summary_19160_20101108.pdf (January 20, 2011).

JAIN, A. (2007), Mechanisms for validation of volunteer data in open web map services. http://www.ncgia.ucsb.edu/projects/vgi/docs/supp_docs/Jain_paper.pdf (March 11 2010).

JAKOBSSON, A. & TSOULOS, L. (2007), The Role of Quality in Spatial Data Infrastructures. 23rd International Cartographic Conference, Moscow.

OORT, P. A. J. V. (2006), Spatial data quality: from description to application. Wageningen: Wageningen Universiteit. PhD, 140 p.

RATCLIFFE, J. H. (2001), On the accuracy of TIGER-type geocoded address data in relation to cadastral and census areal units. Int. J. Geographical Information Science, 15 (5): 473-485.

RATCLIFFE, J. H. (2004), Geocoding crime and a first estimate of a minimum acceptable hit rate. Int. J. Geographical Information Science, 18 (1): 61-72.

STARK, H.-J. (2009), OpenAddresses – Free geocoded street addresses. Applied Geoinformatics for Society and Environment, Stuttgart.

STARK, H.-J. (2010), Quality assurance of crowdsourced geocoded address-data within OpenAddresses. Concepts and implementation. Master thesis. Centre for GeoInformatics (Z_GIS) Salzburg, Salzburg University, MSc, 126 p.

WALSH, J. (2008), The beginning and end of Neogeography. GEOconnexion International Magazine, 7 (4): 28-30.

ZANDBERGEN, P. A. (2007), Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads. BMC Public Health, 7 (37).

ZIMMERMAN, D. L., XIANGMING, F. et al. (2007), Modeling the probability distribution of positional errors incurred by residential address geocoding. International Journal of Health Geographics, 6 (1): 1-16.