# QUALITY ASSESSMENT OF VGI BASED ON OPEN WEB MAP SERVICES AND ISO/TC 211 19100-FAMILY STANDARDS

Hans-Jörg Stark[1]

[1]University of Applied Sciences Northwestern Switzerland, Institute of Geomatics Engineering, Muttenz, Switzerland, hansjoerg.stark@fhnw.ch

## Abstract

*Geocoded address data have high value as reference datasets for a broad range of applications, including express companies, emergency services, business mapping, etc. 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 created from a tremendous amount of spatial data, along with interfaces to customise and use this data infrastructure.*

*Based on ISO/TC 211 standards, OWMS can be used in the quality assessment of OA data, with attribute correctness detectable with a probability of 77% (sample size = 413). Reliable assessment of positional accuracy is more difficult: deviations, such as error distances, can vary greatly. Two constraints are applied for the quality assessment: one regarding deviations, the other OWMS geocoding level information. The first constraint classifies none of the maliciously misreported addresses as correct (n=123). The second constraint correctly identifies 92.7% of addresses with gross positional errors. A web based dynamic interface allows the user to immediately see a chosen address's classification based on the applied constraints, with the option correcting a potentially faulty address immediately.*

*This paper summarises the findings of the master thesis 'Quality assurance of crowdsourced geocoded address-data within OpenAddresses. Concepts and implementation.' (Stark, 2010)*

# 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, 2004; Zandbergen, 2007). In European countries, especially German speaking countries, high quality geodata is available through either public or commercial organisations (Auer and Zipf 2009) but their cost is high. This situation led to the conception and implementation of the Open Geo-data (OGD) 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, with the resulting datasets available 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 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.

## 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 has been described by many authors and is well documented (Fischer, 2008; Flanagin and Metzger, 2008; Coleman 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). Sui (2008: 4) has gone so far as to call VGI 'geography without geographers'.

In the area of *community based VGI,* in which OA is located, the most prominent project is certainly OpenStreetMap (OSM)[1]. But there is also the area of commercially oriented VGI, i.e., enterprises that take advantage of VGI data for commercial gain. A member of this category is People's Map[2].

## 1.3 Geocoded Address Data

In business mapping and geomarketing, high-resolution geocoded address data are often used to analyze spatial distributions, customer densities, etc. Address gazetteers and administrative units also take advantage of these data (Harris et al., 2006).

In health geography and epidemiology micro-geographic analyses are now common (Gatrell and Senior, 2005; Messina et al., 2006). Most importantly, this form of analysis demands not only high spatial accuracy for each application area but also completeness of the reference data (Goldberg, 2008) which is particular aspect of data quality.

## 1.4 Quality Assessment

The term ‘quality’ expresses various unquantifiable characteristics, and no consensus can be found among experts on a single definition. For some people, a high-quality product is one without errors; for others it is one that meets the expectations of a consumer. In the context of spatial data, the term *fitness for use* (Jakobsson and 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).

[1] http://www.openstreetmap.org [viewed July 29 2010]

[2] http://peoplesmap.com [viewed July 29 2010]

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 some of the above mentioned aspects cannot 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 of the true location and - in the case of OWMS assessment - the location provided by the OWMS geocoder or - in the case of OA - 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 error distances can vary greatly for street-based (linear) geocoding algorithms that are used within OWMS.


Figure 1. Map excerpt from parcel map of City of Basel
(source: http://www.stadtplan.bs.ch/geoviewer [viewed July 29 2010])

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. Goodchild (2007) sees a vulnerability of VGI at this point. 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. Further ISO/TC 211:19115 (2002) (Metadata) serves as concept on how to document the results of the quality assessment process.

As mentioned before the three OWMS Bing Maps, Google Maps and Yahoo! Maps must be assessed first in order to serve as reference for the assessment process of OA. A complete dataset of geocoded addresses of the Canton of Solothurn (cadastral data) serves as the reference data for this first quality assessment (OWMS assessment).

**1.5 Open Web Map Services**

All three OWMSs discussed provide application programming interfaces (APIs) both to integrate their maps into web sites and to enable further processing (geo-spatial) information. 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 - the extraction of location information of a particular address as latitude and longitude along with its address parameter values. Since all three OWMS use both different spatial datasets as reference data and different geocoding algorithms their geocoding results are not equal for the same 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. 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.


Figure 2. Map excerpt showing OWMS derived locations versus true locations of addresses in Basel at Gellertstrasse

## 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, stored in a database and investigated on its attribute completeness and its error distance. Table 1 shows the results of the OWMS quality assessment not taking geocoding level into consideration.

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

| Data Quality Subelement | Bing Maps | Google Maps | Yahoo! Maps |
|---|---|---|---|
| Commission (Attribute Completeness) | 88,711 (94.8%) | 90,731 (96.9%) | 89,030 (95.1%) |
| Omission (Attribute Completeness) | 4,911 (5.2%) | 2,891 (3.1%) | 4,592 (4.9%) |
| Positional Accuracy [m] | 2,891.1 | 3,421.4 | 5,498.0 |
| Thematic Accuracy | 49,819 (53.2%) | 54,282 (58.0%) | 41,584 (44.4%) |

None of the three OWMS geocoders achieved 100% attribute completeness. While Google Maps approaches 97%, the rates of the other two are circa 95%. Comparing these figures to those concerning thematic accuracy it becomes clear that completeness does not imply accuracy! For all three OWMS datasets, average error distances as positional accuracy are extremely 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 address level. These constraints lowered the error distances significantly (cf. Table 2).

Table 2. Error distances with applied constraints

| 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% |

A more detailed analysis of the error distances for the three OWMS geocoders is performed to obtain the best possible estimators of threshold values for each OWMS with regard to the OA data quality assessment. To develop an overview of the distribution of the data to be analysed, histograms are created. These show the ranges and distributions of the values (cf. Figures 3 to 5).

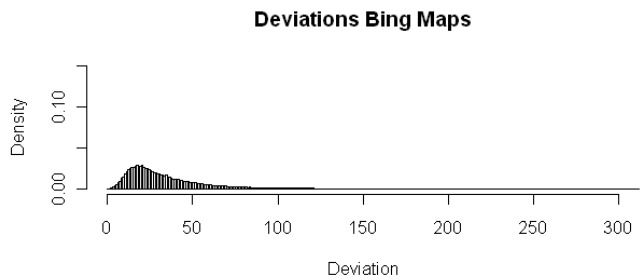



Figure 3. Histogram of distances between Bing Maps' geocoded objects and reference dataset

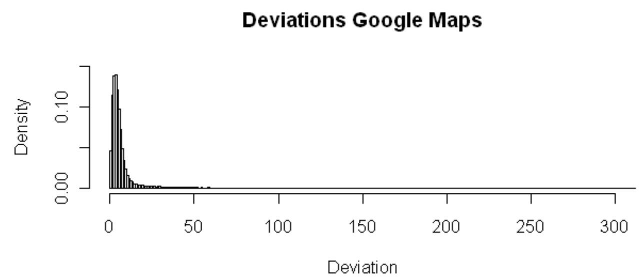



Figure 4. Histogram of distances between Google Maps' geocoded objects and reference dataset

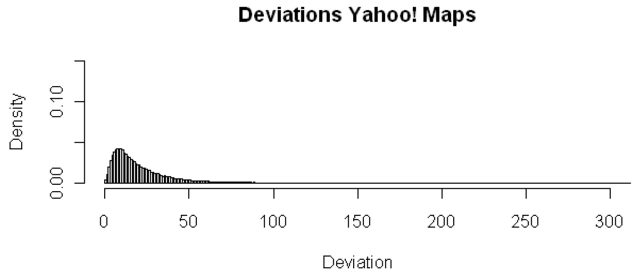



Figure 5. Histogram of distances between Yahoo! Maps' geocoded objects and reference dataset

Google Maps' dataset is spatially closest to the reference data, followed by Yahoo! Maps and finally Bing Maps, which show respectively broader ranges.

So far deviations have been analysed only as Euclidean distances. As these are, by definition, always positive values, deviations convey no directional information, meaning a Gaussian distribution with a mean of 0 is not possible. 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 (cf. Figures 6 to 8).

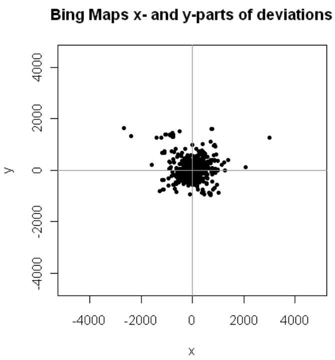



Figure 6. Scatter plot of deviations split into x- and y-directions for Bing Maps

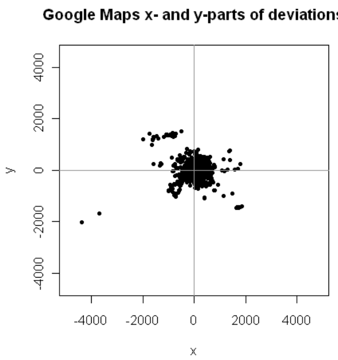



Figure 7. Scatter plot of deviations split into x- and y-directions for Google Maps

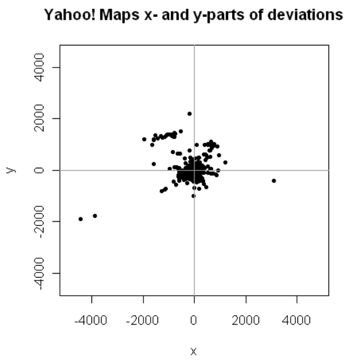



Figure 8. Scatter plot of deviations split into x- and y-directions for Yahoo! Maps

All three scatter plots show distributions around the origin or intersection of the two axes ((0, 0)). Because scatter plots were drawn from a large number of points it is difficult to tell the exact extents of a given percentage of points, or whether the distribution is isotropic. 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. As deviations, the positional uncertainties are calculated as follows:

$$e'_i = \begin{cases} e_i, if\ e_i \leq e_{\max} \\ e_i, if\ e_i \geq e_{\max} \end{cases} \quad [1]$$

with

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

and

$$\bar{e}_{excluding\ outliers} = \frac{1}{N_R} \sum_{i=1}^{N} e'_i \quad [3]$$

$x_{mi}$ 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 (cf. Figures 1 and 2) 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 (cf. Table 3).

Table 3. Determining limits in x- and y-directions in [m] using 95% Quantile

| | x 2.5% | x 97.5% | y 2.5% | y 97.5% | Max deviation[3] | n |
|---|---|---|---|---|---|---|
| Bing Maps | -81.86 | 76.91 | -77.61 | 70.75 | 111.76 | 43.978 |
| Google Maps | -28.95 | 29.78 | -28.65 | 30.37 | 40.81 | 50.603 |
| Yahoo! Maps | -53.00 | 51.38 | -46.49 | 46.47 | 68.41 | 38.187 |

The analysis of the deviations' distribution in x- and y-directions for each OWMS is shown in Figures 9 to 14.

**Bing Maps**

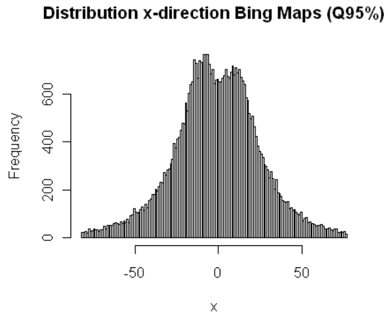



Figure 9. Histogram of x- direction deviations for Bing Maps

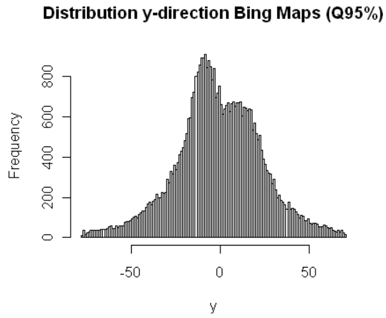



Figure 10. Histogram of y- direction deviations for Bing Maps

**Google Maps**

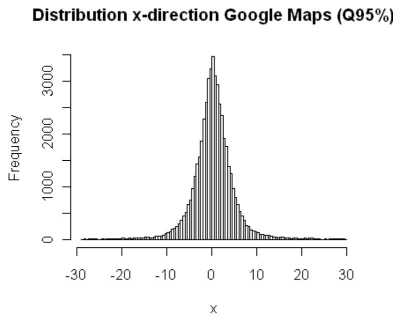



Figure 11. Histogram of x- direction deviations for Google Maps

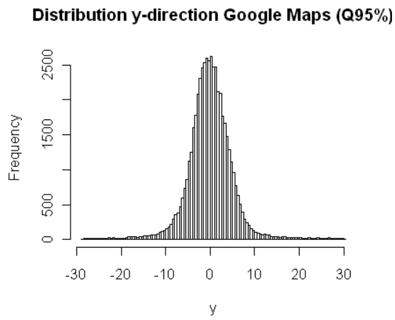



Figure 12. Histogram of y- direction deviations for Google Maps

[3] The values of Max deviations are the empiric ones from the dataset, not the theoretical ones from the analysis

**Yahoo! Maps**

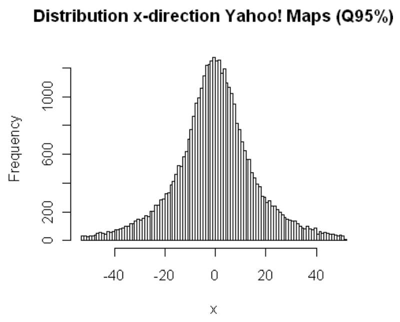



Figure 13. Histogram of x- direction deviations for Yahoo! Maps

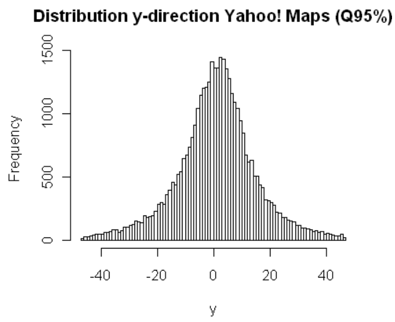



Figure 14. Histogram of y- direction deviations for Yahoo! Maps

Figures 9 to 14 all indicate symmetrical Gaussian distributions in both x- and y-directions, although in Figures 9 and 10, towards the mean, i.e., 0, the frequency decreases. This is due to the geocoding algorithm: as shown in Figure 2 Bing Maps uses an algorithm that aligns the interpolated locations along the street axes with no lateral offset. This means that there is always a deviation between the true location and the computed one: the computed location will never even coincidentally match the true location because all computed locations lie within the street geometry. This fact can be observed in Figures 9 and 10, which show fewer values with a very small deviation in both x- and y-directions. The same effect is very slightly visible in the Histogram of y- direction deviations for Yahoo! Maps (Figure 14).

Redrawing scatterplots similar to Figures 6 to 8 but with the subset described in Table 3 leads to the standard deviation ellipses as presented in Figure 15.

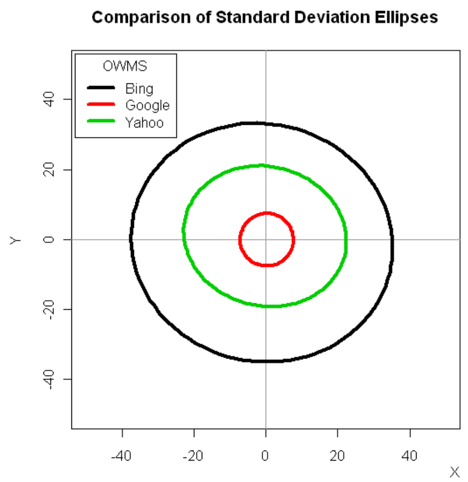



Figure 15. Comparison of Standard Deviation Ellipses

Table 4 gives an overview of the computed parameters of the three OWMS's standard deviation ellipses (cf. Figure 15) based on the x- and y-analysis. The centre x and y values are the coordinates of the ellipse's centre, sigma represents the half-length of each axis in the x- and y-direction, theta is the rotation angle in degrees, and eccentricity describes the flatness of the ellipse.

Table 4. Determining limits in x- and y-directions using 95% Quantile

| | centre x | centre y | sigma x | sigma y | theta | Eccentricity |
|---|---|---|---|---|---|---|
| Bing Maps | -1.14 | -0.86 | 33.9 | 36.6 | 73.1 | 0.38 |
| Google Maps | 0.27 | -0.06 | 7.5 | 7.4 | 167.0 | 0.19 |
| Yahoo! Maps | -0.23 | 0.96 | 19.9 | 22.8 | 74.5 | 0.49 |

The correlation coefficients between x- and y-values are presented in Table 5.

Table 5. Correlation Coefficients of x- and y-values for each OWMS

| | Bing Maps | Google Maps | Yahoo! Maps |
|---|---|---|---|
| Correlation Coefficient between x- and y-values | 0.042 | -0.008 | 0.071 |

The correlation coefficient for Google is nearly 0. Along with Figure 15 it can be assumed that there is no correlation between x- and y-values. The correlation coefficients for Bing Maps and Yahoo! Maps are also very small but show a light correlation. This light correlation may be due to the eccentricity of the standard deviation ellipse. Based on the fact that deviations can vary strongly no further analysis on the correlation of x- and y-values for Bing Maps and Yahoo! Maps are conducted.

The definition $e_{max}$ is derived from the computed values of the 95% Quantile in x- and y-direction for each OWMS. The resulting values are presented in Table 6.

Table 6. $e_{max}$ and concluding $\bar{e}_{excluding\ outliers}$ for each OWMS

| | Bing Maps | Google Maps | Yahoo! Maps |
|---|---|---|---|
| $e_{max}$ | 111.75 | 40.81 | 68.41 |
| $\bar{e}_{excluding\ outliers}$ | 30.46 | 5.53 | 17.64 |

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 7).

Table 7. Threshold values for quality assessment of OA data for positional accuracy

| | 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 the findings of this thesis apply primarily to Switzerland. In other countries data quality of OWMS may vary and thus threshold values should be assessed accordingly.

## 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 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 errors (e.g. the position was defined as slightly outside the building) 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%. This is because of the strict binary comparison algorithm that was applied. Especially when adding characters to house numbers (e.g. '37a') OWMS geocoders do not return identical values and thus user entered input is erroneously classified as potentially wrong. However in only 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 but does not harm the quality of the data.

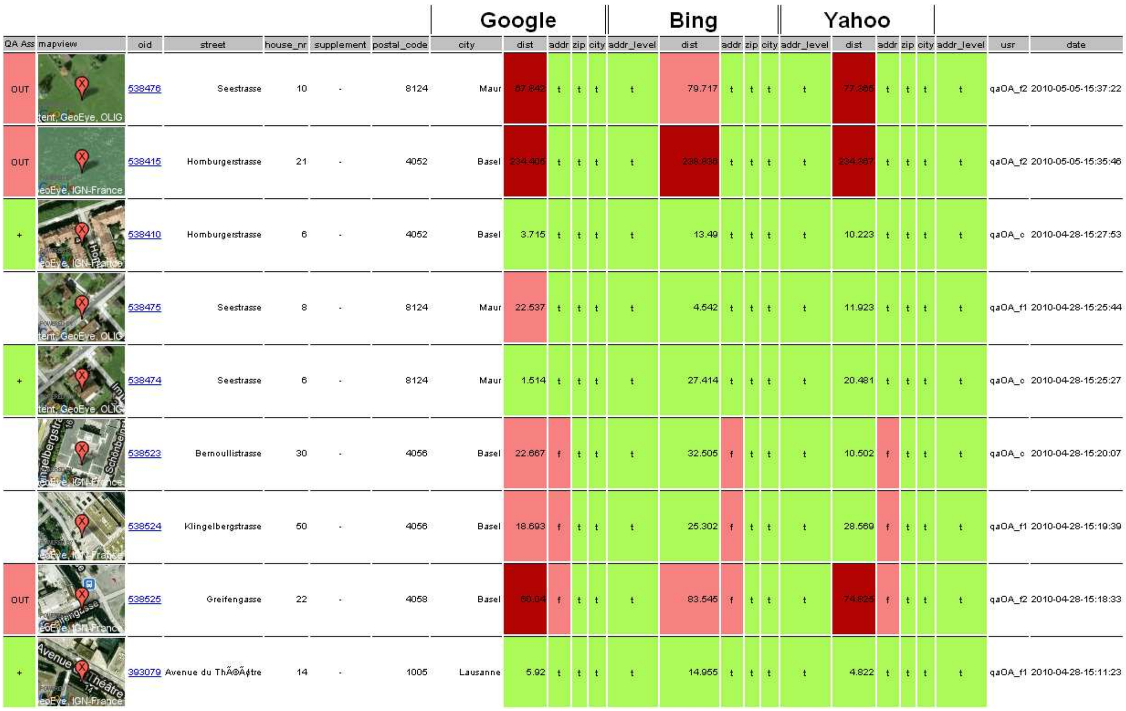
Positional accuracy is more difficult to assess because error distances between true location and OWMS interpolated location can vary greatly. The assessment process uses both introduced threshold values: The threshold of Quantile 95% value to check if the user entered address location seems correct 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 is formulated that 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%. In other words of all user entered addresses a little more than half of them will be investigated on their positional accuracy. 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 (false positives, i.e., an erroneous address is classified as correct) are minimised to 7.3%. With this constraint chances for an error of type I are 34.3%. In other words roughly one third of correctly located addresses are classified as suspicious. 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 is available that lists the latest addresses along with the values of their quality assessment (cf. Figure 16).



| | | | | | | | | Google | | | | | Bing | | | | | Yahoo | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| QA Ass | mapview | oid | street | house_nr | supplement | postal_code | city | dist | addr | zip | city | addr_level | dist | addr | zip | city | addr_level | dist | addr | zip | city | addr_level | usr | date |
| OUT | | 538476 | Seestrasse | 10 | - | 8124 | Maur | 67.842 | t | t | t | t | 79.717 | t | t | t | t | 77.366 | t | t | t | t | qaOA_f2 | 2010-05-05-15:37:22 |
| OUT | | 538415 | Homburgerstrasse | 21 | - | 4052 | Basel | 234.406 | t | t | t | t | 236.836 | t | t | t | t | 234.367 | t | t | t | t | qaOA_f2 | 2010-05-05-15:35:46 |
| + | | 538410 | Homburgerstrasse | 6 | - | 4052 | Basel | 3.715 | t | t | t | t | 13.49 | t | t | t | t | 10.223 | t | t | t | t | qaOA_c | 2010-04-28-15:27:53 |
| | | 538475 | Seestrasse | 8 | - | 8124 | Maur | 22.537 | t | t | t | t | 4.542 | t | t | t | t | 11.923 | t | t | t | t | qaOA_f1 | 2010-04-28-15:25:44 |
| + | | 538474 | Seestrasse | 6 | - | 8124 | Maur | 1.514 | t | t | t | t | 27.414 | t | t | t | t | 20.481 | t | t | t | t | qaOA_c | 2010-04-28-15:25:27 |
| | | 538523 | Bernoullistrasse | 30 | - | 4056 | Basel | 22.667 | f | t | t | t | 32.505 | f | t | t | t | 10.502 | f | t | t | t | qaOA_c | 2010-04-28-15:20:07 |
| | | 538524 | Klingelbergstrasse | 50 | - | 4056 | Basel | 18.693 | f | t | t | t | 25.302 | f | t | t | t | 28.569 | f | t | t | t | qaOA_f1 | 2010-04-28-15:19:39 |
| OUT | | 538525 | Greifengasse | 22 | - | 4058 | Basel | 60.04 | f | t | t | t | 83.545 | f | t | t | t | 74.825 | f | t | t | t | qaOA_f2 | 2010-04-28-15:18:33 |
| + | | 393079 | Avenue du ThÃ©Ã¢tre | 14 | - | 1005 | Lausanne | 5.92 | t | t | t | t | 14.955 | t | t | t | t | 4.822 | t | t | t | t | qaOA_f1 | 2010-04-28-15:11:23 |

Figure 16. Overview of OpenAddresses quality assessment

If the binary comparison of address parameter values was successful a 't'-value is written in the corresponding cell of the table and emphasized with a green colour. If the comparison was not successful a 'f'-value along with red colour indicates a fail. The first column indicates the classification according to the presented constraints. If this testing was not clear - i.e., the address could not be classified as correct nor as outlier - no statement without a colour is given. 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 needs further investigation. If it does need further investigation or even post-editing a click on its oid-value launches the OpenAddresses interface and centers the map at the location of the address and thus makes it easy to alter either attribute values or its position.

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

## 4. Outlook

Since OA is operating globally a concept of "global quality managers" could be evaluated. This means that for certain regions or countries qualified and identified persons act as quality managers. The quality assessment output as in Figure 16 could be adjusted in order that

a) a quality manager sees only the addresses for the region he is responsible for

b) a quality manager is notified via e-mail that new addresses were collected in his area

c) addresses generally require an approval by an authorised quality manager in order to be accepted in the OA database.

Option c) may however contradict to a certain degree the philosophy of VGI and crowdsourcing respectively because it leads to a certain "closed source" community. It will also change OA concept of not requesting an e-mail address or some kind of identification. But in the long run this might be an effective way of applied quality assessment for OpenAddresses in particular and VGI in general.

## References (Times New Roman, 14pt, Bold)


Agichtein, E., C. Castillo, 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. and A. Zipf 2009. 'How do free and Open Geodata and Open Standards fit gogether? From Sceptisim versus high Potential to real Applications.' *The First Open Source GIS UK Conference*, Nottingham.

Coleman, D. J., Y. Georgiadou, et al. 2009. 'Volunteered Geographic Information: The Nature and Motivation of Produsers.' *International Journal of Spatial Data Infrastructures Research* vol 4, no, pp. 332

Elwood, S. 2009. 'Geographic Information Science: new geovisualization technologies - emerging questions and linkages with GIScience research.' *Progress in Human Geography* vol 33, no 2, pp. 256

Fischer, F. 2008. 'Collaborative mapping. How Wikinomics is Manifest in the Geo-Information Economy.' *GEOInformatics* vol 11, no 2, pp. 28

Fisher, P., A. Comber, et al. 2006. 'Approaches to Uncertainty in Spatial Data'. *Fundamentals of spatial data quality*. R. Devillers and R. Jeansoulin. London, ISTE, pp. 43 - 59.

Flanagin, A. J. and M. J. Metzger 2008. 'The credibility of volunteered geographic information.' *GeoJournal* vol no 72, pp. 137

Gatrell, A. C. and M. L. Senior 2005. 'Health and healthcare applications'. *Geographical Information Systems. Principles, Techniques, Management, and Applications*. P. A. Longley, M. F. Goodchild, D. J. Maguire and D. W. Rhind. Hoboken, New Jersey, John Wiley & Sons, pp. 925 - 938.

Goldberg, D. W. 2008. *A Geocoding Best Pracitices Guide*, Springfield, IL: North American Association of Central Cancer Registries.

Goodchild, M. F. 2006. 'Foreword'. *Fundamentals of spatial data quality*. R. Devillers and R. Jeansoulin. London, ISTE, pp. 13 - 16.

Goodchild, M. F. 2007. 'Citizens as sensors: the world of volunteered geography.' *GeoJournal* vol no 69, pp. 211

Goodchild, M. F. 2008. 'Citizens as sensors.' *GIS Trends + Markets* vol no 6, pp. 27

Hancock, C. 2010. 'Address management for emergency services.' *GEOconnexion International Magazine* vol 9, no 2, pp. 20.

Harris, R., P. Sleight, et al. 2006. *Geodemographics, GIS and Neighbourhood Targeting*, John Wiley & Sons, Ltd.

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 211:19115 2002. 'Geographic information/Geomatics - Revised text of 19115 Geographic information - Metadata, as sent to the ISO Central Secretariat for registration as FDIS, International Organization for Standardization (ISO), pp. 1 - 149.

ISO/TC 211:19138. 2006. 'Text for TS 19138 Geographic Information - Data quality measures, as sent to ISO for publication.' viewed March 25 2010, <http://www.isotc211.org/protdoc/211n2029/>.

Jain, A. 2007. 'Mechanisms for validation of volunteer data in open web map services.' viewed March 11 2010, <http://www.ncgia.ucsb.edu/projects/vgi/docs/supp_docs/Jain_paper.pdf>.

Jakobsson, A. and L. Tsoulos 2007. 'The Role of Quality in Spatial Data Infrastructures'. *23rd International Cartographic Conference*, Moscow.

Messina, J. P., A. M. Shortridge, et al. 2006. 'Evaluating Michigan's community hospital access: spatial methods for decision support.' *International Journal fo Health Geographics* vol 5, no 42.

Oort, P. A. J. v. 2006. 'Spatial data quality: from description to application'. viewed March 6 2010.

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* vol 15, no 5, pp. 473

Ratcliffe, J. H. 2004. 'Geocoding crime and a first estimate of a minimum acceptable hit rate.' *Int. J. Geographical Information Science* vol 18, no 1, pp. 61.

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.' Centre for GeoInformatics (Z_GIS) viewed July 29 2010, <http://www.unigis.ac.at/club/bibliothek/pdf/40138.pdf>.

Sui, D. Z. 2008. 'The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS.' *Computers, Environment and Urban Systems* vol 32, no, pp. 1

Walsh, J. 2008. 'The beginning and end of Neogeography.' *GEOconnexion International Magazine* vol 7, no 4, pp. 28

Zandbergen, P. A. 2007. 'Influence of geocoding quality on environmental exposure assessment of children living near high traffic roads.' *BMC Public Health* vol 7, no 37.

Zimmerman, D. L., F. Xiangming, et al. 2007. 'Modeling the probability distribution of positional errors incurred by residential address geocoding.' *International Journal of Health Geographics* vol 6, no 1, pp. 1