

## Master Thesis

submitted within the UNIGIS MSc programme  
at the Centre for GeoInformatics (Z\_GIS)  
Salzburg University

# Quality assurance of crowdsourced geocoded address-data within OpenAddresses Concepts and Implementation

by

**Hans-Jörg Stark**  
up40138

A thesis submitted in partial fulfilment of the requirements of  
the degree of  
Master of Science (Geographical Information Science & Systems) – MSc (GISc)

Advisor:  
Dr. Adrijana Car

Muttenz, June 4, 2010







## **Science Pledge**

By my signature below, I certify that my thesis is entirely the result of my own work, and that I have cited and indicated the origins of all sources used therein.

Muttenz, Switzerland, June 4, 2010

## **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, particularly in terms of positional accuracy. Vgi-based projects face special challenges regarding quality assessment, as no regulatory or legal authority monitors and assesses spatial data quality. Further, vgi projects are inherently dynamic, which implies that quality assessment can only be applied on a subset or feature base level.

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, Google Maps and Yahoo! Maps 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. This thesis investigates whether OWMS can be used to apply ISO/TC 211 19100 series standards to quality-assess OA. As a first step, OWMS geocoding services are assessed based on a reference dataset from the Swiss Canton of Solothurn. Based on the results, thresholds are defined for the assessment of the positional and, to some degree, the thematic accuracy of OA data.

The key finding of this thesis is that, 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, (i.e., differences between user entered positions and true locations) can vary greatly. The main goal of the assessment is to detect inaccurate or

maliciously misreported addresses, thereby allowing the possibility either of correcting inaccurate data or of preventing their inclusion in datasets.

A strategy combining two constraints – one regarding deviations, the other OWMS geocoding level information – is applied. The first constraint classifies none of the maliciously misreported addresses as correct (n=123), while correctly identifying nearly 48.8% of all accurately positioned addresses (n=172). Only 21.2% of addresses with minor positional errors (n=118) are classified as correct, which falls within tolerances. The second constraint correctly identifies 92.7% of addresses with gross positional errors, along with (erroneously) 34.3% of correctly located addresses, as outliers, while 60.2% of addresses with slight positional errors are classified as inaccurate.

A 48.8% rate of correctly classifying accurate addresses seems low. However, using both constraints, the additional effort of rechecking excluded addresses is preferable to the risk of classifying outliers as correct. Hence the results of the assessment can be considered good and the screening strategy both practical and, for the application scenario, robust.

A web based dynamic interface allows the user to immediately see a chosen address's classification based on the applied constraints, with the option of launching OA and correcting a potentially faulty address immediately. Thus, the quality assurance process can readily be integrated into the existing OA application.



## **Acknowledgements**

I would like to thank the UNIGIS team for their support throughout this master's program. I have greatly appreciated their lightning feedback, their generous consultations and their overall friendliness.

I would also like to thank the Canton of Solothurn, who openly provided their spatial data, without which this thesis would not have been possible.

Finally, I thank my family, friends and colleagues for their support, understanding and peer reviews of my ideas and results, especially Chris Shultis for proofing and editing my English writing.

# Contents

<b>SCIENCE PLEDGE</b>	<b>I</b>
<b>ABSTRACT</b>	<b>II</b>
<b>ACKNOWLEDGEMENTS</b>	<b>V</b>
<b>CONTENTS</b>	<b>VI</b>
<b>LIST OF FIGURES</b>	<b>X</b>
<b>LIST OF TABLES</b>	<b>XV</b>
<b>LIST OF ABBREVIATIONS</b>	<b>XVII</b>
<b>LIST OF LISTINGS</b>	<b>XVIII</b>
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Motivation	2
1.2 Task	4
1.3 Approach and Methodology	4
1.4 Expected Results	5
1.5 Intended Audience	6
1.6 Out of Scope	6
1.7 Thesis Structure	7
<b>2 LITERATURE</b>	<b>9</b>

<b>3 RESEARCH BASICS</b>	<b>11</b>
<b>3.1 Basics of geocoded addresses</b>	<b>11</b>
3.1.1 Definition of Geocoding	11
3.1.2 Use of geocoded addresses	12
3.1.3 Structure of geocoded addresses	14
3.1.4 Address Standards	16
<b>3.2 Volunteered Geographic Information</b>	<b>24</b>
3.2.1 VGI, standards and spatial data infrastructures	25
3.2.2 VGI Projects	26
<b>3.3 OpenAddresses</b>	<b>31</b>
3.3.1 Scope and Intention	32
3.3.2 Technological Environment	33
3.3.3 Current State of OpenAddresses	34
<b>3.4 Quality Assessment</b>	<b>34</b>
3.4.1 Definition of 'Quality'	35
3.4.2 Quality Assurance with VGI	37
3.4.3 ISO/TC 211 19100 Series	40
3.4.4 Quality Assurance with OpenAddresses in particular	46
<b>4 ANALYSIS, TECHNOLOGY AND METHODOLOGY</b>	<b>50</b>
<b>4.1 Reference Dataset</b>	<b>50</b>
<b>4.2 Open Web Mapping Services' APIs</b>	<b>52</b>
4.2.1 Functionality	53
4.2.2 Bing Maps	54

4.2.3 Google Maps	56
4.2.4 Yahoo! Maps	57
4.2.5 Comparison of Open Web Map Services	58
<b>4.3 Applied Quality Assurance</b>	<b>62</b>
4.3.1 Assessment of Open Web Map Services' quality	62
4.3.2 Assessment of OpenAddresses' quality	65
<b>5 IMPLEMENTATION OF QUALITY ASSESSMENT</b>	<b>68</b>
<b>5.1 Development Environment</b>	<b>68</b>
<b>5.2 Preparing the Reference Dataset</b>	<b>68</b>
<b>5.3 Implementing Quality Assessment of Open Web Map Services</b>	<b>70</b>
<b>5.4 Implementing Quality Assessment of OpenAddresses</b>	<b>74</b>
<b>6 RESULTS</b>	<b>82</b>
<b>6.1 Evaluation of OWMS geocoders</b>	<b>82</b>
6.1.1 Evaluation according to ISO/TC 211:19113 and ISO/TC 211:19114	82
6.1.2 Defining thresholds for error distances	86
6.1.3 Further Statistical Analysis Based on x- and y-values	91
6.1.4 Conclusion	98
<b>6.2 Evaluation of Quality Assessment for OpenAddresses</b>	<b>99</b>
6.2.1 Null-hypothesis for quality evaluation of OpenAddresses data	99
6.2.2 Test-addresses for Quality Evaluation of OpenAddresses	100
6.2.3 Evaluation according to ISO/TC 211:19113 and ISO/TC 211:19114	101

<b>7 SUMMARY, CONCLUSION AND FUTURE WORK</b>	<b>112</b>
7.1 Summary and Conclusion	112
7.2 Future Work	113
<b>A APPENDIX</b>	<b>117</b>
A.1 Figures	117
A.2 Listings of SQL statements on the evaluation of OWMS geocoding	119
A.3 R Code Listings	121
<b>B LITERATURE (BIBLIOGRAPHY)</b>	<b>128</b>

## List of Figures

Fig. 1 Central role of a geocoding engine .....	12
Fig. 2 Health geography map showing potential community hospital service status (source: Messina et al. (2006), <a href="http://www.ij-healthgeographics.com/content/5/1/42/figure/F12">http://www.ij-healthgeographics.com/content/5/1/42/figure/F12</a> [online March 22, 2010])..	13
Fig. 3 Sample address in OpenAddresses database including spatial information as lon and lat values.....	15
Fig. 4 Hierarchical structure of Swiss addresses.....	15
Fig. 5 Comparison of address location definitions .....	17
Fig. 6 Swiss Data model in UML (source: Schweizerische Normen-Vereinigung (2004, p. 11)) .....	18
Fig. 7 Simplified view of the Swiss address model in UML (source: Schweizerische Normen-Vereinigung (2004, p. 16)).....	19
Fig. 8 General address structure and naming along a street (source: swisstopo (2005, p. 9)).....	20
Fig. 9 General address structure and naming around places/squares (source: swisstopo (2005, p. 9)).....	20
Fig. 10 General address structure and naming within named areas (source: swisstopo (2005, p. 10)).....	21
Fig. 11 General address structure and naming in hamlets (source: swisstopo (2005, p. 24)).....	21
Fig. 12 INSPIRE address data model in UML (source: INSPIRE (2009, p. 16)) ..	22
Fig. 13 Data model used by the Intiendo address matching tool in South Africa based on ISO/TC 211:19112 (2003) (source: Al Rahed et al. (2008, p. 209)). .....	23

Fig. 14 Map view of OpenStreetMap in a web browser (source: <a href="http://www.openstreetmap.org">www.openstreetmap.org</a> [online March 18, 2010]).....	26
Fig. 15 OSM database statistics (source: <a href="http://wiki.openstreetmap.org/w/images/9/91/Osmdbstats1.png">http://wiki.openstreetmap.org/w/images/9/91/Osmdbstats1.png</a> [online March 18, 2010]).....	27
Fig. 16 The five steps of contributing data to the OSM project database (source: <a href="http://wiki.openstreetmap.org/wiki/Beginners%27_Guide">http://wiki.openstreetmap.org/wiki/Beginners%27_Guide</a> [online March 18, 2010]).....	28
Fig. 17 Data collection in OpenAddresses map view container (source: screen capture from <a href="http://www.openaddresses.ch">www.openaddresses.ch</a> [online April 14, 2010]) .....	32
Fig. 18 Data schema of OpenAddresses.....	33
Fig. 19 Tabular overview of data quality elements and data quality subelements with definitions (source: ISO/TC 211:19138 (2006; p. 3)).....	41
Fig. 20 Process to evaluate and report data quality results (source: ISO/TC 211:19114 (2001; p. 3)) .....	42
Fig. 21 UML diagram of metadata application (source: ISO/TC 211:19115 (2002; p. 9)) .....	43
Fig. 22 UML diagram of quality information (source: ISO/TC 211:19115 (2002; p. 22)).....	44
Fig. 23 UML diagram on data quality classes and subclasses (source: ISO/TC 211:19115 (2002; p. 24)) .....	44
Fig. 24 Overview of data quality information of ISO/TC 211:19113 (normative) (source: ISO/TC 211:19113 (2001; p. 5)).....	45
Fig. 25 Map excerpt from parcel map of City of Basel (source: <a href="http://www.stadtplan.bs.ch/geoviewer">http://www.stadtplan.bs.ch/geoviewer</a> [online April 27, 2010]) .....	48

Fig. 26 Metadata of 'Gebäudeadressen' (source: <a href="http://www.sogis1.so.ch/sogis/OnLineData/php/datenbeschreibung.php?id=400454">http://www.sogis1.so.ch/sogis/OnLineData/php/datenbeschreibung.php?id=400454</a> [online April 1, 2010]).....	50
Fig. 27 ERD of topic 'Gebäudeadressen' (source: <a href="http://www.interlis.ch/mo2/diagramme.php?img=Bat&amp;language=d&amp;topic=Gebaueadressen">http://www.interlis.ch/mo2/diagramme.php?img=Bat&amp;language=d&amp;topic=Gebaueadressen</a> [online April 1, 2010]) .....	51
Fig. 28 ERD of 'PLZOrtschaft' topic (source: <a href="http://www.interlis.ch/mo2/diagramme.php?img=Npal&amp;language=d&amp;topic=PLZOrtschaft">http://www.interlis.ch/mo2/diagramme.php?img=Npal&amp;language=d&amp;topic=PLZOrtschaft</a> [online April 1, 2010]) .....	52
Fig. 29 Using Bing Maps API to geocode an address-string .....	55
Fig. 30 Result of the geocoding.....	55
Fig. 31 Using the Google Maps API to geocode an address string.....	57
Fig. 32 Displaying coordinates after geocoding.....	57
Fig. 33 XML listing of the Yahoo! Maps API geocoder .....	58
Fig. 34 Map excerpt showing OWMS derived locations of sample address in MuttENZ.....	59
Fig. 35 Map excerpt showing OWMS derived locations versus true locations of addresses in Basel at Gellertstrasse .....	60
Fig. 36 Creating address data structures in FME .....	69
Fig. 37 Map view of Reference data of the Canton of Solothurn.....	70
Fig. 38 Table with results of OWMS geocoding.....	71
Fig. 39 Flow chart of batch geocoding process for OWMS quality evaluation .....	72
Fig. 40 Data flow of OWMS geocoding quality assessment.....	73
Fig. 41 Flow chart of storage of information for each address after comparison with OWMS .....	77

Fig. 42 Quality evaluation of new or altered address records .....	79
Fig. 43 Missing static map views due to excessive numbers of map requests	81
Fig. 44 Comparison of deviations of all OWMS .....	86
Fig. 45 Detailed view of deviations of all OWMS .....	86
Fig. 46 Histogram of deviations for Bing Maps.....	87
Fig. 47 Histogram of deviations for Google Maps .....	87
Fig. 48 Histogram of deviations for Yahoo! Maps.....	87
Fig. 49 Histogram of distances between Bing Maps' geocoded objects and reference dataset.....	88
Fig. 50 Histogram of distances between Google Maps' geocoded objects and reference dataset.....	88
Fig. 51 Histogram of distances between Yahoo! Maps' geocoded objects and reference dataset.....	88
Fig. 52 Boxplot of deviations for Bing Maps.....	89
Fig. 53 Boxplot of deviations for Google Maps .....	89
Fig. 54 Boxplot of deviations for Yahoo! Maps.....	89
Fig. 55 Zoomed boxplot of deviations for Bing Maps.....	90
Fig. 56 Zoomed boxplot of deviations for Google Maps .....	90
Fig. 57 Zoomed boxplot of deviations for Yahoo! Maps.....	90
Fig. 58 Comparison of wrong location of sample address in Google Maps and correct location according to <a href="http://www.sogis1.so.ch">http://www.sogis1.so.ch</a> .....	90
Fig. 59 Scatter plot of deviations split into x- and y-directions for Bing Maps .....	91

Fig. 60 Scatter plot of deviations split into x- and y-directions for Google Maps.....	91
Fig. 61 Scatter plot of deviations split into x- and y-direction for Yahoo! Maps .....	91
Fig. 62 Histogram of x- direction deviations for Bing Maps.....	94
Fig. 63 Histogram of y- direction deviations for Bing Maps.....	94
Fig. 64 Histogram of x- direction deviations for Google Maps .....	94
Fig. 65 Histogram of y- direction deviations for Google Maps .....	94
Fig. 66 Histogram of x- direction deviations for Yahoo! Maps.....	94
Fig. 67 Histogram of y- direction deviations for Yahoo! Maps.....	94
Fig. 68 Scatterplot of deviations in x- and y-directions for Bing Maps with standard deviation ellipse, Q95%.....	95
Fig. 69 Scatterplot of deviations in x- and y-directions for Google Maps with standard deviation ellipse, Q95%.....	95
Fig. 70 Scatterplot of deviations in x- and y-directions for Yahoo! Maps with standard deviation ellipse, Q95%.....	95
Fig. 71 Comparison of Standard Deviation Ellipses.....	96
Fig. 72 Spatial distribution of test-addresses.....	101
Fig. 73 Adapted quality assessment for OA .....	111
Fig. 74 Visual comparison of OWMS geocoding results in Laupersdorf produced with QGIS .....	117
Fig. 75 Visual comparison of OWMS geocoding results in Olten produced with QGIS .....	118

## List of Tables

Table 1 Swiss structure of geocoded addresses .....	14
Table 2 Comparison of OWMS's geocoding results with one sample address .....	58
Table 3 Listing of OWMS' map data providers.....	59
Table 4 Distance comparison of OWMS geocoder and reference locations for test addresses.....	61
Table 5 Overview of data quality elements and subelements for quality evaluation of OWMS according to ISO/TC 211:19113 (2001; p. 31).....	63
Table 6 Quality evaluation of OWMS: Quality measures and methods for relevant data quality subelements.....	65
Table 7 Data quality elements and subelements for quality evaluation of OA .....	66
Table 8 Quality evaluation of OA: Quality measures and methods for relevant data quality subelements .....	66
Table 9 Overview and description of required files for OWMS batch geocoding process.....	74
Table 10 Query parameters for quality assessment report.....	79
Table 11 Overview and description of required files for OWMS batch geocoding process.....	81
Table 12 Results of OWMS quality evaluation: Applied quality methods .....	82
Table 13 Distinct analysis on thematic accuracy .....	83
Table 14 Geocoding quality of Bing Maps with sample data from the Canton of Solothurn.....	83
Table 15 Geocoding quality of Google Maps with sample data from Canton Solothurn.....	84

Table 16 Geocoding quality of Yahoo! Maps with sample data from Canton Solothurn.....	84
Table 17 Restriction of OWMS to determine RMSE of positional accuracy ...	85
Table 18 RMSE of positional accuracy with applied constraints .....	85
Table 19 Statistical analysis in R of positional accuracy with applied constraints (cf. Table 17).....	87
Table 20 Overview of boxplots statistics of distance evaluation .....	89
Table 21 Determining limits in x- and y-directions using 95% Quantile.....	93
Table 22 Statistics of Standard Deviation Ellipses based on x- and y-clusters	96
Table 23 Correlation Coefficients of x- and y-values for each OWMS .....	97
Table 24 $e_{\max}$ and concluding RMSE / $\bar{e}_{\text{excluding outliers}}$ for each OWMS.....	98
Table 25 Threshold values for quality assessment of OA data for positional accuracy .....	99
Table 26 Number of test-addresses for quality evaluation of OA .....	101
Table 27 Evaluation of non-quantitative attribute correctness for Bing Maps .....	102
Table 28 Evaluation of non-quantitative attribute correctness for Bing Maps .....	102
Table 29 Evaluation of non-quantitative attribute correctness for Bing Maps .....	103
Table 30 Comparison of returned address-values for an address .....	104
Table 31 Evaluation test-addresses of class c on positional accuracy .....	105
Table 32 Evaluation test-addresses of class c on positional accuracy .....	107
Table 33 Evaluation test-addresses of class f2 on positional accuracy .....	108
Table 34 Applying quality assessment constraints on test-classes .....	109

## List of Abbreviations

AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
ERD	Entity Relationship Diagram
GIS	Geographic Information System
GPS	Global Positioning System
HTTP	Hypertext Transfer Protocol
ISO	Organisation for Standardisation
OA	OpenAddresses
OGD	Open Geo-Data
OWMS	Open Web Map Services
OSGEO	Open Source Geospatial Foundation
OSM	OpenStreetMap
PPGIS	Public Participatory Geographic Information System
SDI	Spatial Data Infrastructure
SQL	Structured Query Language
RMSE	Root Mean Square Error
UGC	User-Generated Content
UML	Unified Modeling Language
URL	Uniform Resource Locator
VGI	Volunteered Geographic Information
XML	Extensible Markup Language

## List of Listings

Listing 1 HTML and JavaScript code for geocoding an address with Microsoft Bing Maps .....	55
Listing 2 HTML and JavaScript code for geocoding an address with Google Maps.....	57
Listing 3 Uniform Resource Locator (URL) for geocoding an address with Yahoo! Maps .....	57
Listing 4 SQL Syntax to create the table that contains the values of the OWMS comparison.....	76
Listing 5 Sample call for quality assessment report.....	80
Listing 6 SQL statement for quality report.....	80
Listing 7 SQL statements applied in chapter 6.1 .....	120
Listing 8 R commands applied in chapter 6.1.....	127





## 1 Introduction

*"We are talking about something dramatically different. The new promise of collaboration is that with peer production we will harness human skill, ingenuity, and intelligence more efficiently and effectively than anything we have witnessed previously."*

(Tapscott and Williams 2008, p. 18)

*"Customers have been empowered in new ways. More to the point, they're the people creating the content."*

(Tapscott and Williams 2008, p. 43)

*"But crowdsourcing is no free lunch: communities can be difficult to build and even harder to maintain." (Howe 2008, p. 181)*

*"By motivating individuals to act voluntarily, it is far cheaper than any alternative, and its products are almost invariably available to all."*

(Goodchild 2007, p. 220)

*"What distinguishes the work of the professional from the amateur is its Quality." (Walsh 2008; p. 29)*

*"But missing at this point are the mechanisms needed to ensure quality, to detect and remove errors, and to build the same level of trust and assurance that national mapping agencies have traditionally enjoyed."*

(Goodchild 2008a, p. 30)

*"Thus the task of compiling independently contributed pieces of the patchwork necessarily imposes some degree of quality control."*

(Goodchild 2008b, p. 241)

### **1.1 Motivation**

According to swisstopo (2005), Schweizerische Normen-Vereinigung (2004) and Hancock (2010) geocoded address data are of high value 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, Amelunxen 2009).

In European countries, high quality geodata is available through either public or commercial organisations (Auer and Zipf 2009). In German speaking countries geocoded addresses have been available for some time (TeleAtlas 2008), but their cost has been high. This situation led to the conception and implementation of the Open Geo-data (OGD) OpenAddresses (OA) project in 2007 (Stark 2008, 2009), the aim of which is to collect geocoded addresses via volunteers in a central database, with the resulting datasets available to all at no charge. Scientists and experts have given this approach a variety of names, including, among others, *crowdsourcing* (Howe 2008), *volunteered geographic information* (VGI) (Goodchild 2007), *collaborative mapping* (Fischer 2008) and *participatory mapping* (Aditya 2008). A generic term might be *user-generated content* (UGC), although Howe (2008) distinguishes between UGC and crowdsourcing, seeing crowdsourcing as one type of UGC. After OA, the best-known project of this kind is OpenStreetMap (OSM). A number of others are described below (Section 3.2.2, VGI Projects).

As useful as the integration of volunteers into information collection may be, the quality of the gathered information remains a valid concern (Goodchild 2008a), especially considering the tremendous importance of the final product. Cooper (2009) claims that Geographic Information Systems (GIS) and spatial information are essential ingredients for economic growth and innovation. Acil (2008) confirms this declaration, pointing out that geographic information is being used in increasingly diverse sectors of the

economy, with immediate impacts on productivity. He even considers data quality one of five core issues to be addressed in the development of an Australian Spatial Data Infrastructure (SDI). If this is the case with cadastral and "official" spatial data, how much more is it an issue with VGI! Agichtein et al. (2008, p. 183) explicitly say "The quality of user-generated content varies drastically from excellent to abuse and spam." Even in projects such as OpenStreetMap (OSM), whose contributors are generally motivated and skilled (Stark 2010), quality remains heterogeneous (Haklay 2008, Amelunxen 2009, Auer and Zipf 2009).

The ISO/TC 211<sup>1</sup> 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.

Based on the issues discussed thus far, the motivations for this thesis can be summarised as follows:

- Free and freely available spatial data is a driver for economic growth and innovation.
- The concept of volunteers collecting spatial data and contributing their work to an open and centrally located pool has been approved by many projects in recent years and has considerable potential for future development.
- Acceptance by the user community - be it commercial, public or private - is a key factor influencing data use. This acceptance depends heavily on data quality.
- ISO/TC 211 standards are available on quality assessment and documentation of geo-spatial data.

---

<sup>1</sup> Organisation for Standardisation, Technical Committee 211 Geographic Information (<http://www.isotc211.org/> [online May 12, 2010])

Consequently, this thesis will examine how the ISO/TC 211 19100 family of standards can be used and applied to the OA project.

A reference dataset or service used to assess the quality of OA data 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 available – and preferably be free of charge – the reference resource should also provide international data. Therefore, Open Web Map Services (OWMS) (Jain 2007) are used as the reference data-set here, while the thesis investigates their suitability for the task of quality assessment for OA.

### **1.2 Task**

The main task performed in this thesis is the evaluation of how and to what degree the three Open Web Mapping Services (OWMS) – Google Maps, Bing Maps and Yahoo! Maps – can be used for quality assurance regarding crowdsourced data in the OpenAddresses (OA) project. The main focus will be on positional accuracy but also - if applicable - on attribute correctness. These quality aspects are evaluated and assessed according to the ISO/TC 211 19100 family of standards. However, a second challenge inherent in this research question is whether a VGI dataset claiming high spatial (address-level) accuracy can be quality assessed using services and data of equal or lower quality.

Two basic steps will therefore be necessary to achieve this thesis's primary goal. Firstly the three introduced OWMS will themselves be assessed individually. Secondly it will be determined how the results of the OWMS assessment can be used to appraise each address collected in the OA project.

### **1.3 Approach and Methodology**

Al Rahed et al. (2008) presented the application of ISO 19112 standards to model geocoded addresses in South Africa. Based on their work and the information and concepts of ISO/TC 211:19113 (2001) and Jakobsson and Giversen (2007) the three OWMS introduced above will be quality assessed in a first step. The main goal of this process is to determine these services'

positional accuracy, along with their similarity of address information and spelling. This assessment will be conducted according to ISO/TC 211:19113 (2001), ISO/TC 211:19114 (2001) and ISO/TC 211:19138 (2006), using the official cadastral address data of the Canton of Solothurn – including more than 93'000 addresses – as reference data. Thus it is expected that a precise statement of the positional accuracy will be achieved for each service, along with information on the geocoding quality and correctness of address information. The result will be a global statement on the positional and thematic accuracy for each OWMS regarding the area of interest (the Canton of Solothurn).

In the second step, the selected OWMSs will be used as reference standards to assess addresses collected in OA. Again, ISO/TC 211:19113 (2001), ISO/TC 211:19114 (2001) and ISO/TC 211:19138 (2006) will be used to guide the quality assessment process. However, unlike the first quality assessment of the Mapping application programming interfaces (APIs), which will lead to a global statement, this second step will yield quality information for every address collected in the OA project. It will also comprise the implementation of a dynamic website to inform users about the quality assessment of the collected OA addresses. This approach gives volunteers the option to act either as quality managers in a quality assessment system such as that of Wikipedia (Baumann 2008) or as peer-reviewers (Flanagin and Metzger 2008).

### ***1.4 Expected Results***

The key question to be answered is: Can OWMSs such as Bing Maps, Google Maps and Yahoo! Maps serve as tools to assess VGI in the OA project?

It is also elaborated whether and how the ISO/TC 211 19100 family's quality aspects can be applied to both OWMS and crowdsourced geographic information.

Further questions to be answered are:

Based on the OWMS assessment, can thresholds be defined to serve as indicators in the OA quality assessment process?

## Quality Assurance of OpenAddresses

Is it possible to provide an easy to use and easy to understand interface for volunteers who want to assist in the quality assurance of OA data?

The expected results are both a concept and an application to enable the detection of user entered addresses with low attribute and positional quality. Additionally, good quality addresses will be approved based on comparisons with OWMS data as reference.

### ***1.5 Intended Audience***

The intended audience of this thesis is anyone who has an interest in the following topics:

- Volunteered Geographic Information (VGI)
- Geocoded address data
- Quality aspects of geographic information in general and of OpenAddresses data in particular
- Application of the ISO/TC 211 19100 series of quality standards
- Implementation of quality assurance mechanisms in the area of crowdsourced geographic information

### ***1.6 Out of Scope***

The thesis will not explicitly cover questions of geocoding such as mechanisms, algorithms or resulting geocoding accuracy depending on different geocoding approaches. Many authors have worked on research questions relating to geocoding (Ratcliffe 2001, 2004, North American Association of central cancer registries 2002, Goodchild 2007, Zandbergen 2007, Zimmerman et al. 2007, Al Rahed et al. 2008, Goldberg 2008, 2008a, Harvard University 2008, Amelunxen 2009, Zandbergen and Hart 2009), especially with street geocoding. Researchers generally agree on the importance of geocoded address data as reference data for geocoding purposes; still, it is beyond the scope of this thesis to investigate the entire process of geocoding.

Neither will the thesis discuss the mechanisms or prerequisites to running a successful Open Geo-Data (OGD) project. Finally, data modelling questions or relations to existing spatial data infrastructures (SDI) are also out of the scope of this thesis.

### **1.7 Thesis Structure**

The thesis is structured as follows:

Literature	The literature section gives a high level overview of the main themes and the relevant corresponding sources.
Research Basics	This chapter provides basic information on the essential concepts, projects and technologies applied in this thesis. It shows the current state of research on the main themes introduced in the literature chapter.
Analysis and Methodology	In order to assess OWMS for quality assurance a, high-quality reference dataset is necessary. In this chapter, the reference data of the Canton of Solothurn to be applied, OWMS interfaces and the application of the ISO/TC 211 19100 family of standards are introduced.
Implementation	<p>The assessment of OWMS is performed. This includes development of the batch geocoding environment and setup of the database to store results for further processing.</p> <p>Additionally, the quality assessment of OA for each address is conceptualised and implemented with dummy thresholds until the true values, based on the OWMS assessment analysis, are found.</p>

## Quality Assurance of OpenAddresses

### Results

The statistical analysis of the OWMS assessment is presented, describing findings for the values of the defined quality measures according to the ISO/TC 211 19100 family of standards.

According to the findings of the OWMS analysis, threshold values are applied to the OA data quality assessment framework and tested with sample addresses.

### Summary

The final part summarises the results and findings of the thesis.

### Literature

This section provides the entire list of literature consulted.

### Appendix

The appendix contains additional figures and code listings.

## 2 Literature

Goodchild (2007) introduced the paradigm of volunteered geographic information (VGI). Because the OA project is based on this concept, VGI is one of the main themes of this thesis. As mentioned above in Section 1.1 (Motivation) numerous other terms refer to what Goodchild described; but all refer to more or less the same practice: volunteers - be they professionals or non-experts - contribute geographic information into a centrally stored data repository that is open and freely accessible to all. In addition to Goodchild, Sarah Elwood has conducted research on the phenomenon of allowing non-experts to participate in geographic information projects. This was formerly referred to as Public Participatory GIS (PPGIS) (Elwood 2007, Walsh 2008).

Address basics related to standards and structure - how to model an address and an address database - are mostly country dependent. Since the assessment of OWMS is conducted with Swiss data, the Swiss Association for Standardization (Schweizerische Normen-Vereinigung) (2004) and swisstopo (2005) are major sources for this assessment. Internationally, Coetzee and Cooper (2007) and Coetzee et al. (2008) also provide valuable information on structural investigation of geocoded address data. Their overview of diverse international address repositories and structures helps clarify the global context regarding address data structure and its potential implications vis à vis quality assessment of this data.

When quality is to be assessed, the term 'quality' must first be defined. Various authors have worked on different definitions of the term (Chrisman 2006, Fisher et al. 2006, Devillers and Jeansoulin 2006), greatly simplifying the task of defining quality aspects appropriate for OA. Valuable overviews and introductions into quality components, standards, and the documentation of these aspects in the context of spatial data are offered in Hanguët (2006) and Servigne et al. (2006). These works are particularly significant to the assessment of OA data with regard to the concept of *data quality* presented here.

## Quality Assurance of OpenAddresses

In terms of the *quality assurance* of spatial data captured through the VGI paradigm a small number of articles or studies have been published (Haklay 2008, Maué and Schade 2008, Amelunxen 2009), and Alexander Zipf and his research team are also active in this area (Auer and Zipf 2009, Zipf 2009). However, almost all related investigations refer to OSM, focussing on linear features and not specifically on addresses or other point objects. They mainly give quality information in terms of completeness and spatial accuracy. While such papers are valuable for this thesis from a conceptual point of view, no publication has yet been found that explicitly applied ISO standards to quality assessment of VGI based data.

Quality assessment in general is best derived directly from the application of standards. For geospatial information, the following ISO TC211 standards are most relevant: ISO/TC 211:19113 (2001), ISO/TC 211:19114 (2001), and ISO/TC 211:19138 (2006). Jakobsson and Giversen (2007) describe the process of quality assessment in the context of national mapping and cadastral agencies, which comprise a helpful sample for this thesis. INSPIRE (2009) also refers to these ISO documents, thereby emphasizing their relevance in the European context, which can be significant for the future of OA.

Although geocoding is not the main issue discussed in this thesis, it is relevant because, as reference data, addresses play a key role in the geocoding process. Considerable research has been done in this area. Principally, Goldberg (2008) has created an excellent compendium and covered the issue in detail, while Ratcliffe (2001, 2004), Zandbergen (2007), Zimmerman et al. (2007), Al Rahed et al. (2008), and Zandbergen and Hart (2009) provide further relevant information on the role and importance of address data as geocoding reference data.

Lastly, but no less significantly, OWMS APIs are used in this thesis to assess OA data quality, meaning their application must be studied and applied to the context of this thesis. Fortunately, a wealth of literature can be employed to learn the use of these APIs (Erle et al. 2005, Brown 2006, Gibson and Erle 2006, Purvis et al. 2006).

## 3 Research Basics

### 3.1 Basics of geocoded addresses

The following sections introduce the basic principles of geocoded addresses as they pertain to this thesis.

#### 3.1.1 Definition of Geocoding

Although geocoding is not directly investigated in this thesis, it is nevertheless relevant as the main "interface" or gateway to the application and further analysis of address data.

Geocoding is well documented in research and has various definitions. Some of these, containing the core concept and meaning of geocoding are:

- "the process of associating an address record with a point on a map" (Ratcliffe 2001, p. 473)
- "the assignment of a code – usually numeric – to a geographic location" (Harvard University 2008)
- [used as a verb, geocoding] is the act of transforming aspatial locationally descriptive text into a valid spatial representation using a predefined process (Goldberg 2008, p. 5).

If a company or business unit wants to spatially analyse data based on address information, both a reference dataset and a geocoding engine (geocoder) are essential. According to Goldberg (2008, p. 5), "A geocoder (noun) is a set of inter-related components in the form of operations, algorithms, and data sources that work together to produce a spatial representation for descriptive locational references."

Fig. 1 shows the main principle of geocoding: to match pairs of coordinates to existing textual descriptions of addresses. The matching process is performed through a geocoder, while the reference data provides the necessary information of both the complete address information and the

geocoding metrics, i.e., spatial data usable within a GIS – usually provided either as projected (x, y) or geographic coordinates (latitude, longitude).

The geocoding process thus provides metric location information. The quality of the geocoding process is expressed in a so called match rate, expressed as the percentage of incoming records - addresses – that can be matched to a chosen reference dataset (Cayo and Talbot 2003).

If no address points are available as reference data, linear features, i.e., roads and streets expressed on a map as lines, are often used for geocoding. Most OWMSs use linear reference data for geocoding (cf. Section 4.2). As this kind of geocoding gives information relative, for example, to a point (other than an address) on a road, it is referred to as street-geocoding (North American Association of central cancer registries 2002, Zandbergen 2007).

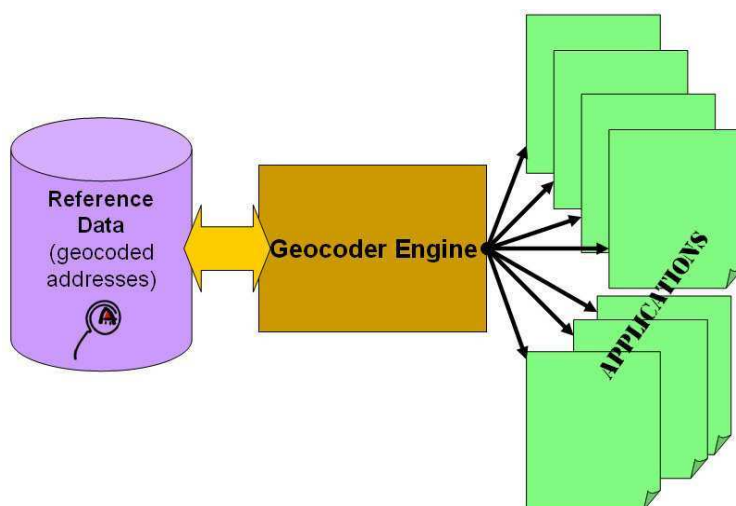


Fig. 1 Central role of a geocoding engine

In other words: a geocoder can, at best, provide geocoding results only as good as the reference data it uses. A hit rate of 100% is rare (Ratcliffe 2004). A very in-depth description of geocoding is provided by Goldberg (2008).

### 3.1.2 Use of geocoded addresses

As mentioned above in Section 1.1 (Motivation), high-resolution and high-quality geocoded address data is valuable for a wide range of applications. Swisstopo (2005) explicitly mentions application areas including emergency services (Hancock 2010), postal and delivery services, taxi, and general

logistic services. 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/epidemiology, the defining example of which is Dr. John Snow's geographic analysis of cholera distribution in the London epidemic of 1854,<sup>2</sup> micro-geographic analyses are now common (Gatrell and Senior 2005, Messina et al. 2006, Piper 2008). Fig. 2 shows a more recent example of spatial analysis in the health sector. 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).

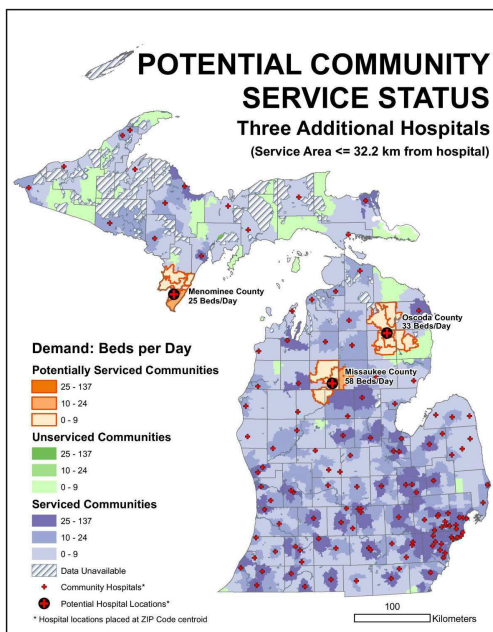


Fig. 2 Health geography map showing potential community hospital service status (source: Messina et al. (2006), <http://www.ij-healthgeographics.com/content/5/1/42/figure/F12> [online March 22, 2010])

<sup>2</sup> cf. [http://en.wikipedia.org/wiki/John\\_Snow\\_%28physician%29](http://en.wikipedia.org/wiki/John_Snow_%28physician%29) and <http://en.wikipedia.org/wiki/File:Snow-cholera-map-1.jpg> [both online March 22, 2010]

### 3.1.3 Structure of geocoded addresses

As the primary area of interest in this thesis, Switzerland provides a national standard (Schweizerische Normen-Vereinigung 2004) discussed below (in Section 3.1.4) in greater detail.

Switzerland's geocoded address records are structured as follows:

Attribute	Description
Street (Strasse)	Name of the street, e.g., "Gründenstrasse"
House number (hnr)	Number of the building, e.g., "40"
Additional address information (adrzusatz)	Additional address information (optional). This may be an alternative to a house number if, e.g., the building is known by a name instead of a street name and house number. In Switzerland, this kind of naming is fairly common, e.g., with cottages or in smaller villages, where streets may be unnamed and houses unnumbered (Schweizerische Normen-Vereinigung 2004, swisstopo 2005), e.g., "Chalet Enzian"
Zip-Code (plz)	Four digit zip or postal code, e.g., "4132"
City (ort)	Name of the locale or city, e.g., "MuttENZ"
Longitude (lon)	Numerical value of the coordinate as longitude, e.g., "7.6387"
Latitude (lat)	Numerical value of the coordinate as latitude, e.g., "47.5339"

Table 1 Swiss structure of geocoded addresses

Regarding the fields of Table 1, only the street name and house number (or, where no street name or house number are available, the “additional information”), zip code and city must be entered when a new address is captured in OA, with the additional address information normally being optional. The metric values of latitude and longitude are determined automatically by the application when indicated (clicked) on the map (cf. Fig. 17).

	<u>strasse</u>	<u>hnr</u>	<u>adrzusatz</u>	<u>plz</u>	<u>ort</u>	<u>lon</u>	<u>lat</u>
<input type="checkbox"/>  	GrÄ¼ndenstrasse	40	-	4132	Muttenz	7.63858795	47.53377151

Fig. 3 Sample address in OpenAddresses database including spatial information as lon and lat values

As shown in Fig. 4, the hierarchical structure of an address is consequently simple. Although the hierarchy presented in Fig. 4 is valid for the major part of Switzerland it may happen in rural areas that for a specific zip code more than one “city” name exist (actually villages or even smaller spatial units). In such cases, six-digit zip codes, assigned mainly for the internal use of Swiss Post, resolve the problem of multiple zip code/city relations (Schweizerische Normen-Vereinigung 2004).

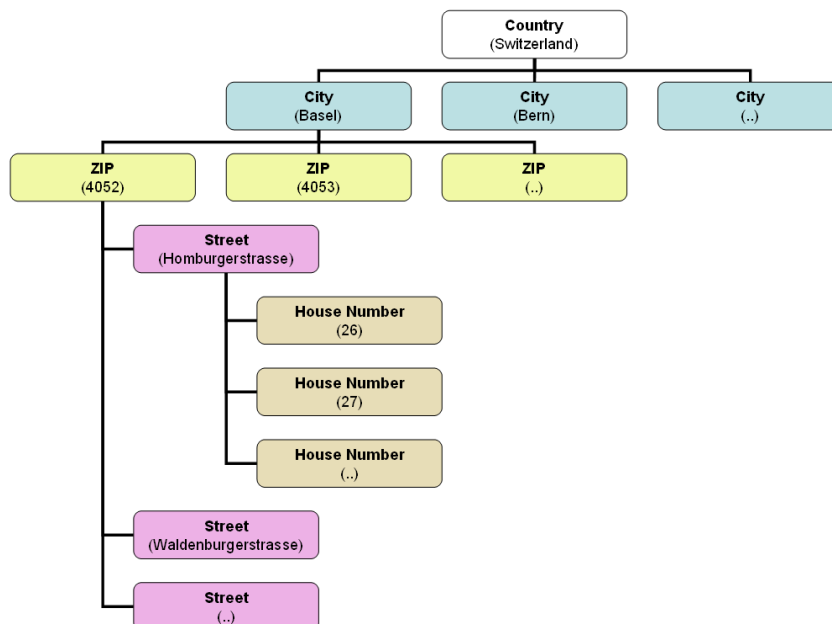


Fig. 4 Hierarchical structure of Swiss addresses

Other countries' geocoded address structures may differ, depending largely on how commonly-used addresses are constructed and whether the spatial referencing system requires a stronger hierarchy than that in Switzerland. It is also significant that if the address model supports a gazetteer-like structure, a location object must be returned even if only part of an address is entered, such as zip and city, with no street or house number. Such an example is depicted in Al Rahed et al. (2008) for South Africa. OA does not support a gazetteer-like structure.

### 3.1.4 Address Standards

In Switzerland the relevant standard for modelling geocoded addresses is that of the Schweizerische Normen-Vereinigung (2004). It is related to ISO/TC 211:19112(2003), which " [...] defines the conceptual schema for spatial references based on geographic identifiers. It establishes a general model for spatial referencing using geographic identifiers, defines the components of a spatial reference system and defines the essential components of a gazetteer."<sup>3</sup>

The Schweizerische Normen-Vereinigung (2004) defines each building's address according to the location of its front entrance. This means that a building with multiple entrances - e.g., a block of row houses - would include multiple locations, each with its own geocoded address. If the block has only one entrance, however, e.g., a shopping mall, it will correspond to only one geocoded location/address, regardless of the number of units it houses. In OA, the locational metrics of a building's number generally represent the location of its centroid. In the case of the block of row houses, the location of each house number must be set manually to approximate the location of that house within the block. The same principle should be applied to a block with only one entrance but housing multiple business or residential units (cf. Fig. 5).

---

<sup>3</sup> [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=26017](http://www.iso.org/iso/catalogue_detail.htm?csnumber=26017) [online March 17, 2010]

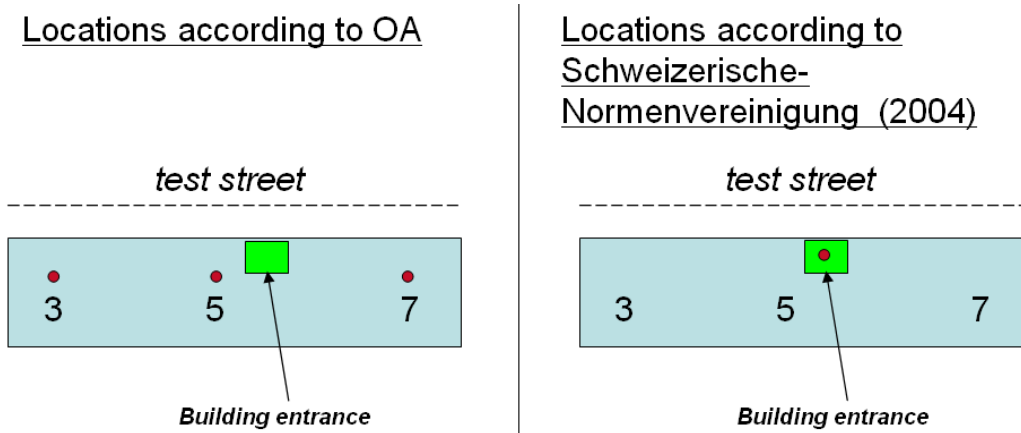


Fig. 5 Comparison of address location definitions

The data model or schema defined by the Schweizerische Normenvereinigung (2004) is presented in Fig. 6. This shows a relatively complex structure that includes maintenance ('Nachführung') and uses the six-digit zip code approach ('PLZ6'). The grey boxes relate to administrative units responsible for the contents of each of the model's zones. E.g., Swiss Post is responsible for zip codes, while the majority of the address information, including the street name, house number and location, must be administered by the municipality in which the building is located.

# Quality Assurance of OpenAddresses

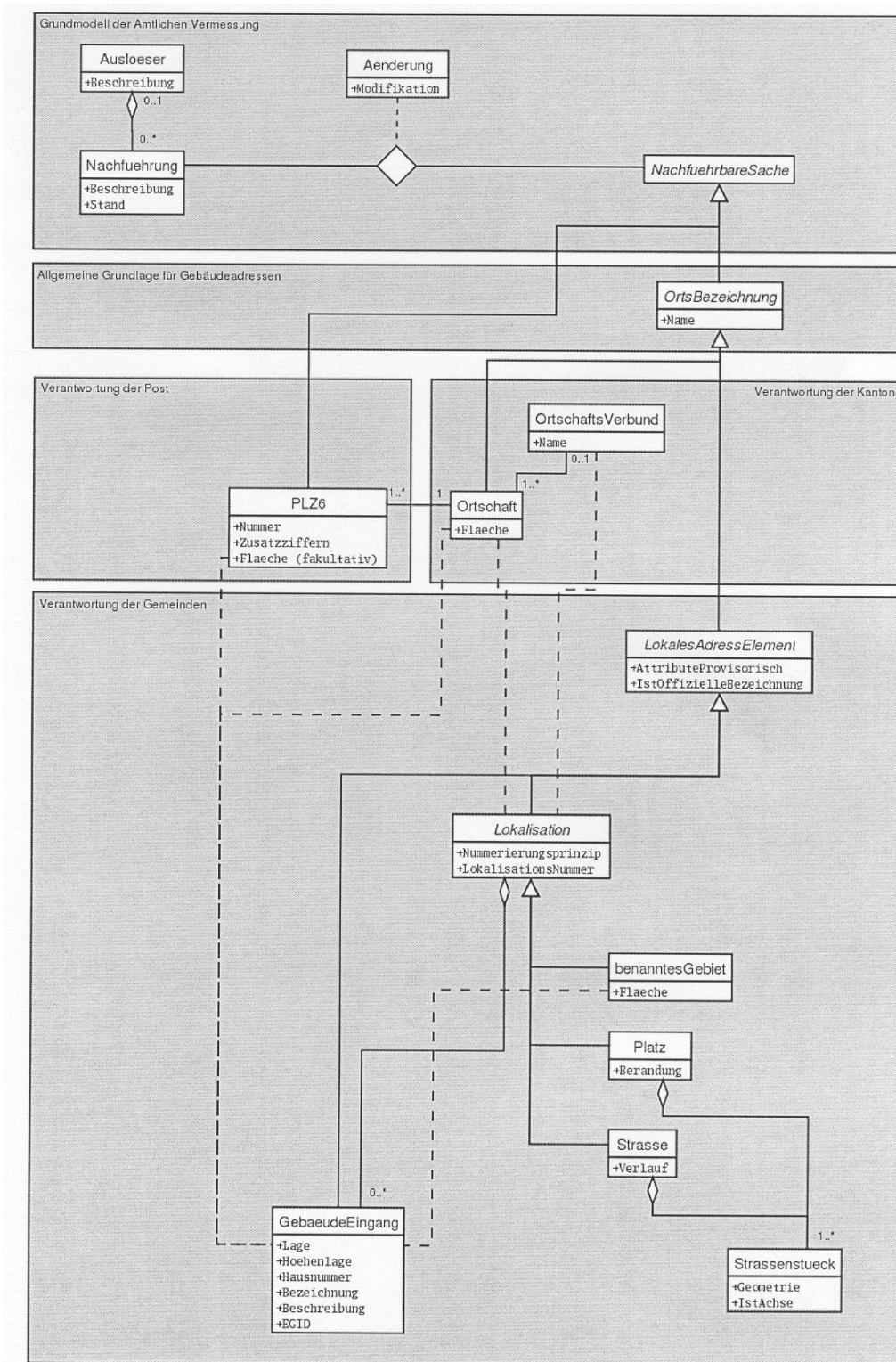


Fig. 6 Swiss Data model in UML<sup>4</sup> (source: Schweizerische Normen-Vereinigung (2004, p. 11))

<sup>4</sup> UML: Unified Modeling Language, cf. <http://www.uml.org/> [online April 6, 2010]

The Schweizerische Normen-Vereinigung (2004) model also deals with incomplete address information that may result from very rural environments or idiosyncratic local addressing systems. A simplified form of this model, also formulated by the Schweizerische Normen-Vereinigung (2004), corresponds closely to the structure of the OA data model (cf. Sections 3.1.3 and 3.3.2).

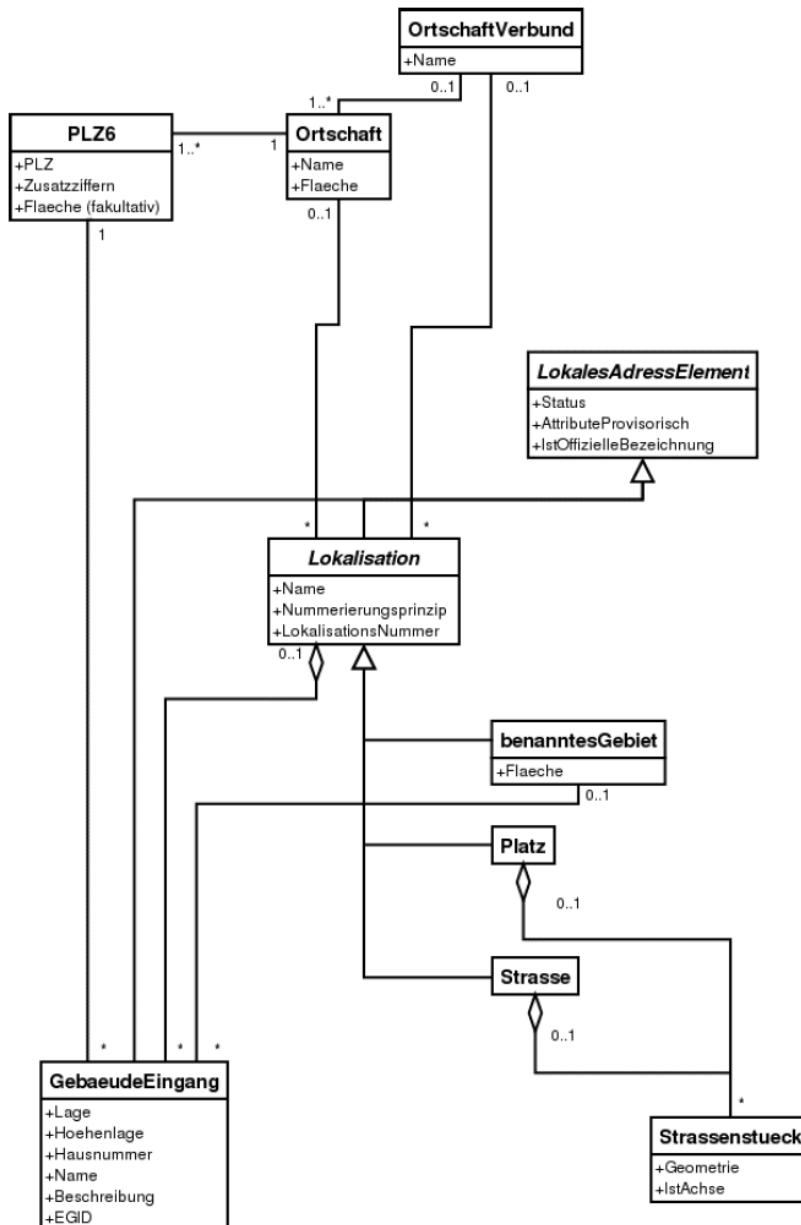


Fig. 7 Simplified view of the Swiss address model in UML (source: Schweizerische Normen-Vereinigung (2004, p. 16))

## Quality Assurance of OpenAddresses

Fig. 8 to Fig. 11 illustrate both the structure and setup of Swiss addresses (all examples used here are taken from swisstopo (2005)). These figures also show different numbering systems any of which may include gaps in the numbering. While street-geocoding algorithms are often incapable of handling such omissions, a data repository such as OA handles them well.

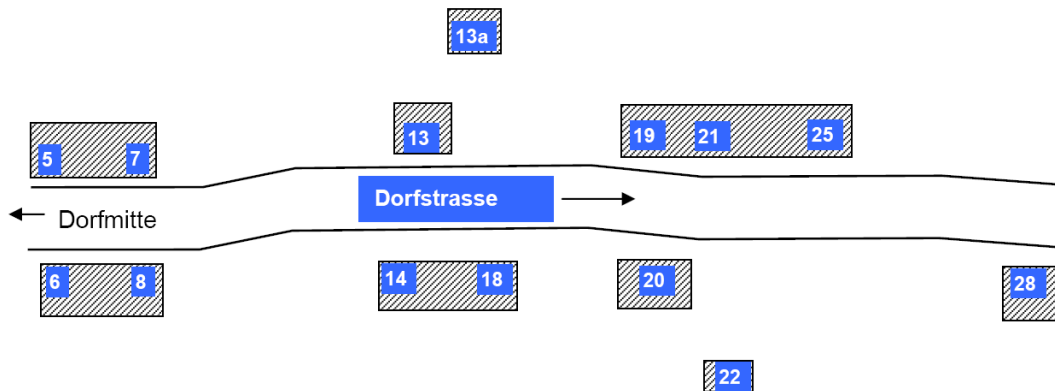


Fig. 8 General address structure and naming along a street (source: swisstopo (2005, p. 9))

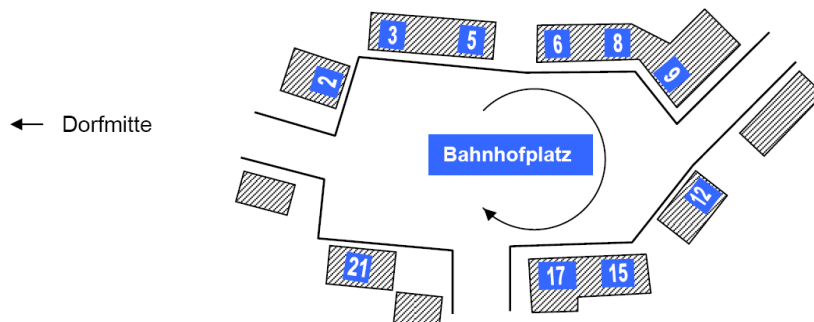


Fig. 9 General address structure and naming around places/squares (source: swisstopo (2005, p. 9))

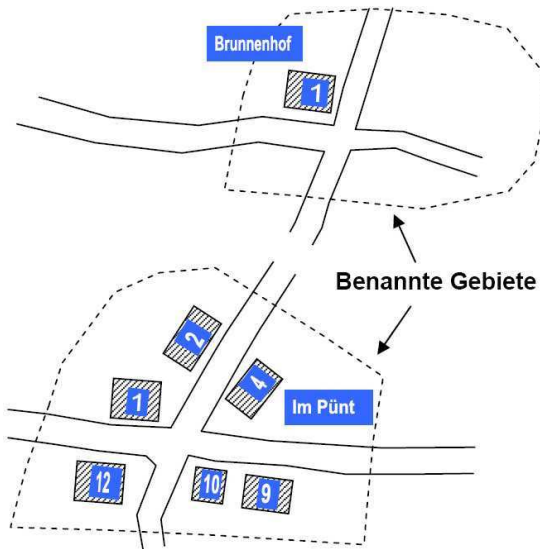


Fig. 10 General address structure and naming within named areas (source: swisstopo (2005, p. 10))

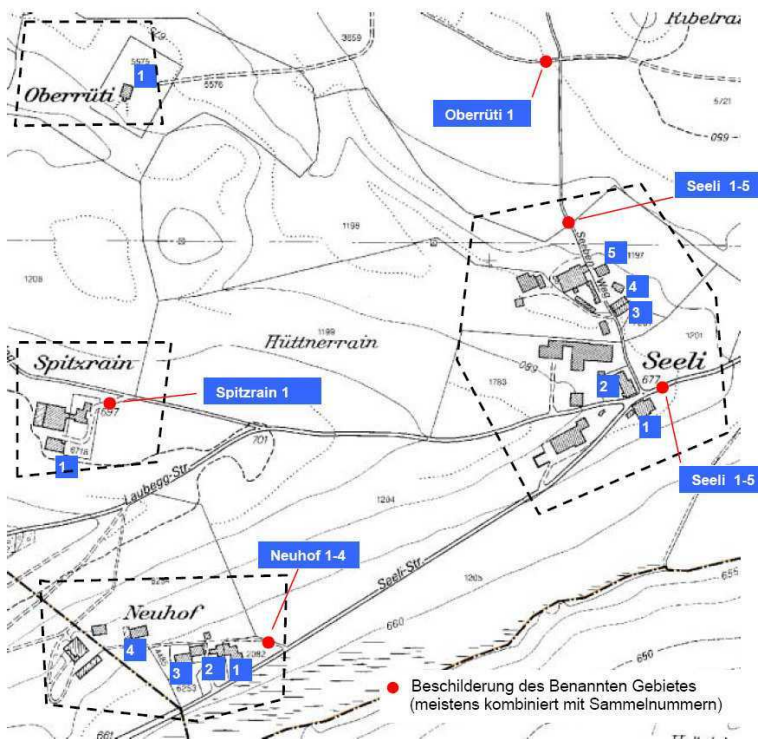


Fig. 11 General address structure and naming in hamlets (source: swisstopo (2005, p. 24))



The data model used in South Africa is presented in Al Rahed et al. (2008) and shown in Fig. 13.

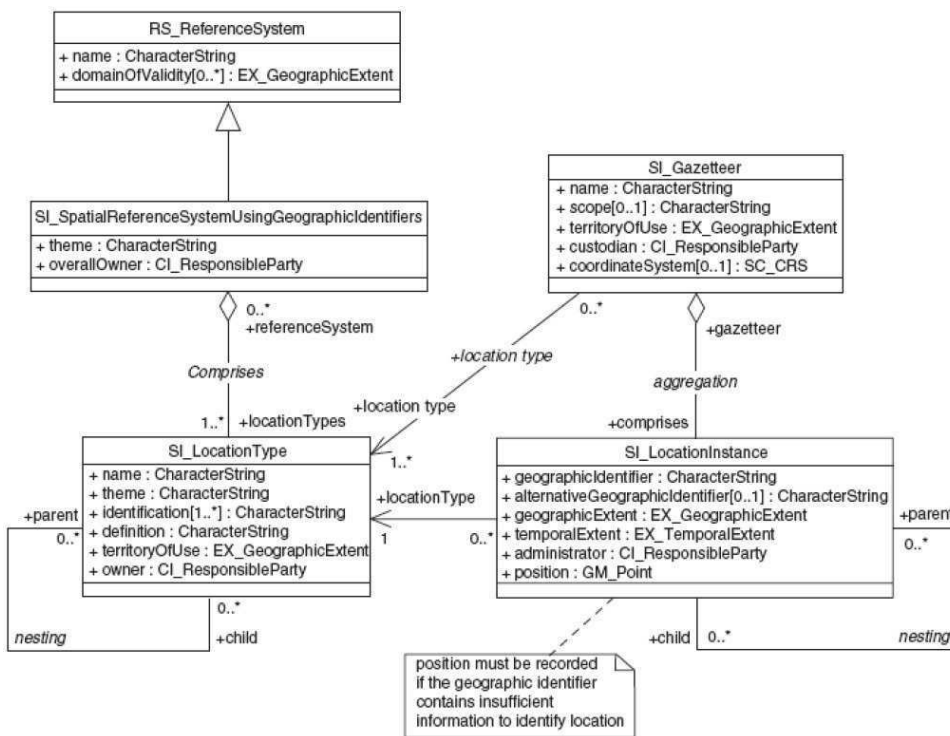


Fig. 13 Data model used by the Intiando address matching tool in South Africa based on ISO/TC 211:19112 (2003) (source: Al Rahed et al. (2008, p. 209)).

Coetzee et al. (2008) give an overview of a range of national and international addressing standards. Not all of these include geo-referencing addresses via co-ordinates and some cover various address types, as opposed to postal addresses only. However, the paper provides an excellent synopsis of the current state of address standardisation, while elaborating the economic impacts and benefits of standardising address structures internationally:

"A European survey on addresses and address data (EUROGI 2005) gives clear evidence that although address systems exist in European countries, with a long history as well, and although address master files or address registers are available in most countries on certain conditions, only very few published standards for address data exist, making the task of 'interoperable and seamlessly accessible' address data sets 'across all of Europe' even more difficult." Coetzee et al. (2008, p. 7)

### **3.2 Volunteered Geographic Information**

The general concept of volunteer-contributed geographic information has been described by many authors and is well documented (Coleman et al. 2009, Elwood 2008b, 2009, Flanagan and Metzger 2008, Fischer 2008, 2009). However, it was Goodchild who coined the term 'volunteered geographic information' (VGI), along with a succinct description of it:

"A remarkable phenomenon ... has become evident in recent months: the widespread engagement of large numbers of private citizens, often with little in the way of formal qualifications, in the creation of geographic information, a function that for centuries has been reserved to official agencies. They are largely untrained and their actions are almost always voluntary, and the results may or may not be accurate. But collectively, they represent a dramatic innovation that will certainly have profound impacts on geographic information systems (GIS) and more generally on the discipline of geography and its relationship to the general public." Goodchild (2007, p. 212)

The basic concept of VGI is just another form of PPGIS (Harris et al. 1995, Elwood 2007), but taking advantage of modern technical infrastructure such as handheld Global Positioning System (GPS<sup>5</sup>) receivers, the internet, and Web 2.0<sup>6</sup> applications incorporating asynchronous JavaScript<sup>7</sup> and XML<sup>8</sup> (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, p. 4) has gone so far as to call VGI "geography without geographers". Coleman et al. (2009) give a typology of overlapping categories of VGI project contributors.

As a phenomenon, of course, large-scale projects based on voluntary contributions are neither new nor restricted to spatial data. The most prominent example is Wikipedia, "a free, web-based, collaborative,

---

<sup>5</sup> <http://www.point-inc.com/glossary.php#gps> [online March 17, 2010]

<sup>6</sup> <http://www.share.uni-koeln.de/?q=en/glossary/29#letterw> [online March 17, 2010]

<sup>7</sup> <http://www.javascript.org> [online March 17, 2010]

<sup>8</sup> Extensible Markup Language, cf. <http://www.w3.org/XML> [online March 17, 2010]

multilingual encyclopaedia project supported by the non-profit Wikimedia Foundation."<sup>9</sup> However, while only a tiny fraction of Wikipedia users are also contributors (Wales 2005), the majority of VGI users are also data producers (cf. Fig. 15), often called either 'producers' (producer-users) (Coleman and Georgiadou 2010) or prosumers (producer-consumers<sup>10</sup>) for VGI projects.

In addition to offering opportunities for new applications and innovations, the use of uncertified contributors raises significant issues of quality and credibility: both Cooper (2009) (cf. Section 3.4.2) and Sui (2008) see VGI as a serious challenge to the traditional authoritative model of geographic information provision.

However, according to Goodchild (2007), VGI is often not only the cheapest but the only source of geographic information. As was shown in the aftermath of the major earthquake that struck Haiti in January, 2010, this effect is particularly evident in crisis mapping.<sup>11</sup>

### **3.2.1 VGI, standards and spatial data infrastructures**

Auer and Zipf (2009) and Zipf (2009) advance arguments that VGI projects must conform to open standards, thereby supporting interoperability and providing favourable conditions for a wide range of open applications. In this context, Zipf (2009) notices a close relationship between VGI and spatial data infrastructures (SDI), proposing a concept of architectures that take advantage of both open standards and VGI based services. Cooper (2009) concurs, emphasizing VGI's implicit economic and innovational potential. To realise this potential, standards are essential, as they empower system independent interfaces. This thesis investigates the suitability of ISO TC211 standards as a possible basis of VGI quality assessment.

---

<sup>9</sup> <http://en.wikipedia.org/wiki/Wikipedia> [online March 17, 2010]

<sup>10</sup> <http://en.wikipedia.org/wiki/Prosumer> [online March 17, 2010]

<sup>11</sup> <http://radar.oreilly.com/2010/01/haiti-osm-and-sat-imagery-for.html>,  
<http://povesham.wordpress.com/2010/01/18/haiti-how-can-vgi-help-comparison-of-openstreetmap-and-google-map-maker>,  
<http://www.fiducialmark.com/2010/01/14/haiti-earthquake-mapping>,  
<http://opengeodata.org/the-first-week-of-humanitarian-openstreetmap>,  
<http://opengeodata.org/bravo-osm-haiti-editors-you-saved-lives> [all online April 2, 2010]

## 3.2.2 VGI Projects

The following section briefly introduces a number of the many current VGI-based projects, with particular attention to OpenStreetMap and OpenAddresses. This section is intended not to give a complete overview or a typology of VGI uses, but to provide a range of samples and describe their main characteristics.

### 3.2.2.1 OpenStreetmap

With its strong media presence<sup>12</sup>, OpenStreetMap (OSM) is currently the best-known VGI project. Accessible free of charge via [www.openstreetmap.org](http://www.openstreetmap.org), it provides online mapping that is interactive, searchable, multilayered, and editable (cf. Fig. 14). OSM's administrators describe it as "like Wikipedia for maps. People gather location data with GPS devices or from free satellite imagery, upload it and add names and other tags."<sup>13</sup>

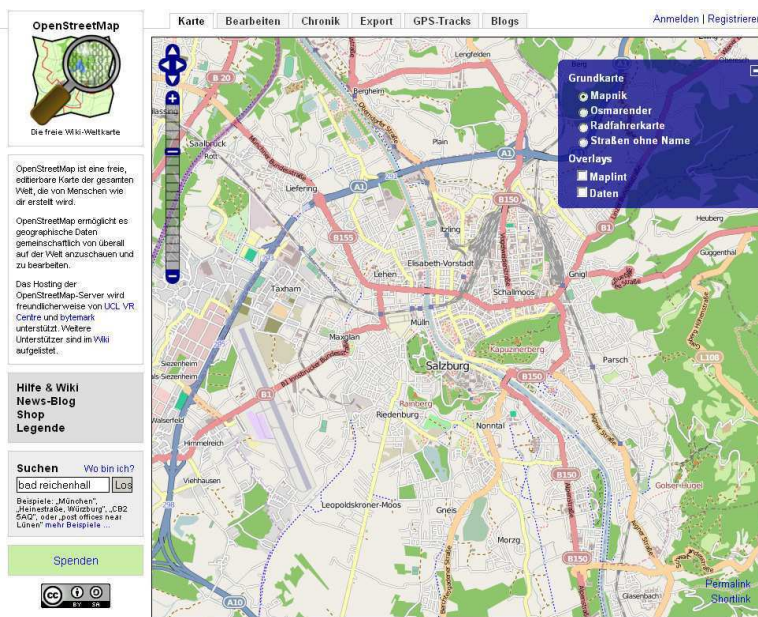


Fig. 14 Map view of OpenStreetMap in a web browser (source: [www.openstreetmap.org](http://www.openstreetmap.org) [online March 18, 2010]).

<sup>12</sup> [http://wiki.openstreetmap.org/wiki/OpenStreetMap\\_in\\_the\\_media](http://wiki.openstreetmap.org/wiki/OpenStreetMap_in_the_media) [online March 18, 2010]

<sup>13</sup> [http://wiki.openstreetmap.org/wiki/Beginners%27\\_Guide](http://wiki.openstreetmap.org/wiki/Beginners%27_Guide) [online March 18, 2010]

The OSM project was initiated by Steve Coast in 2004, and has since spread throughout the world (Ramm and Topf 2006). As shown in Fig. 15, its engagement statistics indicate impressive growth:<sup>14</sup>

Number of users	231,943
Number of uploaded GPS points	1,450,803,913
Number of nodes	571,668,555
Number of ways	42,985,580
Number of relations	387,213

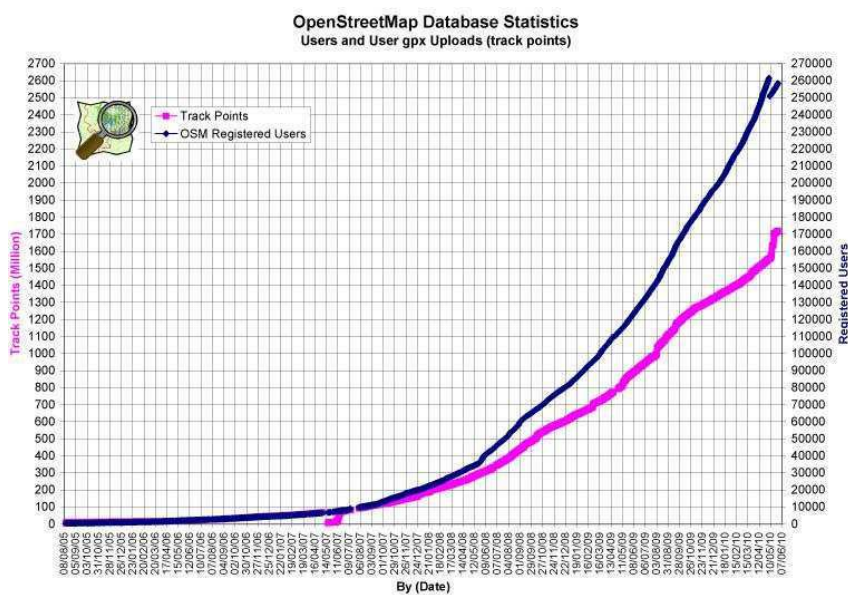


Fig. 15 OSM database statistics (source: <http://wiki.openstreetmap.org/w/images/9/91/Osmdbstats1.png> [online March 18, 2010]).

To participate in OSM, i.e., to contribute data to the OSM database, users must first register, after which a user interface allows them to upload their raw data – such as GPS tracks from field sessions - and process these using tools such as Potlatch.<sup>15</sup> The data are then entered into the OSM database for rendering. The OSM website provides clear documentation to guide the user through the major steps of the process, from initial data collection to final rendering of the resulting map (cf. Fig. 16).

<sup>14</sup> [http://www.openstreetmap.org/stats/data\\_stats.html](http://www.openstreetmap.org/stats/data_stats.html) [online March 18, 2010]

<sup>15</sup> <http://wiki.openstreetmap.org/wiki/Potlatch> [online March 18, 2010]

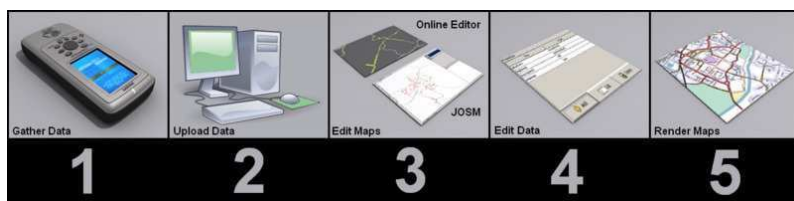


Fig. 16 The five steps of contributing data to the OSM project database (source: [http://wiki.openstreetmap.org/wiki/Beginners%27\\_Guide](http://wiki.openstreetmap.org/wiki/Beginners%27_Guide) [online March 18, 2010]).

The OSM project webpage is generally well documented, and also provides links to further information.

### 3.2.2.2 Other Projects

In addition to OSM, numerous community-based projects use VGI to produce maps. Several of these are summarized in Section 3.2.3.2.1., with Section 3.2.3.2.2 covering commercial applications of VGI. Another project, OpenAddresses (OA), which is built on a basic idea and concept very similar to those of OSM, will be described in detail in Section 3.3, as it is one of the central topics of this thesis. These projects are presented to draw a clearer picture of how crowd-sourced geo data projects operate and what their goals are.

#### 3.2.2.2.1 Community based VGI

Three examples of projects that share the "spirit" of OSM are listed below. Further projects are listed on the Open Source Geospatial Foundation (OSGEO) website ([http://wiki.osgeo.org/wiki/Public\\_Geospatial\\_Data\\_Project](http://wiki.osgeo.org/wiki/Public_Geospatial_Data_Project)).

##### OpenSeaMap

[www.openseamap.org](http://www.openseamap.org)

"An open-source, worldwide project to create a free nautical chart,"<sup>16</sup> OpenSeaMap was founded in 2009, and provides information on marine mapping, buoys, harbors, etc. The project's web site contains a wiki, a frequently asked question (faq) section and other information.

---

<sup>16</sup> <http://www.openseamap.org/index.php?id=faq&L=1> [online March 18, 2010]

**OpenAerialMap**  
[www.openaerialmap.org](http://www.openaerialmap.org)

"Open Aerial Map is a non-profit, open-access meeting place for the aerial imaging community. It exists to provide a freely available image map of the world created solely by community contribution, and to facilitate the free exchange of imagery, technology, and ideas. In order to provide an unrestricted, free, an unbiased view of the world, OpenAerialMap encourages the free exchange of imagery, without restriction on its use."<sup>17</sup>

Because it would allow many other projects to use free Orthophoto images to digitize and extract information, OpenAerialMap's potential is considered high. Unfortunately, at the time of writing, none of the project's proposed services are yet available, and its activities focus on a mailing list to discuss relevant issues.

**GeoNames**  
[www.geonames.org](http://www.geonames.org)

"The GeoNames geographical database is available for download free of charge under a Creative Commons attribution license. It contains over eight million geographical names and consists of 7 million unique features [...including] 2.6 million populated places and 2.8 million alternate names."<sup>18</sup>

Geonames provides not only a freely available database of location names but also a set of related web services. These services are useful for other web applications – including, for example, OSM, which uses freely available geographic data, shows them in a browser using tools such as OpenLayers,<sup>19</sup> and employs Geonames to help locate and depict specific areas.

---

<sup>17</sup> [http://www.openaerialmap.org/Main\\_Page](http://www.openaerialmap.org/Main_Page) [online March 18, 2010]

<sup>18</sup> <http://www.geonames.org/about.html> [online March 18, 2010]

<sup>19</sup> <http://www.openlayers.org> [online March 18, 2010]

### **3.2.2.2.2 Commercially oriented VGI**

The use of VGI carries no requirement that projects be designed, maintained and managed, like those presented above, as *community based* services.

Another form may be labelled as *commercially oriented VGI*, i.e., enterprises that take advantage of VGI data for commercial gain.

One well-known example of this category is Google Inc., which launched Google Map Maker in June 2008.<sup>20</sup> Map Maker offers all users the option of mapping objects as points, lines or polygons. These options can be applied to real-world objects not detailed in Google Maps' existing views. New objects can be landmarks, buildings, paths, or natural objects of interest. Unlike with OSM, though, the collected data is stored within Google's proprietary database and cannot be downloaded by users for further use or individual processing. Further, the resulting views are accessible only through Google applications.

Regarding navigation services, two other commercial users of VGI are Tele Atlas and Navteq, both of which provide consumer navigation data. With the goal of keeping their street network data as current as possible, Tele Atlas invites users to update its database (e.g., to report a newly-opened bridge or road) using its Map Insight feedback and change reporting software.<sup>21</sup> Similarly, Navteq, Tele Atlas's principal competitor, offers users Map Reporter,<sup>22</sup> its own interactive web-based change reporter. As with Google Map Maker, both of these systems store user-supplied data in proprietary databases.

---

<sup>20</sup> <http://www.google.com/Map Maker> [online March 20, 2010]

<sup>21</sup> [http://www.teleatlas.com/WhyTeleAtlas/Pressroom/PressReleases/TA\\_CT023824](http://www.teleatlas.com/WhyTeleAtlas/Pressroom/PressReleases/TA_CT023824) [online March 20, 2010]

<sup>22</sup> <http://mapreporter.navteq.com> [online March 20, 2010]

A mixture of community based and commercially oriented VGI can be seen in the UK-based People's Map project.<sup>23</sup> The project's website introduces it as follows:

"The People's Map is an exciting new mapping concept which enables any individual or organisation to create and maintain maps of Britain by 'drawing' features like roads, land use and point of interests over aerial photography, using simple online editing tools.

The information is checked and compiled to create high quality cartographic maps in various scales that can be used either as a source for finished maps or as the basis for the creation of new maps. The People's Map data is free from third party copyright, with fair and straight forward licensing."<sup>24</sup>

Hence, People's Map is basically community based VGI. However, the commercial model is variable: if the data is used for private or non-commercial purpose it is free, if it is used commercially it is not. This is an interesting approach because it explicitly allows commercial customers to use crowdsourced data without having to conform to the requirements of open licenses such as creative commons.<sup>25</sup> Depending on their definitions and restrictions, open licenses can be somewhat limiting for commercial applications, as they may stipulate that products based on the maps they cover remain the property of the original license holder, i.e., without appropriate licensing, it may be impossible to derive proprietary products directly from open-license projects.

### **3.3 OpenAddresses**

The OpenAddresses (OA) open geo-data project is the subject upon which this thesis's quality assessment will be performed and documented. Thus it is presented in greater detail than those above.

---

<sup>23</sup> <http://peoplesmap.com> [online March 20, 2010]

<sup>24</sup> <http://peoplesmap.com> [online March 20, 2010]

<sup>25</sup> <http://creativecommons.org> [online March 20, 2010]

## Quality Assurance of OpenAddresses

The project was founded in 2007 at the School of Architecture, Civil Engineering and Geomatics of the Northwestern Switzerland University of Applied Sciences (Fachhochschule Nordwestschweiz) by the author of this thesis, who later also collaborated closely in its development. Conforming to the concept of community based VGI, the OA project has been presented in papers and at conferences (Stark 2008, 2009). The following is a brief description of its functionality, focusing on aspects relevant to this thesis.

### 3.3.1 Scope and Intention

The initial aim of OA was to collect geocoded addresses in Switzerland, using a collection process designed for volunteer participation. A very simple interface based on Google Maps was created to offer a three step data collection process:

1. click on the map at the address location
2. enter address information according to Section 3.1.3
3. confirm data



Fig. 17 Data collection in OpenAddresses map view container (source: screen capture from [www.openaddresses.ch](http://www.openaddresses.ch) [online April 14, 2010])

Only limited manipulation of an address is possible. If its position on the map requires adaptation, for example, it is possible to move it to another location, after which the positional information will be updated in the database. If attribute values such as street name or place name must be corrected, however, the entire address must be deleted and completely re-entered. This process has been designed specifically to discourage malicious edits.

Addresses can be captured individually or as listings to be processed sequentially. OA also offers the ability to process GPS exchange (GPX) data, i.e., track points collected via a GPS device. In this case the registered GPS positions are shown on the map and the user must confirm their final position by simply moving the GPS based address location to its correct position on the map. This is necessary because GPS track points are generally registered from in front of or next to a building but not from within it.

### 3.3.2 Technological Environment

OA works as a client-server application. On the client side, JavaScript must be activated to run it. The client's web browser links to `openaddresses.ch` and opens the web page via `http`<sup>26</sup>. Since OA is a Google Maps mashup,<sup>27</sup> the Google Maps JavaScript library, along with the OA functions implemented in JavaScript, will be downloaded to the client.

On the server side, PHP<sup>28</sup> processes client requests, answers them and delivers replies. For its database system, it uses PostgreSQL<sup>29</sup> with PostGIS.<sup>30</sup>

OA uses a very simple data model (cf. Fig. 18).

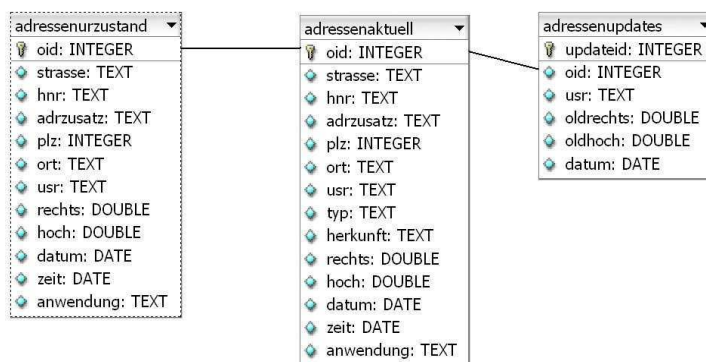


Fig. 18 Data schema of OpenAddresses

<sup>26</sup> `http`: hypertext transfer protocol; <http://www.beonet.rs/?strana=220#HTTP> [online March 20, 2010]

<sup>27</sup> [http://www.quackit.com/web\\_design/web\\_design\\_glossary.cfm](http://www.quackit.com/web_design/web_design_glossary.cfm) [online March 20, 2010]

<sup>28</sup> PHP: hypertext preprocessor; <http://shopscrip.hu/glossary.html#php> [online March 20, 2010]

<sup>29</sup> <http://www.postgresql.org> [online March 20, 2010]

<sup>30</sup> <http://postgis.refrains.net> [online March 20, 2010]

## Quality Assurance of OpenAddresses

Besides the information presented in Section 3.1.3 above, a user name, date and time, and the coordinates supplied by Google Maps are stored in the OA database. If an address is entered for the first time, it is stored both in 'adressenurzustand' (original condition of address) and 'adressenaktuell' (current address) tables. 'Adressenaktuell' contains all addresses in their current state while 'adressenurzustand' stores the first information entered, along with the position where the user initially clicked.

If an address is moved on the map, its last position before the move is stored in the 'adressenupdates' (address updates) table and the fields 'rechts' (right) and 'hoch' (top), containing longitude and latitude values, are updated in the 'adressenaktuell' table.

If an address is deleted it is not removed from the database, but its state (field 'typ' in table 'adressenaktuell') is set to 'delete'.

### **3.3.3 Current State of OpenAddresses**

OA offers not only its collection of addresses but its entire database for download. Between the start of the project in 2007 and March 2010, OA incorporated about 174,000 addresses, the vast majority of which were donated by cadastres such as the Cantons of Zug and Solothurn and municipalities such as Morges. Besides address data, OA offers two free tools<sup>31</sup> – one for data capture (GeoDataSnapper), and one for geocoding address data (Geocoder).

## **3.4 Quality Assessment**

This section defines '*quality*' in the context of OA and its quality assessment, after which experiences with quality assessment of VGI are presented, leading to an introduction to the family of ISO/TC 211 19100 standards. Its conclusion discusses these matters' implications on geocoded addressing.

---

<sup>31</sup> In the download section of OA website

### 3.4.1 Definition of 'Quality'

The term 'quality' expresses various unquantifiable characteristics, and no consensus can be found among experts on a single definition as will be presented in this section. For some people, for example, 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 latter understanding leads to another important term: 'fitness for use' (Jakobsson and Tsoulos 2007), the basic implication of which is that, used in different contexts, the same product may conform to one context's quality requirements but not to another's. Thus a declaration of quality is only possible in the context of a specific application: the 'fitness for use' approach requires no judgement of a product's overall abstract quality but simply reveals, for example, the concrete results of a set of tests (Chrisman 2006). In order to assess 'fitness for use', then, Oort (2006) suggests a multi-step evaluation procedure, taking the spatial quality into account.

According to Goodchild (2006) the quality of spatial data is crucial to its effective use. He defines it 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; p. 13).

Devillers and Jeansoulin (2006) distinguish further between internal and external quality. "Internal quality corresponds to the level of similarity that exists between the data produced and the 'perfect' data that should have been produced" (Devillers and Jeansoulin 2006; p. 37). Assessing quality as spatial accuracy, internal quality must be measured by some external tool comparing the dataset to an external reference. Internal quality can be described with various criteria listed in ISO/TC 211:19113 (2001). These criteria are elaborated in Section 3.4.3.

## Quality Assurance of OpenAddresses

In contrast, external quality is much closer to the definition of fitness for use: its concept "[...] corresponds to the level of concordance that exists between a product and user needs, or expectations, in a given context" (Devilleers and Jeansoulin 2006; p. 39). The implication is that external quality is not an absolute value or measure, but rather one measurable only in the context of use.

Flanagin and Metzger (2008) distinguish between credibility as a subjective measure of quality, and accuracy as an objective measure: experts are seen as credible (although they may occasionally be wrong) if they produce accurate data. With OA, the goal must be to provide verifiably accurate data, which may lead to widespread trust both within the community and the consumers. To achieve this goal, one useful tool is a reasonable and functional quality assessment based on international standards, such as the one presented in this thesis.

Fisher et al. (2006) list six principal areas of data quality from a perspective related to uncertainty in spatial data:

- *lineage* - information on the origin and history of data
- *accuracy* - expressed as positional accuracy and that of attribute values
- *completeness* - does the data cover and contain all features from the real world
- *logical consistency* - such as topology
- *semantic accuracy* - refers to classification of data
- *currency* - how up-to-date is the data?

Oort (2006) extends this list to include usage, purpose, constraints, variation in quality, meta-quality and resolution. Some of the aspects listed have been incorporated into ISO/TC 211:19113 (2001). While not all aspects of Fisher et al. (2006) and Oort (2006) are taken into account in this thesis during the evaluation process, positional and attribute accuracy are investigated (cf. Section 4.3).

This thesis's quality assessment of OA data looks to the concepts of both internal and external quality. Firstly, three OWMSs data are evaluated against a reference dataset from the Canton of Solothurn (internal quality); secondly, OA is assessed based on the results of the OWMSs' evaluations (external quality).

The result is information for each feature (address) in OA as an implicit measure of fitness for use – one appraisable by each user individually. This information should also comply with the recommendations of Boin and Hunter (2006), who found that quality information regarding spatial data – in the sense of fitness for use – must be communicated in a way that the consumer is able to both understand and apply the results to his needs. Hence, a web page will inform interested users about the OWMS based quality assessment of each OA record, along with a recommendation as to whether this address is acceptably accurate or needs revision.

Finally, the issue of malicious data entry must be addressed. There is a potential within any VGI project that data will be 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. This thesis evaluates whether and how, with the use of OWMS, such "malicious data" can be detected or at least indicated in OA.

#### **3.4.2 Quality Assurance with VGI**

Since the core concept of VGI is to collect data via volunteers, i.e., "the crowd", the same population is generally also responsible for the quality of the data produced. According to Agichtein et al. (2008) the quality of user generated content varies from excellent to bad. Haklay (2008), Amelunxen (2009), and Auer and Zipf (2009) all agree. As no superior authority is responsible for the assessment of geospatial data quality, the community approach is inherently more vulnerable to general quality management issues than data products created by professionals or experts (Goodchild 2007). Additionally, because VGI data is generally in a constant state of flux, there is basically nothing similar to a version or release, as one would find in commercial products, which also makes quality assessment more difficult.

## Quality Assurance of OpenAddresses

Quality and trust are two of the main issues faced by VGI projects, both in research and in practice (Goodchild 2007, 2008b, Sinclair 2007, Baumann 2008, Elwood 2008a, 2009, Flanagan and Metzger 2008, Bovard 2009, Coleman et al. 2009). Further, Haklay (2008) found in his investigations with OSM data that errors in spatial data in VGI are not randomly distributed, but depend on individual contributors' diligence. This insight emphasizes the need for VGI-targeted quality assurance mechanisms.

Concerning quality issues with crowd sourced data, the example of Wikipedia is often cited (Goodchild 2007, Baumann 2008). Wikipedia is seen as a pioneer in its use of volunteers to perform quality assurance: acting as volunteers, experts comment on or correct articles in their fields. Wikipedia also uses electronic tools such as WikiScanner to assess and safeguard quality to a certain degree<sup>32</sup> (Coleman and Georgiadou 2010). This tool matches articles or edits by unregistered contributors with the IP addresses from which they are sent. This information can be evaluated statistically to assess contributor credibility. Although no registration is necessary in OA, a user name must be entered for statistical evaluation. The WikiScanner approach could potentially be applied to OA: if a certain user consistently entered addresses showing inaccurate spatial values, that user's contributions could automatically be reported to an OA administrator for testing. Currently, however, no 'OpenAddressesScanner' has yet been implemented; while such a development is outside the scope of this thesis, it should nevertheless be kept in mind for future quality assurance studies.

---

<sup>32</sup> <http://de.wikipedia.org/wiki/Wikipedia:WikiScanner> [online March 26, 2010]

Bovard (2009) suggests three VGI quality assurance methods that can be applied by volunteers to improve potential users' sense of trust in VGI projects:

- A *photo* or *other image* of the real-world object could be attached to its digital representation for visual quality assessment. For OA this could be a screenshot of the digitised address and its surroundings, preferably using a hybrid map consisting of an aerial image and semi-translucent vector data with labels of street names.
- A *number* of volunteers could collect the same information. This approach is based on the assumption that the more volunteers document the same feature the higher the probability that the feature actually exists and is correctly located. For OA this strategy would not be appropriate because there would be little motivation for volunteers to collect addresses that had already been entered.
- *Registration*. This strategy would require the registration of anyone contributing data to OA. A registered user's trustworthiness would then be measured on both the number of addresses he contributed and their quality. Since OA currently neither requires nor supports user registration this method would be unfeasible.

Cooper (2009) considers the quality of VGI uncertain, as no metadata is associated with it and its sources lack authority. The idea of metadata is applied to a certain degree within this thesis to provide estimates, so that potential OA consumers can determine whether OA data fits their needs.

Referring to OSM, tools are currently available to support quality assurance<sup>33</sup> but "[...] there is no integrated quality assurance mechanism that allows participants to rate the quality of the contributions of other participants" Haklay (2008; p. 21).

---

<sup>33</sup> [http://wiki.openstreetmap.org/wiki/Quality\\_Assurance](http://wiki.openstreetmap.org/wiki/Quality_Assurance) [online March 26, 2010]

Hanguët (2006) asks generally - not specifically in the context of VGI - whether spatial data, as a representation of features of the real world, is faithful to what should be represented. In the context of OA this could be paraphrased, “Is the provided position of each given address correct and does it satisfy spatial accuracy requirements?” Providing metadata on positional accuracy as a reference dataset for each address may help answer this question.

### **3.4.3 ISO/TC 211 19100 Series**

This section briefly introduces the ISO/TC 211 standards relevant to this thesis. This introduction is useful to understand the context of how ISO/TC 211 standards work together and assist in the assessment of spatial data quality. Their application with regard to OA is presented in Section 4.3.

Technical committee 211 of the International Organisation for Standardisation (ISO) has developed standards for geographic information in the 19100 family. Some of those that are relevant to this thesis are introduced in the following sections.

#### **3.4.3.1 Geographic information - Quality principles**

ISO/TC 211:19113 (2001) establishes principles for describing the quality of geo-spatial data. It further specifies two types of components for reporting quality information and the organisation of information about data quality: data quality elements and data quality subelements (cf. Fig. 19).

data quality element	data quality subelement	definition
completeness	commission	excess data present in a dataset
	omission	data absent from a dataset
logical consistency	conceptual consistency	adherence to rules of the conceptual schema
	domain consistency	adherence of values to the value domains
	format consistency	degree to which data is stored in accordance with the physical structure of the dataset
	topological consistency	correctness of the explicitly encoded topological characteristics of a dataset
positional accuracy	absolute or external accuracy	closeness of reported coordinate values to values accepted as or being true
	relative or internal accuracy	closeness of the relative positions of features in a dataset to their respective relative positions accepted as or being true
	gridded data position accuracy	closeness of gridded data position values to values accepted as or being true
temporal accuracy	accuracy of a time measurement	correctness of the temporal references of an item (reporting of error in time measurement)
	temporal consistency	correctness of ordered events or sequences, if reported
	temporal validity	validity of data with respect to time
thematic accuracy	classification correctness	comparison of the classes assigned to features or their attributes to a universe of discourse (e.g. ground truth or reference dataset)
	non-quantitative attribute correctness	correctness of non-quantitative attribute
	quantitative attribute accuracy	accuracy of quantitative attributes

Fig. 19 Tabular overview of data quality elements and data quality subelements with definitions (source: ISO/TC 211:19138 (2006; p. 3))

For each data quality sub-element, seven descriptors are defined to record quality information where applicable: data quality scope, measure, evaluation procedure, result, value type, value unit, and date.

The standard does not define a minimal acceptable level of quality for geo-spatial data, nor does it specify how the quality should be measured.

#### 3.4.3.2 Geographic information - Quality evaluation procedures

ISO/TC 211:19114 (2001) provides a framework of procedures for evaluation and determination of geo-spatial dataset quality. The applied procedures must be consistent with the quality principles of ISO/TC 211:19113 (2001) (cf. Section 3.4.3.1).

The provided framework of ISO/TC 211:19114 (2001) serves also for the evaluation and reporting of geo-spatial data quality results. These can be provided as metadata or as a separate quality evaluation report. The process for evaluating data quality is presented in Fig. 20.

## Quality Assurance of OpenAddresses

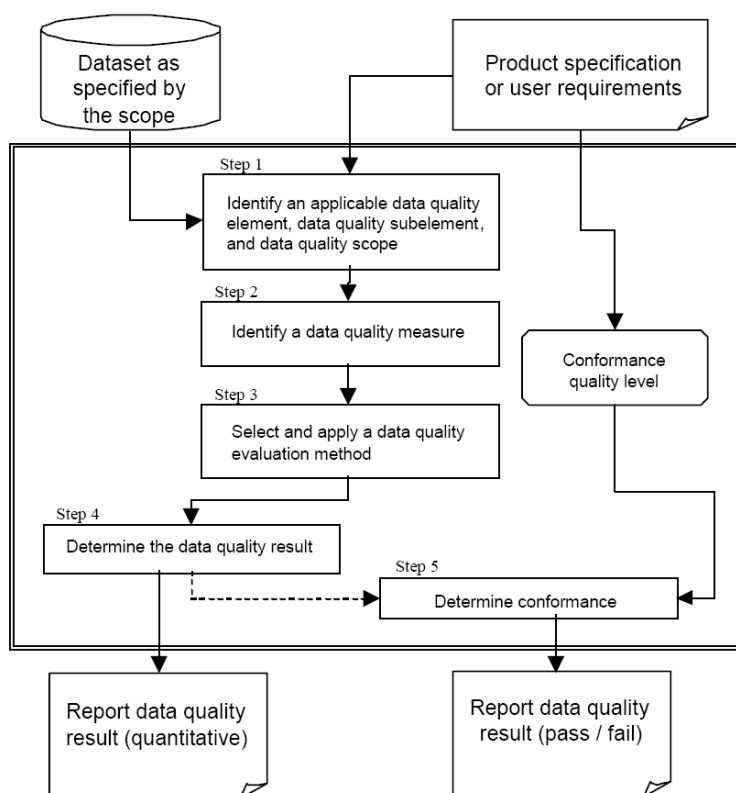


Fig. 20 Process to evaluate and report data quality results (source: ISO/TC 211:19114 (2001; p. 3))

ISO/TC 211:19114 (2001) divides data quality evaluation methods into two basic groups: direct and indirect, with direct methods subdivided into internal and external groups. To test positional accuracy it is necessary to use an external reference dataset or conduct a new survey (direct, external methods).

Indirect evaluation methods are based on external knowledge, and their application is recommended only if no direct evaluation method is usable. In the case of OA, in compliance with ISO/TC 211:19114 (2001), both direct, external and indirect methods are applied. In a first step, the selected OWMSs' attributes and spatial accuracy are evaluated against a reference dataset provided by the Canton of Solothurn (direct, external method) (cf. Section 4.3.1). This evaluation is applied to the entire reference dataset. In a second step, based on the results of the OWMS accuracy evaluation (indirect evaluation method), OA data are evaluated against the OWMS data (cf. Section 4.3.2). This evaluation is performed on each new or altered feature of the OA dataset.

### 3.4.3.3 Geographic information - Metadata

ISO/TC 211:19115 (2002) defines a schema for describing geo-spatial data and services. It provides information about the quality but also about the identification, the extent, the spatial and temporal schema, spatial references, and distribution of digital geo-spatial data. For ISO/TC 211:19113 (2001) and ISO/TC 211:19114 (2001), the UML schemas are part of the document. The data quality elements and subelements of ISO/TC 211:19113 (2001) are documented in data quality information packages. Fig. 21 shows a UML class diagram defining the classes of geographic information to which metadata apply.

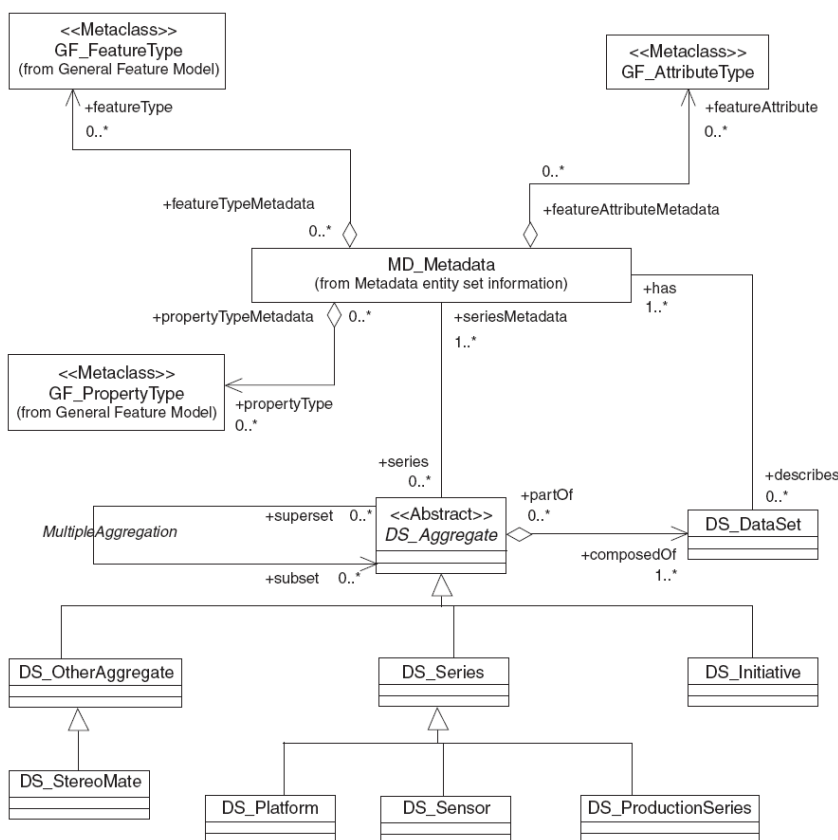


Fig. 21 UML diagram of metadata application (source: ISO/TC 211:19115 (2002; p. 9))

MD\_Metadata is an aggregate of several entities, including the optional DQ\_DataQuality entity, which contains information on data quality. DQ\_DataQuality is itself an aggregate of DQ\_Element and LI\_Lineage. DQ\_Element can be specified as DQ\_PositionalAccuracy, DQ\_ThematicAccuracy, DQ\_TemporalAccuracy, DQ\_Completeness and DQ\_LogicalConsistency (cf. Section 3.4.3.1).

## Quality Assurance of OpenAddresses

In Fig. 22 the metadata required to give a general assessment of a resource's quality are defined, while Fig. 23 shows the quality classes and subclasses used.

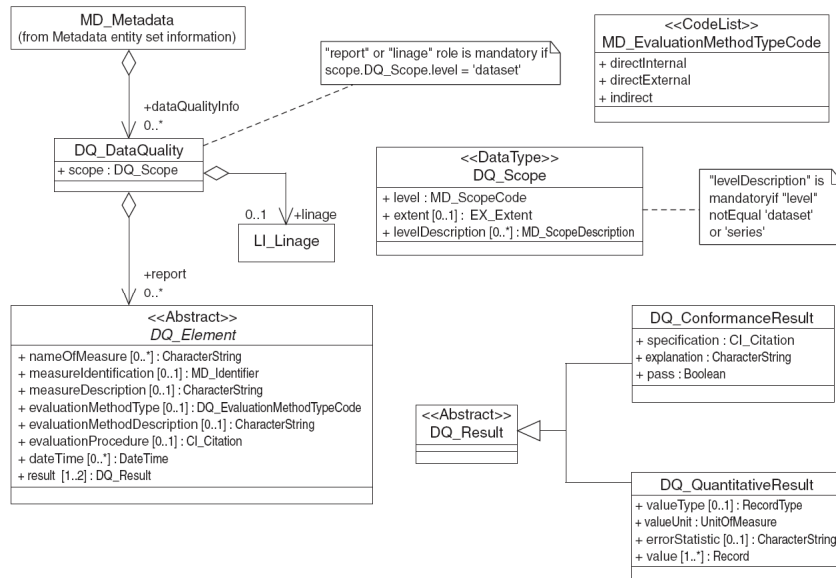


Fig. 22 UML diagram of quality information (source: ISO/TC 211:19115 (2002; p. 22))

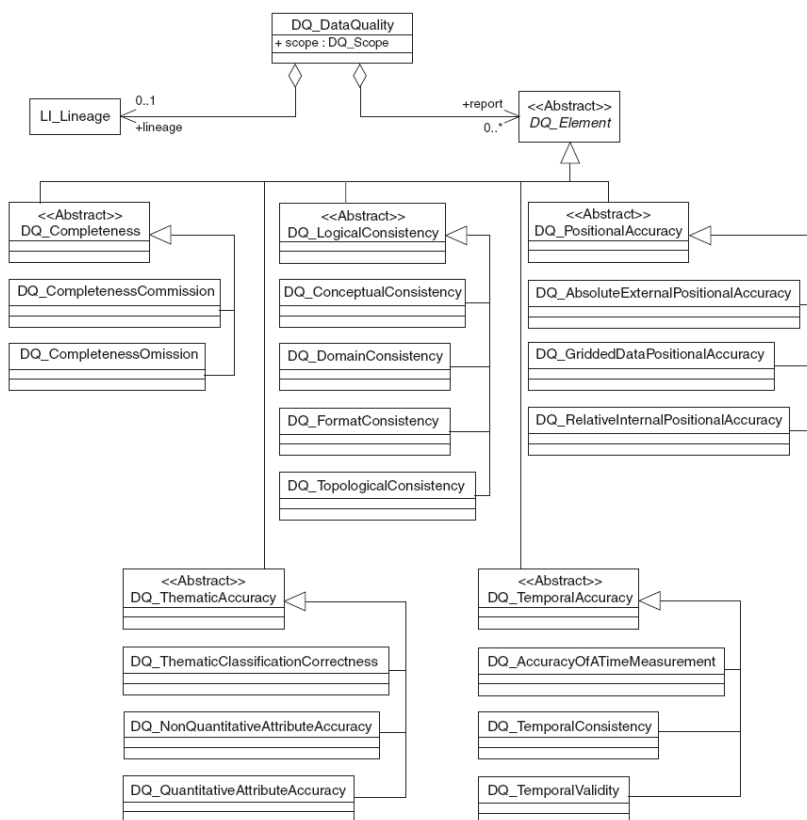


Fig. 23 UML diagram on data quality classes and subclasses (source: ISO/TC 211:19115 (2002; p. 24))

ISO/TC 211:19113 (2001) and (ISO/TC 211:19114 (2001) work together closely, while ISO/TC 211:19115 (2002) and ISO/TC 211:19139 (2007)<sup>34</sup> serve as Metadata-XML-schema implementation standards on how to structure the results of quality assessments in separate quality reports. Fig. 24 provides a schematic view of how ISO/TC 211:19113 (2001) is connected to ISO/TC 211:19114 (2001) and ISO/TC 211:19115 (2002).

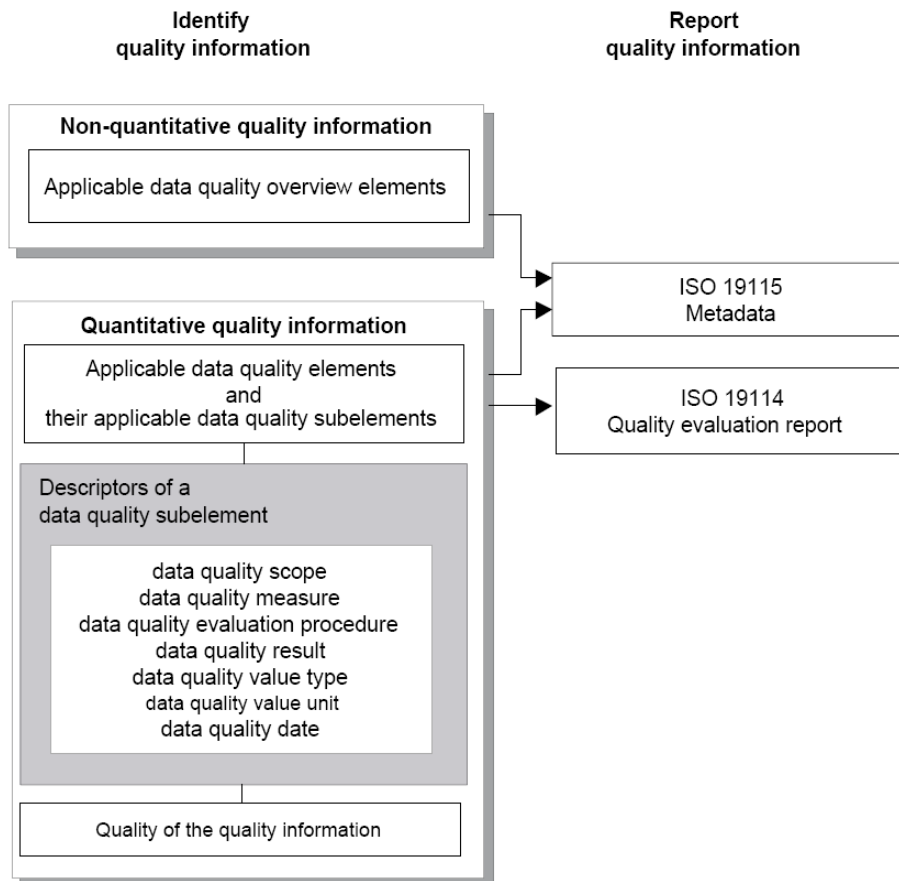


Fig. 24 Overview of data quality information of ISO/TC 211:19113 (normative) (source: ISO/TC 211:19113 (2001; p. 5))

Although reporting quality information according to ISO/TC 211:19115 (2002) is beyond the scope of this thesis, it is relevant for OA and thus introduced here. Quality assessment of OA data is currently handled mostly internally or via the web page introduced in Section 5.4. The feature-based modelling of quality assessment according to ISO/TC 211:19115 leads to a vast amount of data and complexity in the export file structure, calling for

<sup>34</sup> [http://www.iso.org/iso/catalogue\\_detail.htm?csnumber=32557](http://www.iso.org/iso/catalogue_detail.htm?csnumber=32557) [online April 16, 2010]

tools capable of handling such large amounts of data quickly and efficiently.<sup>35</sup> Therefore, at this stage of the project it is omitted. In the future, however, it may be possible to add quality information, if needed, to OA records in an export file. A sample of such a metadata report based on ISO/TC 211:19115 (2002) is provided on the attached CD-ROM.

### **3.4.3.4 Geographic information - Data quality measures**

ISO/TC 211:19138 (2006) defines, on the one hand, a set of data quality measures for reporting the subelements identified in ISO/TC 211:19113 (2001). On the other hand, its objective is to guide a data producer through the process of choosing the correct data quality measures for their reporting. This allows a data producer to add his own steps to the provided quality elements and measures. These data quality measures are applied in this thesis.

### **3.4.4 Quality Assurance with OpenAddresses in particular**

#### **3.4.4.1 General Considerations**

Mäs et al. (2005) suggests three constraints during mobile online data acquisition. Transforming these constraints to the OA project would mean:

- Every feature or object must have the *correct geometry type*. This condition is fulfilled with OA, since it supports the acquisition of addresses only as point objects.
- All necessary attributes and relations are considered. As OA uses JavaScript on the client side to check if all mandatory fields are completed when data is entered, this condition is also fulfilled.
- Attribute values must conform to the defined attribute data types. In OA this constraint is difficult to check, since all fields are type strings<sup>36</sup>.

---

<sup>35</sup> cf. <http://www.cprm.gov.br/33IGC/1344798.html> [online April 20, 2010]

<sup>36</sup> if OA is to store international addresses, zip codes must also be modelled as strings (cf. [http://www.freesearching.com/zip\\_codes\\_intl.htm](http://www.freesearching.com/zip_codes_intl.htm) and [http://en.wikipedia.org/wiki/Postal\\_code](http://en.wikipedia.org/wiki/Postal_code) [online March 26, 2010])

Based on Goodchild's quality definition of (cf. Section 3.4.1), OA's goal must be to collect, store and provide geocoded address data as accurately as possible regarding the geographic locations they represent. In other words, the semantic of an address - attribute values such as those introduced in Section 3.1.3 - must comply with the corresponding values of the real-world address. The position of an address in OA must lie at least within the ground view of the corresponding building. Both aspects - attribute accuracy and spatial accuracy - can only be evaluated with a reference dataset.

#### **3.4.4.2 Attribute Correctness**

Attribute correctness has a direct impact on a geocoder's reliability: the better the quality of the attribute values, the better the geocoder match rate, i.e., the better the overall positional accuracy of the geocoding entry (Goldberg et al. 2008b). When it comes to the assessment of attributes, OWMS could be used for comparison purposes: when an address is sent to an OWMS geocoder, both its spatial position and the details of the reference address used by the OWMS geocoder are transferred to the client (cf. Section 4.2). Currently, OWMS APIs can handle typographic errors and alternate spellings quite well<sup>37</sup>. Thus, using OWMS for comparison of address attribute values is possible and may help indicate whether a volunteer-entered address value is correct.

#### **3.4.4.3 Spatial Accuracy**

The situation regarding spatial accuracy assessment is slightly more difficult. Goldberg et al. (2008a) clearly show that the spatial accuracy of a geocoder's reference dataset is probably the most critical factor to the spatial accuracy of its output. Currently most OWMS systems use street geocoding data, the positions of which are essentially interpolated navigation data, and do not explicitly consider the exact location of buildings (cf. Fig. 35). Thus, the position derived from an OWMS can only be a (statistical) estimate of closeness to the position but can never be considered an exact reference.

---

<sup>37</sup> Example of different spelling with the same geocoding result of Google Maps API: 'Kasimir-Pfyffer-Strasse' vs. 'Kasimir Pfyffer Strasse' vs. 'Kasimir-Pfyffer Strasse' vs. 'Kasimir Pfyffer-Strasse' vs. 'Kasimir Pfyfer Strasse' vs. 'Kassimir Pfyffer Strasse' etc. in Lucerne

## Quality Assurance of OpenAddresses

According to Servigne et al. (2006), an OWMS derived address position is precise but inaccurate.

Fig. 25 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; some are oriented along the street axis, others are perpendicular to it; some buildings are detached, others share walls. Such characteristics have a direct impact on the quality of street geocoding results.



Fig. 25 Map excerpt from parcel map of City of Basel (source: <http://www.stadtplan.bs.ch/geoviewer> [online April 27, 2010])

The deviation between a street geocoding-derived location and its true location is smaller for a building close to the street axis than for a building that is detached, far removed from the street axis and not part of a homogeneous numbering pattern. The location for a building close to the street axis is placed approximately at the location that an algorithm will find when interpolating between known street number values

Since OWMS systems are open and freely available they are potentially useful as estimators. Although, as noted above, street geocoded addresses are subject to certain inherent problems, the following analysis tries to determine the feasibility of using them to estimate the spatial accuracy of OA addresses.

Both Haklay (2008) and Amelunxen (2009) show that completeness<sup>38</sup> in OSM has not yet been achieved. Likewise, although OA's long-term goal is to offer a complete dataset, it does not claim to do so at this stage. Thus, completeness is beyond the scope of the following analysis.

---

<sup>38</sup> cf. Maué, P. and S. Schade (2008). Quality of Geographic Information Patchworks. 11th AGILE International Conference on Geographic Information Science, Girona.

## 4 Analysis, Technology and Methodology

This chapter introduces the dataset, tools and technologies applied for the OA quality assessment. The implementation of the assessment process and its results are presented in Chapters 5 and 6.

### 4.1 Reference Dataset

The website of the bureau of the Canton of Solothurn that is responsible for the cadastre of spatial data (SOGIS) can be found at <http://www.so.ch/departemente/bau-und-justiz/sogis.html>. Under the Canton's specific license conditions,<sup>39</sup> it provides many datasets at no charge,<sup>40</sup> including geocoded building addresses (called 'Gebäudeadressen').<sup>41</sup> The metadata for each dataset are available when the dataset is chosen from a list (cf. Fig. 26).

SOIGIS Dateninventar:	
<b>Thema:</b>	AV - Gebäudeadressen
<b>Projekt:</b>	Rasche Aufnahme der Daten der amtlichen Vermessung - RADAV
<b>Beschreibung:</b>	Die Informationsebene Administrative Einteilungen ist ein Bestandteil des Grunddatensatzes nach Artikel 6 VAV. Das Thema "Gebäudeadressen" umfasst die Hausnummer, deren Lokalisation und Strassenzugehörigkeit.
<b>Datengrundlage:</b>	Technische Verordnung über die amtliche Vermessung (TVAV) vom 10. Juni 1994
<b>Erhebungsmethode:</b>	Private Vermessungsbüros
<b>Erfassungsmethode:</b>	Datenimport via AVS-Interlis
<b>Zeitstände (von - bis):</b>	28.04.2006 - 31.03.2010
<b>Maßstab:</b>	1:1
<b>Verbindlichkeit:</b>	Amtliche Vermessung (AV93)
<b>Bearbeitet durch:</b>	AGI / Stefan Ziegler
<b>Nachführung:</b>	laufend
<b>Datenherr:</b>	Amt für Geoinformation
<b>Auskunft fachlich:</b>	Ziegler, Stefan 032 627 75 96
<b>Auskunft GIS:</b>	Ziegler, Stefan 032 627 75 96
<b>Autor Doku:</b>	Ziegler, Stefan 032 627 75 96
<b>SIK Thema:</b>	Daten der amtlichen Vermessung
<b>Datentyp:</b>	PostGIS Layer (Punkt)
<b>Status:</b>	realisiert
<b>Abgabebedingungen:</b>	freie Datenabgabe über SOI ONLINE
<b>Abhängigkeiten:</b>	SOIGIS-Metadaten - 01.04.2010

Fig. 26 Metadata of 'Gebäudeadressen' (source: <http://www.sogis1.so.ch/sogis/OnLineData/php/datenbeschreibung.php?id=400454> [online April 1, 2010])

<sup>39</sup> cf. [http://www.sogis1.so.ch/sogis/OnLineData/etc/nutzungs\\_und\\_lizenzbedingungen.pdf](http://www.sogis1.so.ch/sogis/OnLineData/etc/nutzungs_und_lizenzbedingungen.pdf) [online April 1, 2010]

<sup>40</sup> cf. <http://www.so.ch/departemente/bau-und-justiz/sogis/sogis-daten.html> [online April 1, 2010]

<sup>41</sup> Download page: [http://www.sogis1.so.ch/sogis/OnLineData/php/datenbeschreibung\\_auswahl.php](http://www.sogis1.so.ch/sogis/OnLineData/php/datenbeschreibung_auswahl.php) [online April 1, 2010]

Geodata are generally described and delivered in INTERLIS<sup>42</sup> format but can also be ordered from SOGIS in ESRI format. The applied data model is the current Swiss Cadastre model,<sup>43</sup> referred to as 'DM.01-AV-CH, Version 24'. This model is divided into topics, containing classes, each with specific features. The 'Gebäudeadressen' topic contains the data of the geocoded addresses. Its entity relation diagram (ERD) is presented in Fig. 27.

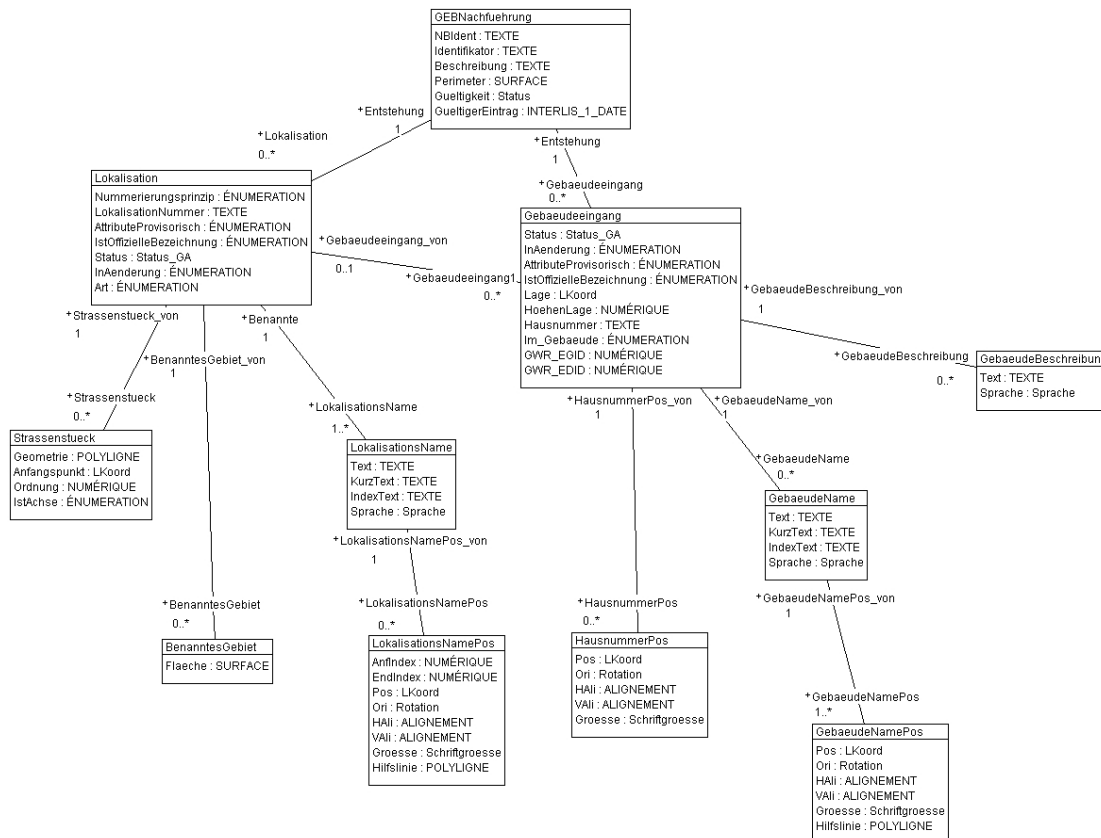


Fig. 27 ERD of topic 'Gebäudeadressen' (source:

<http://www.interlis.ch/mo2/diagramme.php?img=Bat&language=d&topic=Gebaeudeadressen> [online April 1, 2010])

The topic 'Gebäudeadressen' contains no information on zip codes or city names. This information is available under the heading of 'PLZOrtschaft' (cf. Fig. 28).

<sup>42</sup> cf. <http://www.interlis.ch> [online April 1, 2010]

<sup>43</sup> cf. <http://www.interlis.ch/mo> [online April 1, 2010]

## Quality Assurance of OpenAddresses

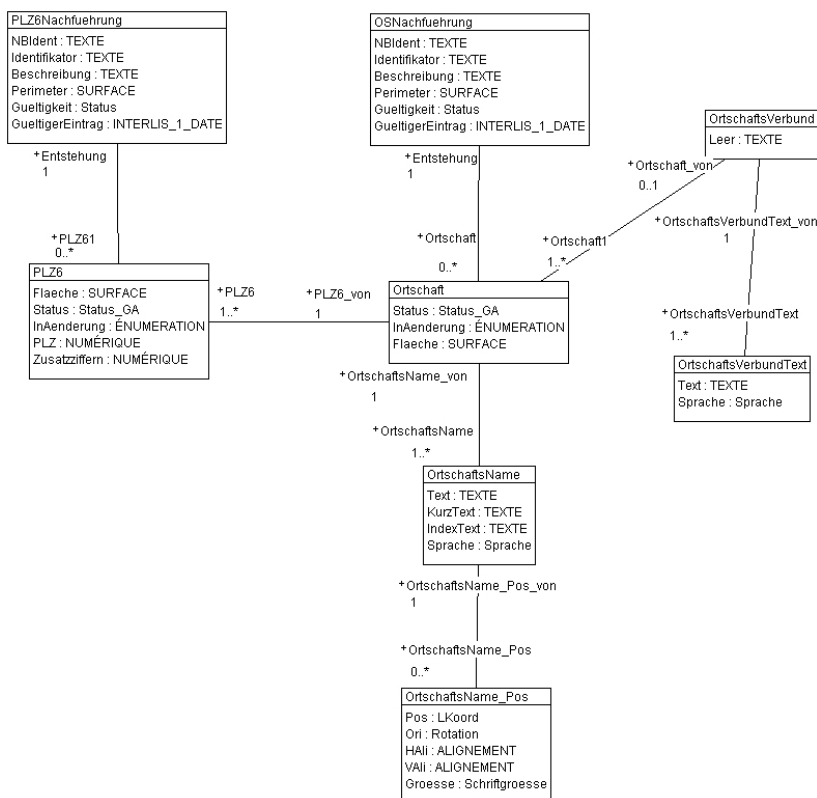


Fig. 28 ERD of 'PLZOrtschaft' topic (source: <http://www.interlis.ch/mo2/diagramme.php?img=Npal&language=d&topic=PLZOrtschaft> [online April 1, 2010])

### 4.2 Open Web Mapping Services' APIs

This section briefly introduces the application programming interfaces (APIs) of Microsoft Bing Maps<sup>44</sup>, Google Maps<sup>45</sup> and Yahoo! Maps.<sup>46</sup> APIs such as these have empowered a great many web mapping applications, also sometimes referred to as 'mashups'.<sup>47</sup>

"Mashups are web applications combining content and functionality from different online sources via publicly available interfaces (e.g., API, RSS<sup>48</sup>). This allows end-users to create new websites that dynamically combine services of existing providers." Novak and Voigt (2006; p. 1)

<sup>44</sup> <http://www.bing.com/maps> [online April 6, 2010]

<sup>45</sup> <http://maps.google.com> [online April 6, 2010]

<sup>46</sup> <http://maps.yahoo.com> [online April 6, 2010]

<sup>47</sup> [http://en.wikipedia.org/wiki/Mashup\\_%28digital%29](http://en.wikipedia.org/wiki/Mashup_%28digital%29) [online April 6, 2010]

<sup>48</sup> <http://liblearn.osu.edu/tutor/glossary.html#r> [online April 6, 2010]

OA is one example of such a mashup. Novak and Voigt (2006) found that more than a third of mashups are mapping applications. This figure is quite impressive and shows the popularity of web-based mapping, which has become possible thanks to APIs such as those introduced below.

### 4.2.1 Functionality

All three OWMSs discussed here provide APIs both to integrate their maps into web sites and to enable further processing of (geo-spatial) information or customisation of their maps. All three use AJAX<sup>49</sup> to provide their map data and further information in a way that the website does not have to be reloaded completely every time a user modifies it (e.g., zooming or panning a map). Only the part that has changed will be updated. Thus, individual elements of a web site can be changed and manipulated individually. This paradigm is very useful to create attractive and interactive web sites.

All three of these APIs also provide well documented interfaces with comprehensive functionality offering a range of actions to be taken by the client, including geocoding. A client - in this case a user working on a PC running a web-browser - sends address information to the API's server. The server handles the request; if a match is successfully created, the server returns both the location - as numerical latitude and longitude values - and complete address information: street name, house number, zip code and city name, along with geocoding quality data. This process is well documented. Brown (2006), Gibson and Erle (2006) and Purvis et al. (2006) focus on the Google Maps API, while Erle et al. (2005) also introduce other Mapping APIs, and Miller et al. (2009b) provides a compendium of web links to developer resources for several OWMS APIs. A comparison of functionality and data-sets used for each of the three OWMS used in this thesis is presented in Miller et al. (2009a). The researchers mentioned here helped create the automated geocoding system presented below in Section 5.3.

---

<sup>49</sup> <http://www.adaptivepath.com/ideas/essays/archives/000385.php> [online April 6, 2010]

## Quality Assurance of OpenAddresses

To assess the quality of the OWMS geocoders, their geocoding functionality is used and explored. Their individual uses are presented in the following three paragraphs.

Use of the OWMS APIs for Google Maps and Yahoo! Maps, requires an API key, i.e., an access code. Bing Maps does not require an API key at present.

### 4.2.2 Bing Maps

The API documentation for Microsoft Bing Maps is available at <http://www.microsoft.com/maps/isdk/ajax>. To geocode an address a map container must be inserted in an html (Hypertext Markup Language) page and JavaScript used to call the Microsoft Bing Maps server (cf. Fig. 29 and Fig. 30). Listing 1 shows a simple html page with a text field to input an address and to send the address string to the Bing Maps API geocoder.

```
<html>
<head>
  <title>Geocoding with Microsoft Bing Maps</title>
  <script type="text/javascript"
    src="http://dev.virtualearth.net/mapcontrol/mapcontrol.ashx?v=6.2">
  </script>
  <script type="text/javascript">
    //define map variable
    var map = null;

    function initiate(){
      //Necessary map definition and display -
      //otherwise the geocoder does not work
      map = new VEMap('myMap');
      map.LoadMap(new VELatLong(47, 8.3), 6 );
    }
    function startGeocoding() {
      //this function sends an address-string to the
      //bing maps server and receives the result of the geocoding
      //read address from text-field
      cur_address = document.getElementById('address').value;
      //send address-string to the map server
      map.Find(null,cur_address,null,null,0,20,false,false,false,false,
        function(layer, resultsArray, places,
          hasMore, veErrorMessage){
          //if geocoding was successful...
          if (places.length>0) {
            var place = places[0];
            //extract latitude and longitude values
            lat = place.LatLong.Latitude;
            lng = place.LatLong.Longitude;
            //update html-tags
            document.getElementById('long').innerHTML = lng;
            document.getElementById('lat').innerHTML = lat;
            //zoom to location
            map.SetCenterAndZoom(new VELatLong(lat, lng), 14);
          }
        }
      )
    }
  </script>
</head>
<body>
  <input type="text" id="address" value="<input type="text" id="address" value="</body>
</html>
```

```

</script>
</head>

<body onload="initiate()">
<h1>Bing maps</h1>
<div id='myMap' style="position:relative; width:450px; height:300px;"></div>
<table>
  <tr>
    <td>enter address:</td>
    <td>
      <input id="address" type="text" size="50" maxlength="50"
        value="Gründenstrasse 40, 4132 Muttenz, Switzerland">
    </td>
  </tr>
</tr>
<tr>
<td></td>
<td>
      <input type="button" value="geocode..."
        onclick="startGeocoding()" />
    </td>
  </tr>
</tr>
<tr>
<td>Longitude:</td>
<td id="long">-</td>
  </tr>
</tr>
<tr>
<td>Latitude:</td>
<td id="lat">-</td>
  </tr>
</tr>
</table>

</body>
</html>

```

Listing 1 HTML and JavaScript code for geocoding an address with Microsoft Bing Maps

**Bing maps**

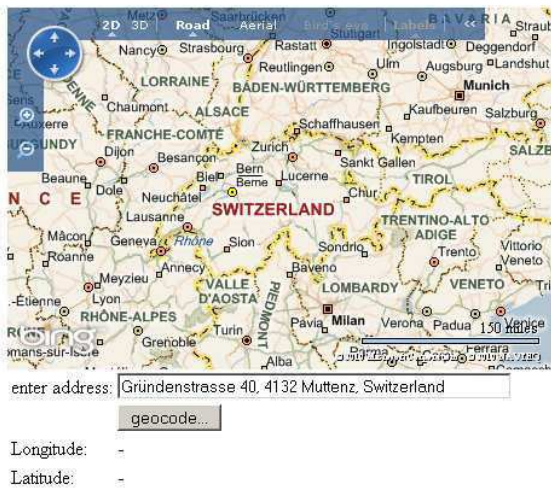


Fig. 29 Using Bing Maps API to geocode an address-string

**Bing maps**



Fig. 30 Result of the geocoding

### 4.2.3 Google Maps

The API of Google Maps is documented in detail at <http://code.google.com/intl/en-EN/apis/maps/documentation> (English version) and is available in various languages. The code for the same geocoding process as introduced in Section 4.2.2 is shown in Listing 2. Unlike Bing Maps, Google Maps does not require an embedded map container (cf. Fig. 31 and Fig. 32).

```
<html>
<head>
  <title>Geocoding with Google Maps</title>
  <script src=http://maps.google.com/maps?file=api&v=2.x&key=ABCD...
    type="text/javascript"></script>

  <script type="text/javascript">
    //define map variable
    var map = null;

    function initiate(){
      //Necessary map definition and display -
      //otherwise the geocoder does not work
      map = new VEMap('myMap');
      map.LoadMap(new VELatLong(47, 8.3), 6 );
    }

    function startGeocoding() {
      //this function sends an address-string to the
      //google maps server and receives the result of the geocoding
      //read address from text-field
      cur_address = document.getElementById('address').value;
      //send address-string to the map server
      var geocoder = new GClientGeocoder();
      geocoder.getLocations(cur_address,
        function getcoords(response) {
          place = response.Placemark[0];
          lng = place.Point.coordinates[0]
          lat = place.Point.coordinates[1];
          //update html-tags
          document.getElementById("lat").innerHTML = lat;
          document.getElementById("long").innerHTML = lng;
        }
      )
    }
  </script>
</head>

<body>
<h1>Google maps</h1>
<table>
  <tr>
    <td>enter address:</td>
    <td>
      <input id="address" type="text" size="50" maxlength="50"
        value="Gründenstrasse 40, 4132 MuttENZ, Switzerland">
    </td>
  </tr>
</table>
</body>
</html>
```

```

<td></td>
<td>
  <input type="button" value="geocode..."
    onclick="startGeocoding()" />
</td>
</tr>
<tr>
<td>Longitude:</td>
<td id="long">-</td>
</tr>
<tr>
<td>Latitude:</td>
<td id="lat">-</td>
</tr>
</table>

</body>
</html>

```

Listing 2 HTML and JavaScript code for geocoding an address with Google Maps

### Google maps

enter address:

Longitude: -

Latitude: -

Fig. 31 Using the Google Maps API to geocode an address string

### Google maps

enter address:

Longitude: 7.6385571

Latitude: 47.5339211

Fig. 32 Displaying coordinates after geocoding

## 4.2.4 Yahoo! Maps

The Yahoo! Maps API is documented at

<http://developer.yahoo.com/maps/ajax>. In order to geocode an address, Yahoo! Maps provides a special geocoding interface,<sup>50</sup> which uses a REST<sup>51</sup> like schema to return the coordinates of address strings. The result is an XML listing containing both the address information and the coordinates (cf. Fig. 33). Listing 3 shows the syntax of the call with the address information used in 4.2.3 and 4.2.4.

```

http://local.yahooapis.com/MapsService/V1/geocode?appid=ABCD...
&street=Gruendenstrasse+40
&zip=4132
&city=Muttenz
&country=Switzerland

```

Listing 3 Uniform Resource Locator (URL) for geocoding an address with Yahoo! Maps

<sup>50</sup> <http://developer.yahoo.com/maps/rest/V1/geocode.html> [online April 6, 2010]

<sup>51</sup> [http://bitworking.org/news/How\\_to\\_create\\_a\\_REST\\_Protocol](http://bitworking.org/news/How_to_create_a_REST_Protocol) [online April 6, 2010]

## Quality Assurance of OpenAddresses

```
-<ResultSet xsi:schemaLocation="urn:yahoo:maps http://api.local.yahoo.com/MapsService/V1/GeocodeResponse.xsd">
- <Result precision="address">
  <Latitude>47.533367</Latitude>
  <Longitude>7.638402</Longitude>
  <Address>Gründenstrasse 40</Address>
  <City>4132 Muttenz</City>
  <State>Switzerland</State>
  <Zip/>
  <Country>CH</Country>
</Result>
</ResultSet>
- <!--
ws11.ydn.ac4.yahoo.com compressed/chunked Tue Apr 6 08:08:45 EDT 2010
-->
```

Fig. 33 XML listing of the Yahoo! Maps API geocoder

### 4.2.5 Comparison of Open Web Map Services

Sections 4.2.2 to 4.2.4 above introduced three OWMS APIs. Their use is fairly similar but still differs on certain points. Comparing the result of the sample address sent to each of the OWMS geocoders shows that each returns a slightly different location result:

OWMS API	Longitude	Latitude
Microsoft Bing Maps	7.638601	47.533175
Google Maps	7.638557	47.533921
Yahoo! Maps	7.638402	47.533367

Table 2 Comparison of OWMS's geocoding results with one sample address

Table 2 raises the question of whether one of the OWMS used here is more accurate than the others, or whether the "true" location is a mean or a weighted mean of all three. However, it shows clearly that if OWMSs are applied for quality assessment of OA data, they must first be quality assessed themselves (cf. Section 5.3).

Visualising the returned coordinate values from Table 2 answers the question above: While Google Maps locates the address exactly on the building, both Bing Maps and Yahoo! Maps use street geocoding - Yahoo! Maps also applies a lateral offset (cf. Fig. 34).

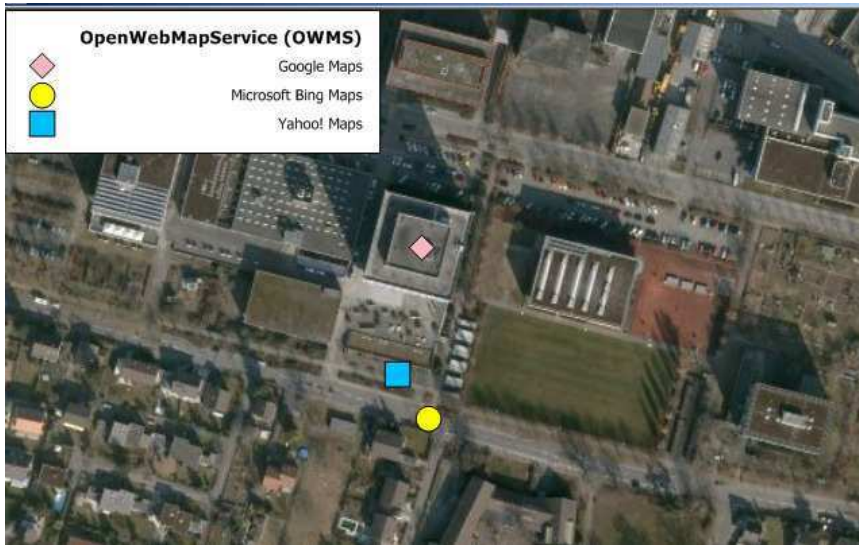


Fig. 34 Map excerpt showing OWMS derived locations of sample address in MuttENZ

In Switzerland, Google Maps uses the Swiss Post’s high-precision Geopost address dataset,<sup>52</sup> while Microsoft Bing Maps<sup>53</sup> and Yahoo! Maps<sup>54</sup> use NAVTEQ navigation data and interpolation techniques to provide coordinates for geocoding addresses. Table 3 presents the general data providers from international and general points of view according to Miller et al. (2009a; p. 74).

	Microsoft Bing Maps	Google Maps	Yahoo! Maps
Map data providers	MAPIT, TeleAtlas, DigitalGlobe, MDA Federal	NAVTEQ, TeleAtlas, i-cubed, Public domain	NAVTEQ, Intermap, Pictometry, NASA

Table 3 Listing of OWMS' map data providers

<sup>52</sup> [http://www.google.com/intl/en\\_ALL/help/legalnotices\\_maps.html](http://www.google.com/intl/en_ALL/help/legalnotices_maps.html) [online April 7, 2010]

<sup>53</sup> <http://www.bing.com/maps/Help/en-us/About.htm> and <http://msdn.microsoft.com/en-us/library/dd435699%28v=MSDN.10%29.aspx> [online April 7, 2010]

<sup>54</sup> <http://info.yahoo.com/legal/us/yahoo/maps/mapstou/mapstou-278.html> [online April 7, 2010]

## Quality Assurance of OpenAddresses

To emphasize the differences of street-geocoded locations versus their true locations Fig. 35<sup>55</sup> presents a number of sample addresses in Basel's Gellertstrasse (cf. Fig. 25), showing clearly that Google Maps provides the best spatial accuracy, with data very close to the true building locations.

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. Table 4 shows the computed error distances (deviations) – ranging from small (0.8m) to fairly large (252.5m) values – between the OWMS geocoder's locations and the true locations. In addition to showing that a fairly wide range of deviations is possible, this indicates that Google Maps has the narrowest range of deviations, followed by Yahoo! Maps and finally Bing Maps. This is plausible when the locations of geocoded addresses are considered in Fig. 35.

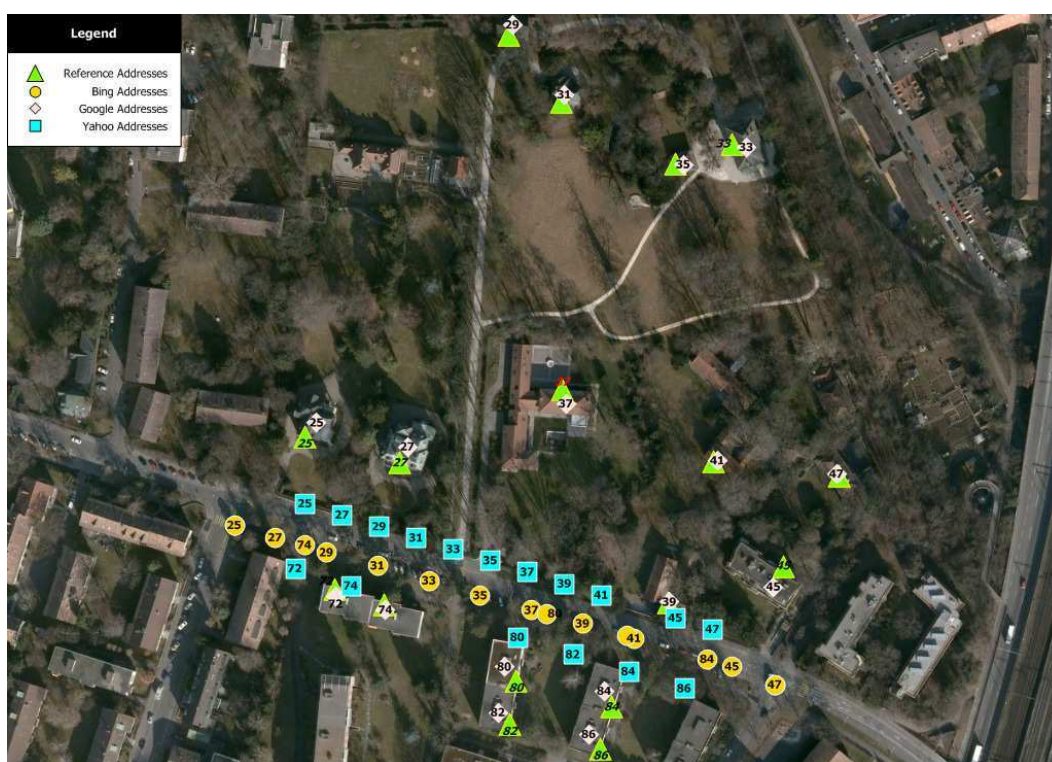


Fig. 35 Map excerpt showing OWMS derived locations versus true locations of addresses in Basel at Gellertstrasse

<sup>55</sup> The map was created in Manifold GIS Version 8.0 with Bing Maps as backdrop map

Test-address	Deviation Bing	Deviation Google	Deviation Yahoo
Gellertstrasse 25	51.9	8.6	30.7
Gellertstrasse 27	66.9	8.3	35.7
Gellertstrasse 29	252.5	4.7	233.9
Gellertstrasse 31	229.2	3.8	211.1
Gellertstrasse 33	244.5	6.8	226.4
Gellertstrasse 35	218.1	3.7	201.6
Gellertstrasse 37	102.5	6.8	85.6
Gellertstrasse 39	40.5	0.8	48.6
Gellertstrasse 41	89.2	1.7	80.3
Gellertstrasse 45	51.7	10.5	55.0
Gellertstrasse 47	100.6	1.4	91.7
Gellertstrasse 72	54.6	3.6	20.4
Gellertstrasse 74	45.7	2.6	17.6
Gellertstrasse 80	33.8	8.4	20.0
Gellertstrasse 82	67.6	7.6	43.4
Gellertstrasse 84	49.4	8.1	18.2
Gellertstrasse 86	85.6	8.4	47.4

Table 4 Distance comparison of OWMS geocoder and reference locations for test addresses

The deviation ranges within the individual OWMSs are:

Bing Maps: 33.8m to 252.5m

Google Maps: 0.8m to 10.5m

Yahoo! Maps: 17.6m to 233.9m

### **4.3 Applied Quality Assurance**

Section 3.4.3 introduced several ISO/TC 211 19100 series standards. This section explains how these standards are applied for quality assessment of OA data.

In order to assess the quality of OA data using OWMS, two steps are necessary:

1. Assessment of OWMS quality
2. Assessment of OA data based on findings of OWMS quality assessment

As presented in Sections 4.3.1 and 4.3.2 below, these quality assessments require similar but individual approaches.

#### **4.3.1 Assessment of Open Web Map Services' quality**

As stated in Section 3.4.3.2, OWMS assessments were conducted using direct, external methods. Since all three OWMSs provide data globally but only a local reference dataset was available, however, the results of the OWMS quality assessments cannot be regarded as comprehensive and valid for the entire OWMS data range. However, it can be regarded as representative for Switzerland, as each of the three OWMSs uses the same dataset throughout Switzerland. Annex E of ISO/TC 211:19114 (2001) refers to this spatial sampling method as *area-guided sampling of a predefined area*.

In terms of spatial accuracy, one user requirement is that the position of the location returned by the OWMS must be within the ground view of the building it represents. To judge whether a location's coordinates indicate a point within the same building as indicated by the reference dataset, Ahlers and Boll (2008) define the tolerance distance as 20m. Because both Bing Maps and Yahoo! Maps' datasets are based on street geocoding, neither fulfils this requirement consistently for the sample given (cf. Table 4).

According to Annex C of ISO/TC 211:19138 (2006), basic data quality measures for dealing with uncertainty are based on three assumptions, one of which one is not fulfilled with the assessment of OWMS. This assumption is that "uncertainties are homogeneous for all observed values" (ISO/TC

211:19138 2006; p. 16). Hence a data quality measure is applied that does not fully comply with the prescribed standard, i.e., rather than treating uncertainties as homogeneous, it works with a mean value of positional uncertainties excluding outliers (ISO/TC 211:19138 2006; p. 42). To facilitate understanding and discussion, this mean value will be referred to below as root mean square error (RMSE). The data quality scope is the entire dataset (cf. ISO/TC 211:19113 (2001)).

As shown in Fig. 20, step 1 in the process of data quality evaluation involves the identification of data quality elements and subelements. Since the quality assessment has no specific user requirements, this parameter has no influence on the evaluation process. The identification of the data quality elements and subelements for quality assessment of OWMS are presented in Table 5. They are adapted from samples provided by National Mapping and Cadastral Agencies (Jakobsson and Giversen 2007) and are based on ISO/TC 211:19113 (2001).

Data Quality Element	Data Quality Subelement	Relevant?
Completeness	commission (Feature completeness)	No
	omission (Feature completeness)	No
	commission (Attribute completeness)	Partly
	omission (Attribute completeness)	Partly
Logical consistency	conceptual consistency	No
	domain consistency	No
	format consistency	No
	topological consistency	No
Positional accuracy	absolute or external accuracy	Yes
	relative or internal accuracy	No
	gridded data position accuracy	No
Temporal accuracy	accuracy of a time measurement	No
	temporal consistency	No
	temporal validity	No
Thematic accuracy	classification correctness	No
	non-quantitative attribute correctness	Partly
	quantitative attribute accuracy	No

Table 5 Overview of data quality elements and subelements for quality evaluation of OWMS according to ISO/TC 211:19113 (2001; p. 31)

## Quality Assurance of OpenAddresses

Step 2 in the process of data quality evaluation identifies the quality measures to be used, while in step 3 the data quality evaluation methods are defined and applied. Table 6 presents both quality measures and methods for each data quality subelement identified in Table 5.

<b>Data Quality Subelement</b>	<b>Quality Measure</b>	<b>Quality Method</b>
commission (Attribute completeness)	Pass-Fail	Check whether OWMS returns complete address information and count the number of records of completely returned values for 'street name & house number', 'zip code' and 'city name'.
omission (Attribute completeness)	Pass-Fail	Check whether OWMS returns complete address information and count the number of records for which the OWMS did not return a value for either 'street name & house number', 'zip code' or 'city name'.
absolute or external accuracy	RMSE mean value of positional uncertainties excluding outliers (2D); identifier 29 (ISO/TC 211:19138 (2006; p. 42))	For each record (address) the error distance or deviation is computed as a Euclidean distance between the location of the point indicated by the reference dataset and the location returned by the OWMS. Finally the mean value of positional uncertainties is computed as the overall average error distance (excluding outliers) for each OWMS.

non-quantitative attribute correctness	Pass-Fail	Compare whether OWMS address information is equal to the address information of the reference dataset. The comparison is applied to the joined string of 'street name & house number', 'zip code' & 'city name' for each OWMS.
--	-----------	--

Table 6 Quality evaluation of OWMS: Quality measures and methods for relevant data quality subelements

Step 4 determines the result of the quality evaluation process, which will be presented in Chapter 6. The determination of conformance defined in step 5 cannot be applied because specific user requirements (e.g. positional accuracy in metric values) are missing.

The evaluation and the quality assessment reports are valid for each OWMS as a whole, and not for individual features (addresses) of an OWMS.

#### 4.3.2 Assessment of OpenAddresses' quality

The procedure presented in detail in Section 4.3.1 is also applied for the quality assessment of OA data. OA is a dynamic dataset, i.e., it changes continuously as either new addresses are entered into the database or existing ones are altered or deleted. According to ISO/TC 211:19114 (2001; p. 11) "only indirect or internal direct methods may be applied" for dynamic datasets. This fact implicitly defines the scope of the dataset: unlike with OWMSs, where the scope of quality assessment is the entire dataset, in OA it is the individual feature. This also implies that no sampling is applied but that quality assessment is applied to each new feature-instance (address) stored or manipulated in the OA database. Thus, for dynamic datasets, ISO/TC 211:19114 makes quality assessment part of the data-collection process.

In this quality assessment process, nearly the same data quality subelements are used as listed in Section 4.3.1. Only the data quality element 'completeness' and its subelements are omitted, as OA defines the entry of all address fields as mandatory. Table 7 shows data the quality elements and

## Quality Assurance of OpenAddresses

subelements for quality evaluation of OA. Corresponding quality measures and methods are presented in Table 8.

Data Quality Element	Data Quality Subelement	Relevant?
Logical consistency	conceptual consistency	No
	domain consistency	No
	format consistency	No
	topological consistency	No
Positional accuracy	absolute or external accuracy	Yes
	relative or internal accuracy	No
	gridded data position accuracy	No
Temporal accuracy	accuracy of time measurement	No
	temporal consistency	No
	temporal validity	No
Thematic accuracy	classification correctness	No
	non-quantitative attribute correctness	Partly
	quantitative attribute accuracy	No

Table 7 Data quality elements and subelements for quality evaluation of OA

Data Quality Subelement	Quality Measure	Quality Method
absolute or external accuracy	Error Distance	For each new or altered record (address) in OA the error distance or deviation is computed as a Euclidean distance between that position and the corresponding positions in all three OWMSs (cf. Section 4.2).
non-quantitative attribute correctness	Pass-Fail	Compare whether address information of a new or altered address is identical with non-geomatic address information returned by all three OWMS. This test is performed individually on 'street name & house number', 'zip code' and 'city name' for each OWMS.

Table 8 Quality evaluation of OA: Quality measures and methods for relevant data quality subelements

Quality assessment reporting is at present handled internally only. It is stored in a separate table in the database and is presented using a dynamic webpage (cf. Fig. 42), which also includes a colour-coded classification based on threshold values resulting from the OWMS quality assessments. These threshold values are used to judge the entered addresses as records of good quality, records that need further investigation, or outliers.

Chapter 4 introduced both the reference data and the methods and measures used to assess the quality of OWMSs and also OAs. The implementation of these concepts is presented in chapter 5.

## 5 Implementation of Quality Assessment

### 5.1 Development Environment

Both quality evaluation procedures (cf. Sections 4.3.1 and 4.3.2) involve the following technical components:

Client side: HTML and JavaScript

Server side: PHP (Hypertext Preprocessor) and  
PostgreSQL/PostGIS

The communication between client and server is via http. AJAX provides a dynamic data exchange between the two.

Both JavaScript and PHP are scripting languages. Thus a simple text editor is sufficient for the coding. The sample data is stored in a PostgreSQL database, with the additional PostGIS module used to handle spatial data.

### 5.2 Preparing the Reference Dataset

In order to realise the address structure as introduced above in Section 3.1.3, the spatial objects of address locations and the regions of zip codes for the reference dataset of the Canton of Solothurn must first be combined using GIS functionality.

When the data was downloaded for analysis for this thesis, only the data of topic 'Gebaeudeadressen' (cf. Section 4.1 and Fig. 27) was completely available and fully updated, while an update of the data for 'PLZOrtschaft' (cf. Section 4.1 and Fig. 28) was underway. Thus Swiss Post data were used. The matching of the data from different sources into the address structure of Section 3.1.3 was created using Safe Software's FME<sup>56</sup> (File Manipulation Engine) software package.

The features of 'Gebaeudeadressen' are points, and are provided with an individual number that will later serve as object-identifier. The features of

---

<sup>56</sup> <http://www.safe.com/> [online April 1, 2010]

'PLZOrt' are regions and had to be completed using FME's 'joiner' function. After a spatial overlay and a re-projection, the data is complete and ready to be loaded into PostgreSQL/PostGIS, where it is stored for further processing. The re-projection into the WGS-84 system is applied because all three OWMS geocoders return their geocoding results as WGS-84 coordinate values. The process is presented in Fig. 36.

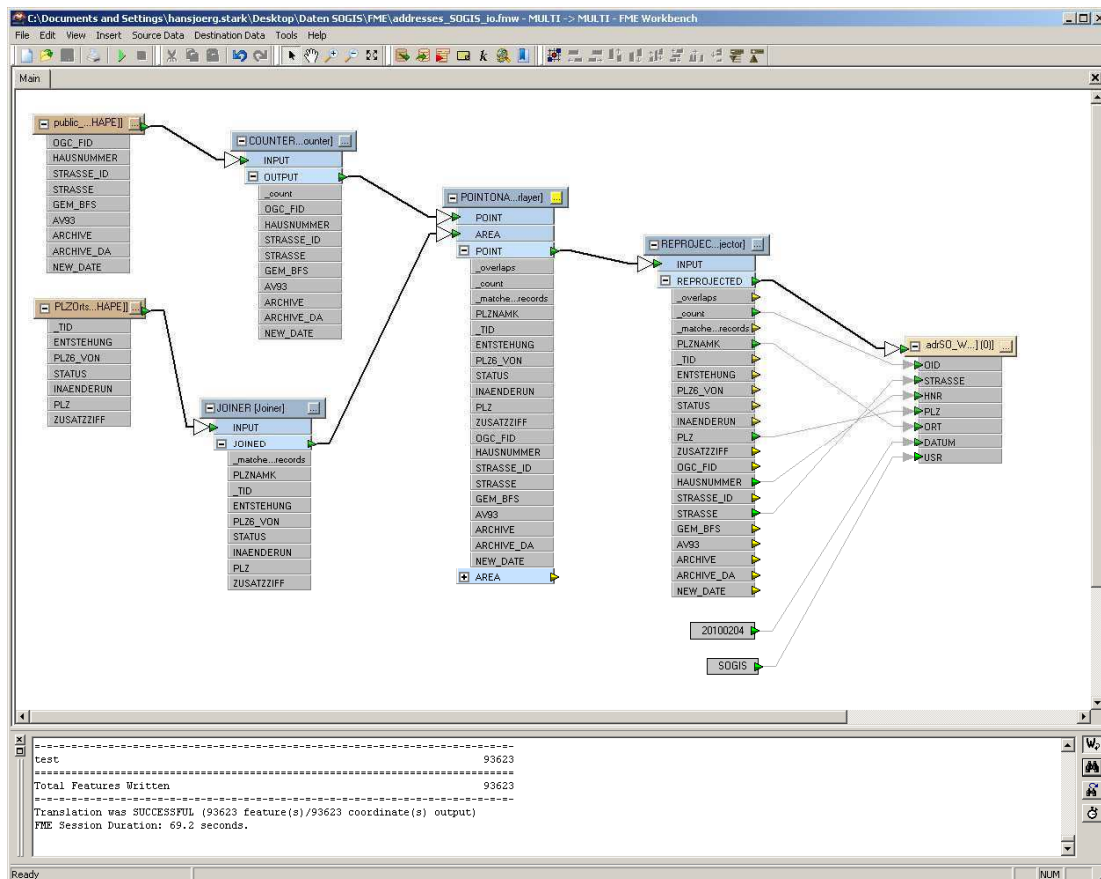


Fig. 36 Creating address data structures in FME

The total number of records is 93,623. Fig. 37 shows the spatial distribution of the data.

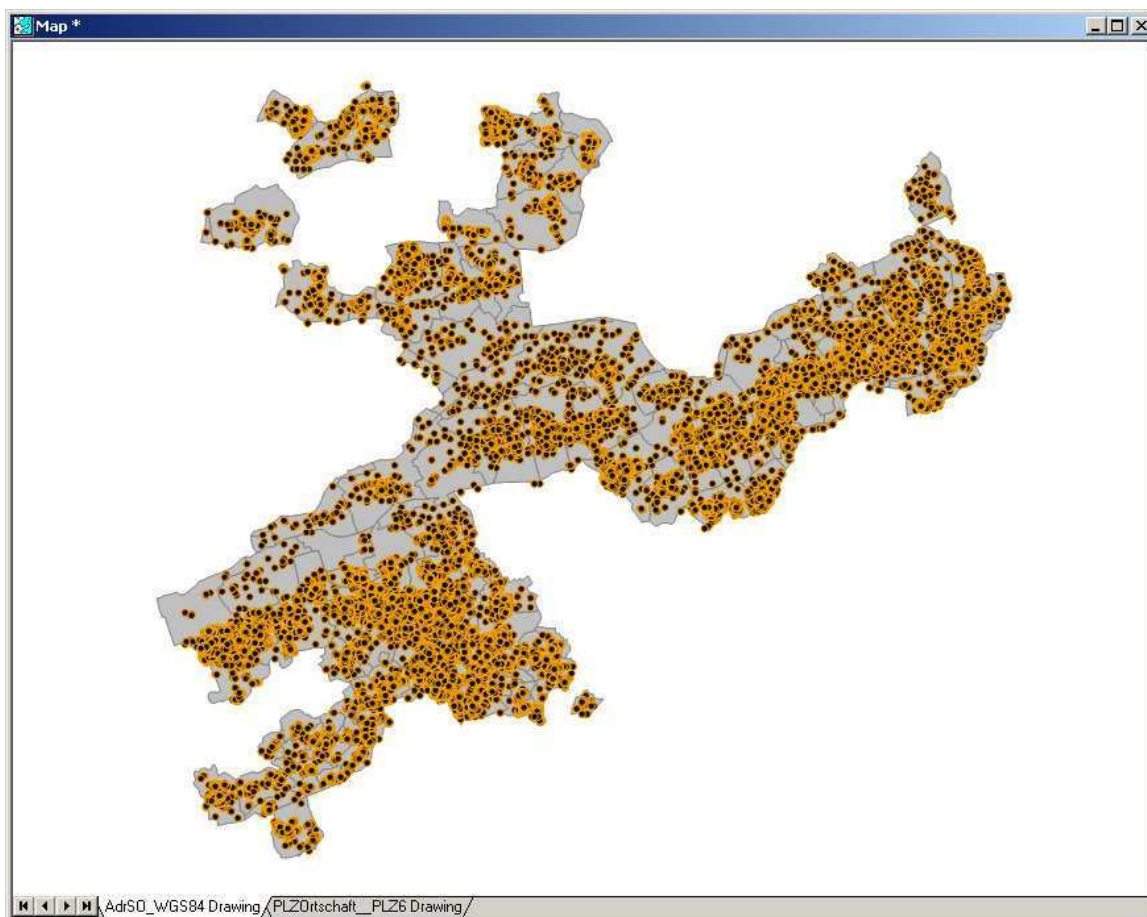


Fig. 37 Map view of Reference data of the Canton of Solothurn

### ***5.3 Implementing Quality Assessment of Open Web Map Services***

In order to evaluate the quality of OWMS geocoding results the complete sample dataset of Solothurn is geocoded by each of the three OWMSs, which requires the development of a batch geocoding procedure for each OWMS. For each OWMS an additional table, listing that OWMS's geocoder results for each record (address), is created in the PostgreSQL database. The structure of these tables is identical for all OWMSs. Besides the standard address information, the positions as geometric objects and information on the quality of the geocoding process for each record is stored. Fig. 38 presents an example structure and part of the data within the `bing_geocoding` table. These data are the basic material for the OWMS geocoders' quality evaluations. The process of batch geocoding is presented in Fig. 39.

## 5 Implementation of Quality Assessment

	oid integer	oid_orig integer	street character var	zip character var	city character var	coordinates geometry	warning character var	geocodingprecision character varying(50)	datum character(19)
1	535481	1	Zwinglistrasse 2	2540	Grenchen	0101000020E61	"	Good / High / Interpolated	20100204-042552
2	535482	2	Zwingliweg 14	4562	Biberist	0101000020E61	"	Good / High / Interpolated	20100204-042553
3	535484	3	Zwingliweg 13	4562	Biberist	0101000020E61	"	Good / High / Interpolated	20100204-042554
4	535486	4	Zwingliweg 9	4562	Biberist	0101000020E61	"	Good / High / Interpolated	20100204-042554
5	535487	5	Zwingliweg 3	4562	Biberist	0101000020E61	"	Good / High / Interpolated	20100204-042556
6	535488	6	Zwingliweg 12	4562	Biberist	0101000020E61	"	Good / High / Interpolated	20100204-042556
7	535489	7	Zwingliweg 5	4562	Biberist	0101000020E61	"	Good / High / Interpolated	20100204-042557
8	535490	8	Zwingliweg 5	4562	Biberist	0101000020E61	"	Good / High / Interpolated	20100204-042558
9	535491	9	Zwingliweg 7	4562	Biberist	0101000020E61	"	Good / High / Interpolated	20100204-042559
10	535492	10	Zwingliweg 3	4562	Biberist	0101000020E61	"	Good / High / Interpolated	20100204-042559
11	535540	11	Zurmattenstrass	4500	Solothurn	0101000020E61	"	/ Medium / Interpolated	20100204-042907
12	535542	12	Zurmattenstrass	4500	Solothurn	0101000020E61	"	Good / High / Interpolated	20100204-042907
13	535543	13	Zurmattenstrass	4500	Solothurn	0101000020E61	"	Good / High / Interpolated	20100204-042908
14	535545	14	Zurmattenstrass	4500	Solothurn	0101000020E61	"	/ Medium / Interpolated	20100204-042909
15	535546	15	Zurmattenstrass	4500	Solothurn	0101000020E61	"	/ Medium / Interpolated	20100204-042910

Fig. 38 Table with results of OWMS geocoding

Metadata regarding these evaluations are not reported directly to the OWMSs because the quality assessment results are only valid for the area of investigation and further for Switzerland alone (i.e., not globally). They are presented below in Section 6.1.

## Quality Assurance of OpenAddresses

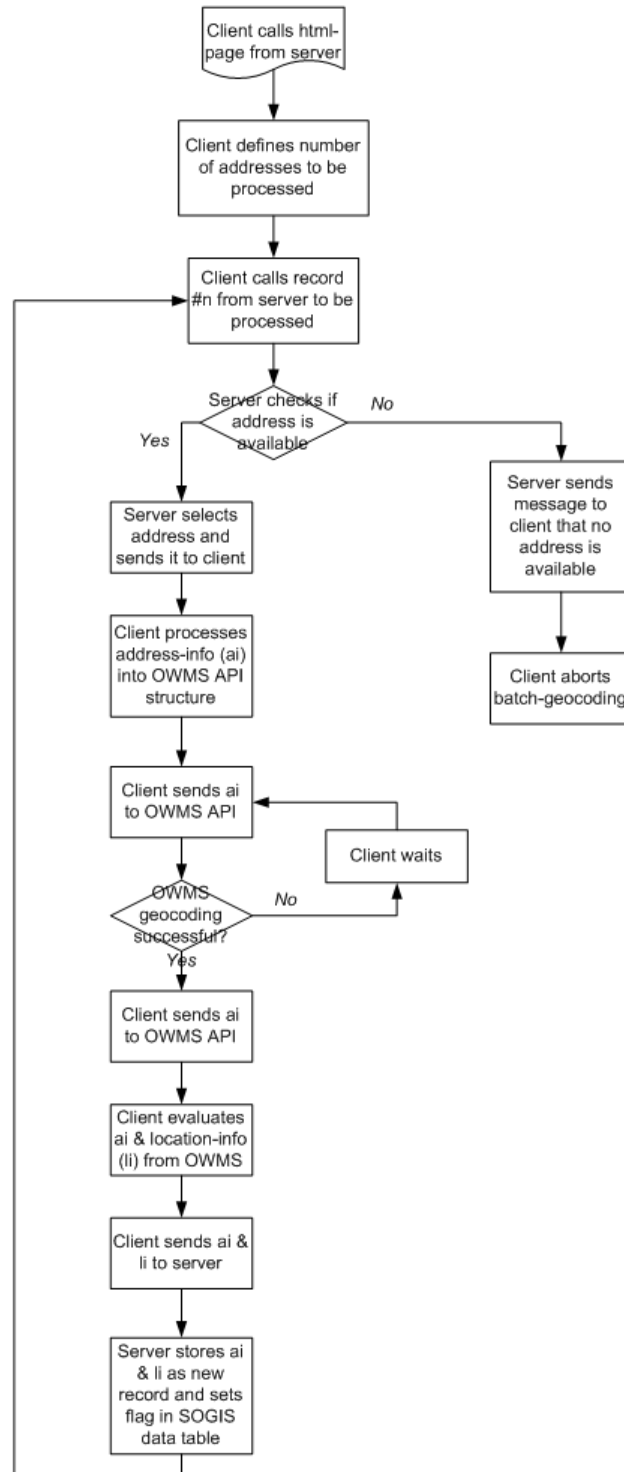


Fig. 39 Flow chart of batch geocoding process for OWMS quality evaluation

Fig. 40 shows the general data flow of the batch geocoding process and lists files created for the OWMS geocoding assessments. The colour coding of the files indicates which files belong together: Black file-names indicate that all

OWMSs use these files. Unlike Bing Maps and Google Maps, which use calls from JavaScript for geocoding, Yahoo! Maps geocoding is initiated via PHP calls. Thus the server files the geocoding requests instead of the client. Table 9 lists all files along with a brief description.

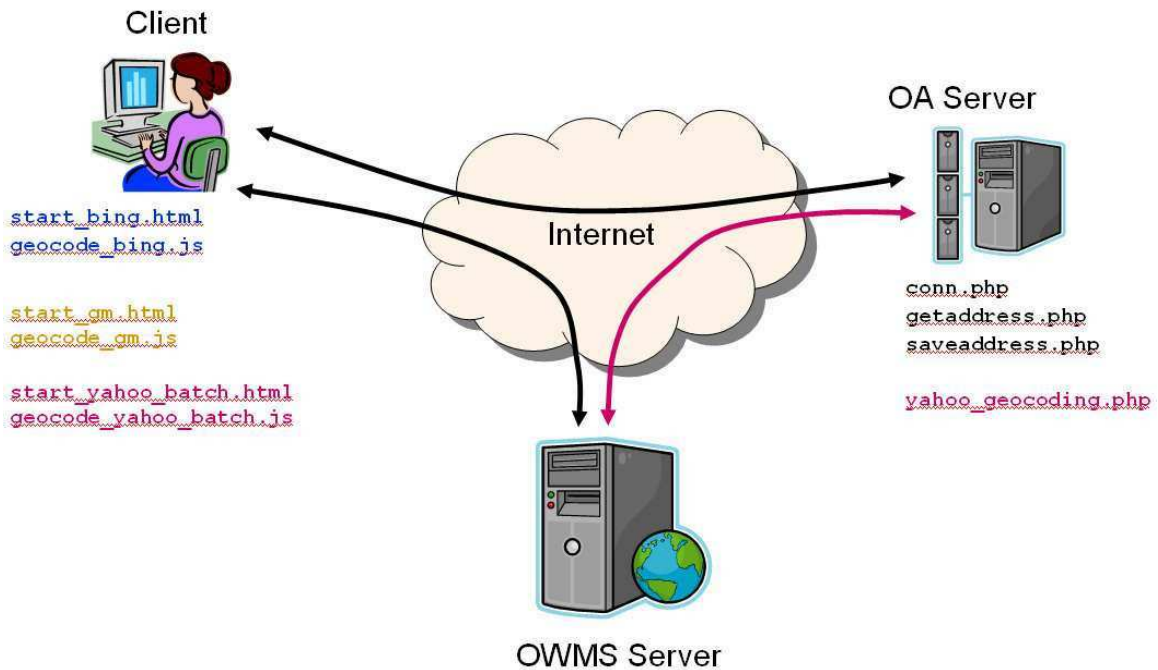


Fig. 40 Data flow of OWMS geocoding quality assessment

File-name	Description
<code>start_bing.html</code> <code>start_gm.html</code> <code>start_yahoo_batch.html</code>	These files are called to initiate the batch geocoding process.
<code>geocode_bing.js</code> <code>geocode_gm.js</code> <code>geocode_yahoo_batch.js</code>	These files provide the functionality of the batch-geocoding process. They call the server to extract an un-geocoded address from the database and forward it to the OWMS geocoder. The result returned by the OWMS server is processed and sent back to the server to be stored along with the OWMS geocoding information.

## Quality Assurance of OpenAddresses

<code>conn.php</code>	Establishes a connection with the database via which OA records and OWMS geocoding replies are stored.
<code>getaddress.php</code>	Fetches an address to be sent to the client.
<code>saveaddress.php</code>	Saves all information returned by the OWMS geocoder and processed by the client for storage in the database.
<code>yahoo_geocoding.php</code>	Yahoo! Maps geocoding is called from the server. The results are sent to the client, processed and sent back to the server for storage ( <code>saveaddress.php</code> ).

**Table 9 Overview and description of required files for OWMS batch geocoding process**

All OWMS APIs limit the number of geocoding requests each user can make per day.<sup>57</sup> Therefore, the process must be designed so that a user can enter a number of addresses to be processed at one time. Further, all three APIs use different syntax, meaning separate processes had to be designed and coded for each. The complete code listings of all three batch geocoding procedures can be found on the attached CD-ROM. Two sample figures of the geocoding results are presented in Appendix A.1.

### ***5.4 Implementing Quality Assessment of OpenAddresses***

The comparison of OA records with those of OWMS uses the fields street, house number, zip code and city name. The additional address information is purposely omitted because it normally contains a description of a building that cannot be evaluated and processed by OWMS.

The implementation of the OA quality assessment requires several additions to both the existing code and the database.

---

<sup>57</sup> Microsoft Bing Maps' MapPoint API limits the number of Geocoding requests to 10,000 per day (<http://www.bing.com/community/blogs/maps/archive/2008/07/28/overcoming-pushpin-limitations-in-mappoint-2009.aspx>), Google Maps 15,000 ([http://code.google.com/intl/de-DE/apis/maps/faq.html#geocoder\\_limit](http://code.google.com/intl/de-DE/apis/maps/faq.html#geocoder_limit)) and Yahoo! Maps 5,000 (<http://developer.yahoo.com/maps/rest/V1/geocode.html>)

In the `index.php` file, which is loaded immediately when the application is called, a map container for Microsoft Bing Maps has to be integrated in order to be able to use Bing Maps geocoding.

The `map_functions.js` file requires several changes: the `storeAddress()` function is expanded with one command line that calls the mechanism for OWMS value comparisons when a new address is collected. When an existing address is manipulated - regarding either attribute or positional values - it must first be determined whether information already exists on this address in table `qaOA`. If no such record is found a dataset must first be created. If a record exists, it must be updated according to the changes made. Thus an additional command line is entered in the `editAddress()` and `shiftAddress()` functions. If an address is deleted its status is set to 'delete' in table `qaOA`, which stores information on the OWMS based quality evaluation.

On the server side, a new `qaOA` table has to be created to store the values of the OWMS comparisons.<sup>58</sup> The relevant syntax is presented in Listing 4. The entire process is presented in Fig. 41.

```
CREATE SEQUENCE qaOA_seq
  INCREMENT 1
  MINVALUE 1
  MAXVALUE 9223372036854775807
  START 1
  CACHE 1;
ALTER TABLE qaOA_seq OWNER TO postgres;

CREATE TABLE qaOA
(
  oid bigint,
  bing_dist double precision,
  bing_addr boolean,
  bing_zip boolean,
  bing_city boolean,
  google_dist double precision,
  google_addr boolean,
  google_zip boolean,
  google_city boolean,
  yahoo_dist double precision,
  yahoo_addr boolean,
  yahoo_zip boolean,
  yahoo_city boolean,
  qa_oid bigint NOT NULL DEFAULT nextval('qaOA_seq'::regclass),
```

<sup>58</sup> The coding of these commands, as well as of the sequence and trigger were created and applied by the database manager of the unit responsible for database maintenance

## Quality Assurance of OpenAddresses

```
bing_precision boolean,  
google_precision boolean,  
yahoo_precision boolean,  
type character(10),  
date character(20)  
)  
WITH (oids=FALSE);  
ALTER TABLE qaOA OWNER TO postgres;  
  
CREATE TRIGGER f_add_qaOA  
AFTER INSERT  
ON addresses  
FOR EACH ROW  
EXECUTE PROCEDURE hj_add_qaOA();  
  
CREATE OR REPLACE FUNCTION hj_add_qaOA()  
RETURNS trigger AS  
$BODY$  
BEGIN  
    INSERT INTO qaOA  
    (oid)  
    VALUES  
    (  
        NEW.oid  
    );  
    RETURN NULL;  
END;  
  
$BODY$  
LANGUAGE 'plpgsql' VOLATILE  
COST 100;  
ALTER FUNCTION hj_add_qaOA() OWNER TO postgres;
```

**Listing 4 SQL Syntax to create the table that contains the values of the OWMS comparison**

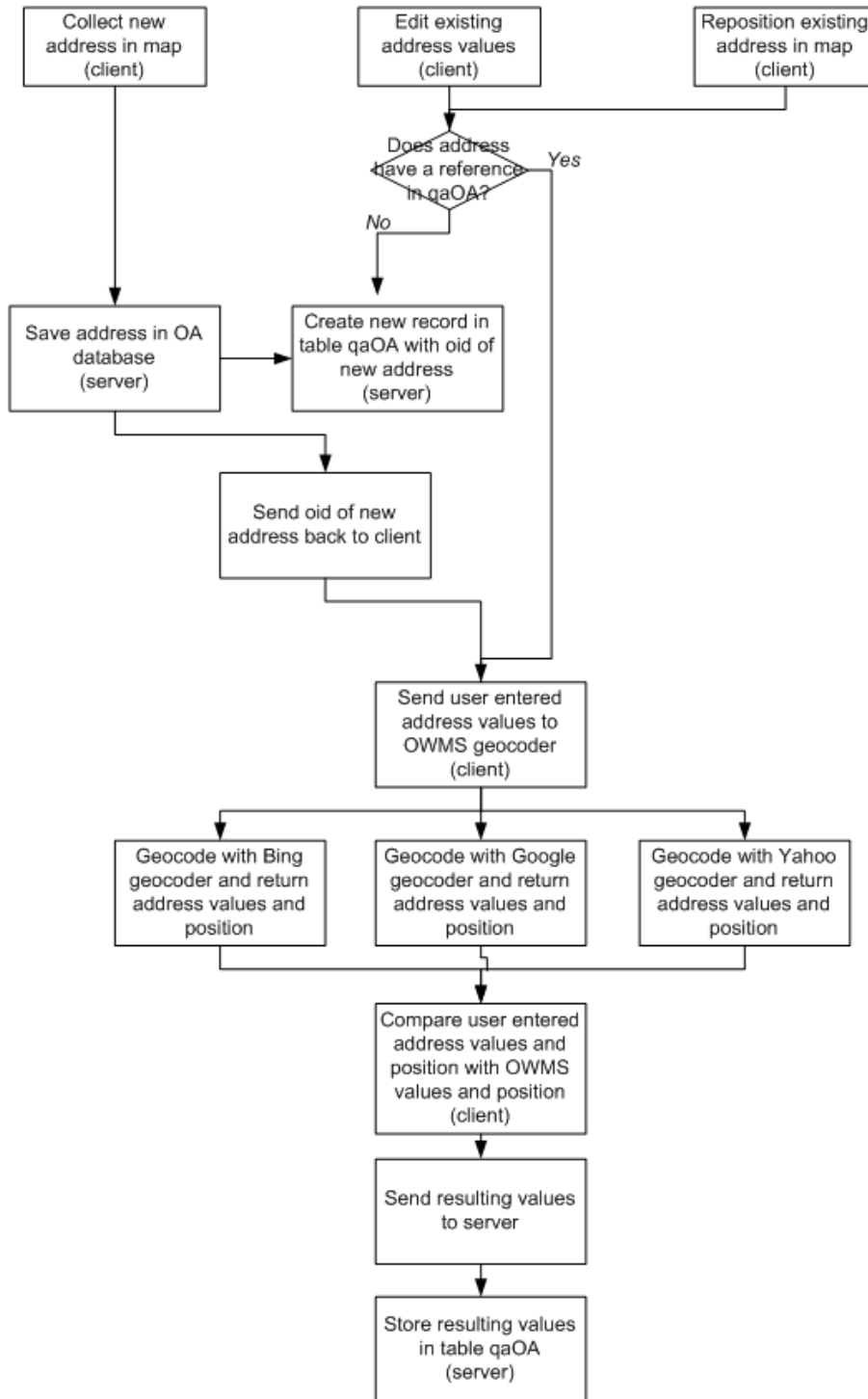


Fig. 41 Flow chart of storage of information for each address after comparison with OWMS

The JavaScript and PHP functionalities implemented here can be found on the attached CD-ROM. These assure that every change to the OA address database is recorded in the qaOA table.

## Quality Assurance of OpenAddresses

To evaluate all new records or alterations, a PHP script documents each change and produces an html file that viewable in a standard web browser. This combines information from the `qaOA` table and the native address table to produce a quality report. Besides attribute information, it uses colour coding to inform a user whether addresses are likely to be correct or if there is a potential need to check and update them.

Fig. 42<sup>59</sup> shows the result of such a quality report in a web browser. A static Google Maps picture helps to show if the address position is correctly placed on a building. Green colouring indicates concordance of user-entered values with OWMS geocoding data. Light red indicates that the distance between a user-defined location and the corresponding reference data point exceeds the threshold defined in the OWMS assessment. Dark red indicates outliers, i.e., positions that are very likely misplaced. In such cases, the OWMS reference dataset may contain no such address, possibly because of a malicious entry.

Thus, if the data fields of an address record are all or mainly green, it can be assumed that the address is correct in terms of both attribute values and position. If the record has red-toned fields it should be checked manually.

---

<sup>59</sup> Fig. 42 uses dummy threshold values for colour coding because the figure was created during the development phase

## Quality assurance for OpenAddresses

This page shows quality indicators for OpenAddresses.org addresses.

Distance values are in [m]. 'addr', 'zip', 'city' indicate as binary values whether the user entered values are identical with the ones from OWMs. 'addr\_level' indicates as binary value whether the user entered address values could be geocoded by the OWMs to address-level.

In order to change the position or address values of an address simply click on its oid. OpenAddresses launches in a new window at the address' location.

mapview	oid	street	house_nr	supplement	postal_code	city	Google				Bing				Yahoo				usr	date			
							dist	addr	zip	city	addr_level	dist	addr	zip	city	addr_level	dist	addr			zip	city	addr_level
	538502	Thunstrasse	124	.	3074	Muri bei Bern	2.597	t	t	t	t	43.022	t	t	t	t	25.888	t	t	t	t	qaOA_e	2010-04-17-10:59:06
	538501	Thunstrasse	122	.	3074	Muri bei Bern	42.385	t	t	t	t	56.211	t	t	t	t	56.999	t	t	t	t	qaOA_f1	2010-04-17-10:58:48
	538502	Thunstrasse	120	.	3074	Muri bei Bern	92.115	t	t	t	t	97.227	t	t	t	t	106.705	t	t	t	t	qaOA_r2	2010-04-17-10:58:32
	538494	Dorf	.	.	3764	Weissenburg	31.739	f	f	f	f	72.711	f	t	f	t	46.746	f	t	f	f	qaOA_e	2010-04-17-10:52:24
	538499	Rutschweid	.	.	3413	Kaltacker	92.054	f	f	f	f	746.195	f	f	f	f	697.098	f	f	f	f	qaOA_r2	2010-04-17-10:51:38
	538497	Alpenstrasse	16	.	3800	Interlaken	35.839	t	t	t	t	41.207	t	t	t	t	46.759	t	t	t	t	qaOA_f1	2010-04-17-10:49:43

Fig. 42 Quality evaluation of new or altered address records

The quality report offers three optional parameters (presented in Table 10) that allow customization of the report's contents. The three parameters are:

Parameter	Description
number	Maximum number or records to be displayed in the report; integer value (e.g. number=50)
order	Sequence of presentation; the value equals the syntax of the SQL <sup>60</sup> order by-statement (e.g. order=date desc)
since date	Timeframe; indicates the oldest date (in the format yyyyymmdd <sup>61</sup> ) from which the address will be listed (e.g. since date=20100416)

Table 10 Query parameters for quality assessment report

<sup>60</sup> SQL (structured query language); cf. <http://en.wikipedia.org/wiki/SQL> [online April 15, 2010]

<sup>61</sup> y=year, m=month, d=day

## Quality Assurance of OpenAddresses

A sample call to create a quality report is shown in Listing 5.

```
http://geoweb05.cti.ac.at/OpenAddressesv1Beta/htdocs/_mthshj/_qa/report_qaOA.php?number=100&order=date%20desc&since=20100416
```

### Listing 5 Sample call for quality assessment report

The complete SQL statement for the report is shown in Listing 6. This does not contain the parameters introduced in Table 10: if these are evaluated, PHP is necessary to adapt the SQL statement accordingly.

```
select a.oid, a.street, a.house_nr, a.supplement, a.postal_code, a.city,
qaOA.google_dist, qaOA.google_addr, qaOA.google_zip, qaOA.google_city,
qaOA.google_precision, qaOA.bing_dist, qaOA.bing_addr, qaOA.bing_zip,
qaOA.bing_city, qaOA.bing_precision, qaOA.yahoo_dist, qaOA.yahoo_addr,
qaOA.yahoo_zip, qaOA.yahoo_city, qaOA.yahoo_precision, a.usr, qaOA.date,
ST_X(ST_Transform(a.coordinates, 4326)) as
lng, ST_Y(ST_Transform(a.coordinates, 4326)) as lat
from addresses as a, qaOA
where a.status <> 'delete' and qaOA.oid=a.oid
order by qaOA.date desc
```

### Listing 6 SQL statement for quality report

The static map view integrated in the report generator has a limit of 1,000 unique image requests per day per viewer<sup>62</sup>. If activity exceeds this limit these map views are omitted and a dummy map view is displayed, as shown in Fig. 43 (below). If the limit is exceeded often and regularly, Google may even block access to the Static Maps API<sup>63</sup>.

The application of the concept can be tested using the URL

[http://geoweb05.cti.ac.at/OpenAddressesv1Beta/htdocs/\\_mthshj/\\_qa/index2.php](http://geoweb05.cti.ac.at/OpenAddressesv1Beta/htdocs/_mthshj/_qa/index2.php)

to collect addresses. The URL

[http://geoweb05.cti.ac.at/OpenAddressesv1Beta/htdocs/\\_mthshj/\\_qa/report\\_qaOA.php](http://geoweb05.cti.ac.at/OpenAddressesv1Beta/htdocs/_mthshj/_qa/report_qaOA.php)

shows the result of the applied quality assessment on the collected addresses.

---

<sup>62</sup> <http://code.google.com/intl/en-EN/apis/maps/documentation/staticmaps>  
[online April 16, 2010]

<sup>63</sup> <http://code.google.com/intl/en-EN/apis/maps/documentation/staticmaps>  
[online April 16, 2010]

id	name	type	city	lat	lon	col1	col2	col3	col4	col5	col6	col7	col8	col9	col10	col11	col12	col13	col14	col15	col16	col17	col18	col19	col20	col21	col22	col23	col24	col25	col26	col27	col28	col29	col30	col31	col32	col33	col34	col35	col36	col37	col38	col39	col40	col41	col42	col43	col44	col45	col46	col47	col48	col49	col50	col51	col52	col53	col54	col55	col56	col57	col58	col59	col60	col61	col62	col63	col64	col65	col66	col67	col68	col69	col70	col71	col72	col73	col74	col75	col76	col77	col78	col79	col80	col81	col82	col83	col84	col85	col86	col87	col88	col89	col90	col91	col92	col93	col94	col95	col96	col97	col98	col99	col100
538765	Bietenweg	2a	3882 Innerkirchen	47.748	10.111	f	t	t	t	4000000	f	f	f	f	1000000	f	t	t	t	t	qaOA_2	2010-04-23-23:34:20																																																																																			
538764	Bietenweg	2	3882 Innerkirchen	43.074	10.111	t	t	t	t	3000000	f	f	f	f	30073	t	t	t	t	t	qaOA_2	2010-04-23-23:34:10																																																																																			
538763	Bietenweg	2b	3882 Innerkirchen	77.873	10.111	t	t	t	t	3000000	f	f	f	f	75.834	f	t	t	t	t	qaOA_2	2010-04-23-23:34:03																																																																																			
538762	Bietenweg	3	3882 Innerkirchen	90.176	10.111	t	t	t	t	4000000	f	f	f	f	80.105	t	t	t	t	t	qaOA_2	2010-04-23-23:33:58																																																																																			
538761	Bietenweg	1	3882 Innerkirchen	85.194	10.111	t	t	t	t	4000000	f	f	f	f	92.203	t	t	t	t	t	qaOA_2	2010-04-23-23:33:49																																																																																			

Fig. 43 Missing static map views due to excessive numbers of map requests

To facilitate corrections of address values or positions, a link is installed behind each `oid` value, centred on the location of the address to be corrected, to launch the OpenAddresses web-page (cf. Fig. 42). This link provides an easy way of correcting attribute values, repositioning or even deleting the address. To implement this option the initializing function (`initializeMap()`) had to be adapted slightly, including the insertion of a function to read and use the positional parameter values to centre the map.

Table 11 presents the files created for the process of comparing individual addresses with OWMS.

File-name	Description
<code>_qaOA.js</code>	This file provides the central functionality of the comparison process. Its logic is presented in Fig. 41 (above).
<code>updateTable.php</code>	This file receives several parameter values that update table <code>qaOA</code> .
<code>report_qaOA.php</code>	This file creates a report of the comparison results of individual addresses against OWMS data (cf. Fig. 42)

Table 11 Overview and description of required files for OWMS batch geocoding process

The complete code listings can be found on the attached CD-ROM.

## 6 Results

### 6.1 Evaluation of OWMS geocoders

#### 6.1.1 Evaluation according to ISO/TC 211:19113 and ISO/TC 211:19114

The geocoders of Bing, Google Maps and Yahoo! Maps are evaluated according to the measures defined in Section 4.3.1. The results are presented in Table 12. This first evaluation involves the analysis of all records, independent of their geocoding level.

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

Table 12 Results of OWMS quality evaluation: Applied quality methods

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!

One point that may partly explain the low thematic accuracy rates is that, for some city names, the reference data include the two-letter Canton codes, e.g. Aeschi SO. As these codes are not returned by the OWMS geocoders, they lead to mismatches regarding binary comparisons of the reference and OWMS data values.

If the thematic accuracy is reduced to a match of the joined strings of street name & house number & zip code only, i.e., omitting the city name, the accuracy values improve significantly (cf. Table 13).

	Bing Maps	Google Maps	Yahoo! Maps
Values of street name & house number & zip code match	58,754 (62.8%)	62,649 (66.9%)	49,070 (52.4%)

Table 13 Distinct analysis on thematic accuracy

However, with the relaxation of constraints, as in Table 13, chances of mismatches rise, especially with erroneous combinations of zip codes and city names.

For all three OWMS datasets, RMSEs for positional accuracy are extremely high. As explained above in 4.2.1, all OWMS also report the geocoding quality of their data. Taking this quality information into account allows for a more differentiated analysis of each OWMS's geocoding quality (cf. Table 14, Table 15 and Table 16).

Number of records	Geocoding Precision	Comments
76,630	Good / High / Interpolated	-
16,797	/ Medium / Interpolated	-
152	Good / High / Rooftop	-
42	//	-
1	Good / Medium / Interpolated	-

Table 14 Geocoding quality of Bing Maps with sample data from the Canton of Solothurn

## Quality Assurance of OpenAddresses

Number of records	Geocoding Precision	Comments
87,206	address	-
4,621	street	-
1,072	zip	-
566	city	-
105	-	-
41	undefined	Unknown address
11	undefined	Server error

Table 15 Geocoding quality of Google Maps with sample data from Canton Solothurn

Number of records	Geocoding Precision	Comments
74,388	address	-
11,058	address	The street number could not be found. Here is a nearby location.
4,373	city	The exact location could not be found. Here is the centre of the ZIP code.
1,837	address	The exact location could not be found. Here is the closest match:
1,305	street	The street name might have been changed
444	street	-
162	zip	The exact location could not be found. Here is a nearby neighborhood.
28	zip	-
27	city	-

Table 16 Geocoding quality of Yahoo! Maps with sample data from Canton Solothurn

For each OWMS geocoder, further constraints are applied to yield the best possible RMSE value not biased by either bad geocoding quality or bad thematic accuracy (cf. Table 17).

OWMS	Restriction
Bing Maps	Values of street name & house number & zip code & city name must match and the Geocoding Precision value must be 'Good / High'
Google maps	Values of street name & house number & zip code & city name must match and the Geocoding Precision indicator must be 'address'
Yahoo! Maps	Values of street name & house number & zip code & city name must match, Geocoding Precision indicator must be 'address' and there must be no additional comments

Table 17 Restriction of OWMS to determine RMSE of positional accuracy

The application of constraints in Table 17 leads to the results in Table 18.

OWMS	RMSE [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%

Table 18 RMSE of positional accuracy with applied constraints

Table 12, Table 13 and Table 18 show that the overall quality of Google Maps is the best of the three investigated OWMSs, followed by Bing Maps, then Yahoo! Maps. This hierarchy is noteworthy when OWMSs are used as indicators on data quality regarding positional accuracy of OA.

According to Amelunxen (2009), Fig. 44 shows the entire value range of the error distances for all OWMS after the application of constraints in Table 17, while Fig. 45 presents a detailed view of the upper 10 percentile (x axis) and

## Quality Assurance of OpenAddresses

lower 500m deviations (y axis). From a global perspective the percentiles of all three OWMS look similar. With focus both on percentile and deviations it becomes more obvious that Google Maps shows the lowest deviations up to the 97th percentile. There is a striking kink around 98 and 99 percentile for all OWMS, after which all error distances rise sharply. Data beyond this threshold must be considered outliers.

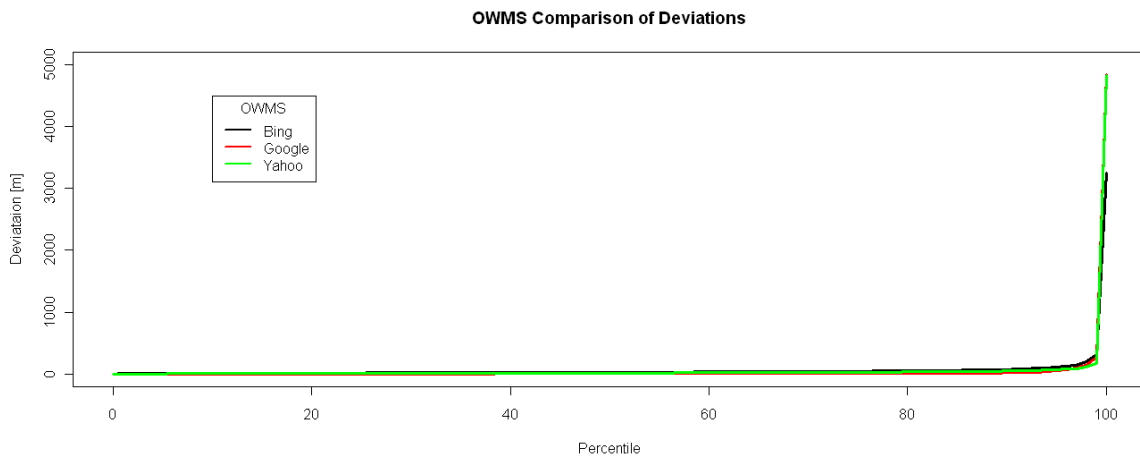


Fig. 44 Comparison of deviations of all OWMS

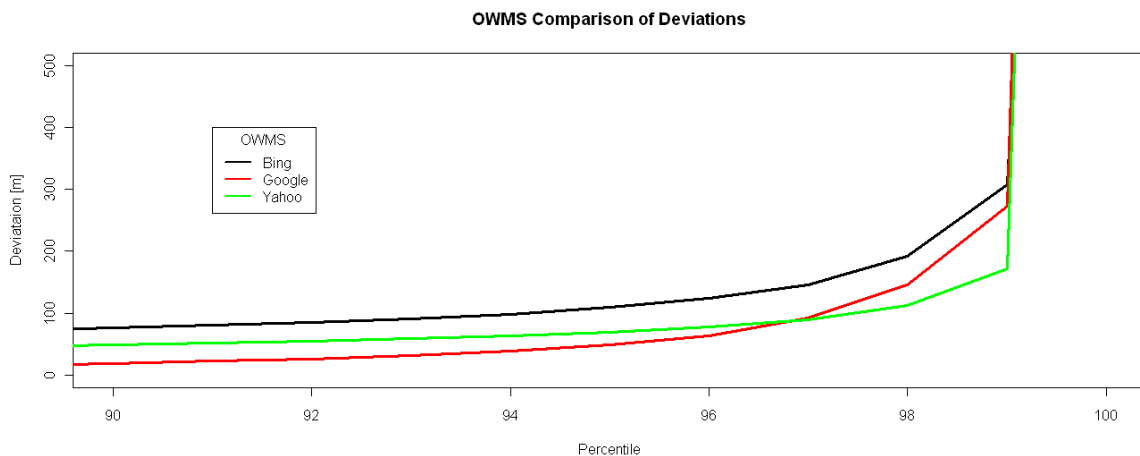


Fig. 45 Detailed view of deviations of all OWMS

### 6.1.2 Defining thresholds for error distances

A more detailed analysis of the error distances for the three OWMS geocoders is performed using the R statistical software package. This analysis is applied to obtain the best possible estimators of threshold values for each OWMS with regard to the OA data quality assessment.

### 6.1.2.1 Basic Statistical Values on Positional Accuracy

Standard statistical values for deviations (minimum (min), maximum (max), mean, median and standard deviation (stdev)) are computed and presented in Table 19, with “n” representing the number of processed records. In the following analysis, all deviation values are expressed in meters.

OWMS	min	max	mean	median	stdev	n
Bing Maps	0.03	3,255.27	43.50	27.86	77.21	47,786
Google maps	0.03	4,840.42	16.51	4.52	81.24	54,281
Yahoo! Maps	0.09	4,829.61	27.17	15.79	76.18	41,524

Table 19 Statistical analysis in R of positional accuracy with applied constraints (cf. Table 17)

### 6.1.2.2 Histograms

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. Fig. 46 to Fig. 48).

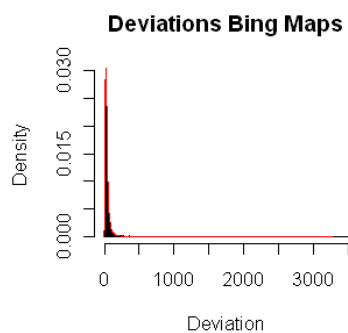


Fig. 46 Histogram of deviations for Bing Maps

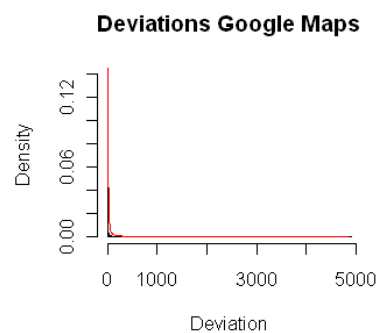


Fig. 47 Histogram of deviations for Google Maps

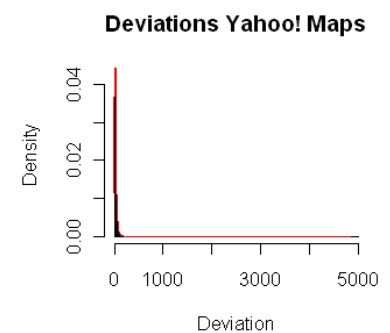


Fig. 48 Histogram of deviations for Yahoo! Maps

All histograms are skewed strongly toward the left. This distribution is logical because it can be expected that most deviations are small, while large ones can be considered outliers. In order to clarify and compare the distribution densities of the deviations, histograms using a single range for all three datasets are presented below in Fig. 49 to Fig. 51.

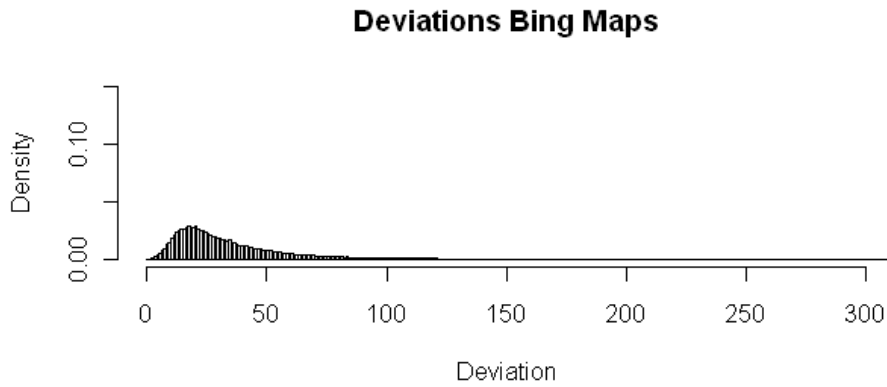


Fig. 49 Histogram of distances between Bing Maps' geocoded objects and reference dataset

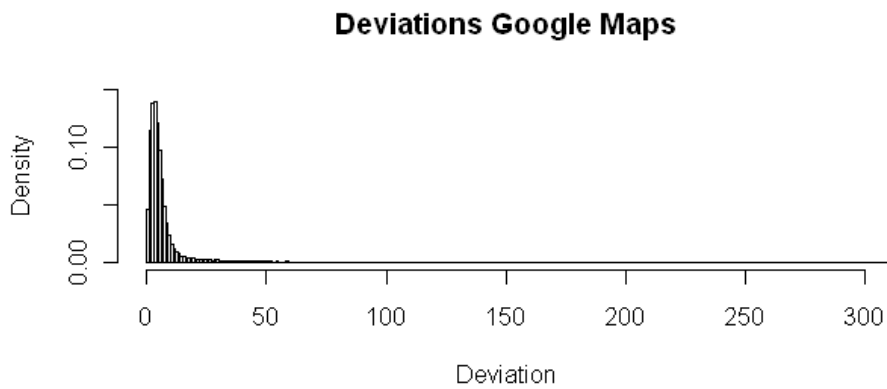


Fig. 50 Histogram of distances between Google Maps' geocoded objects and reference dataset

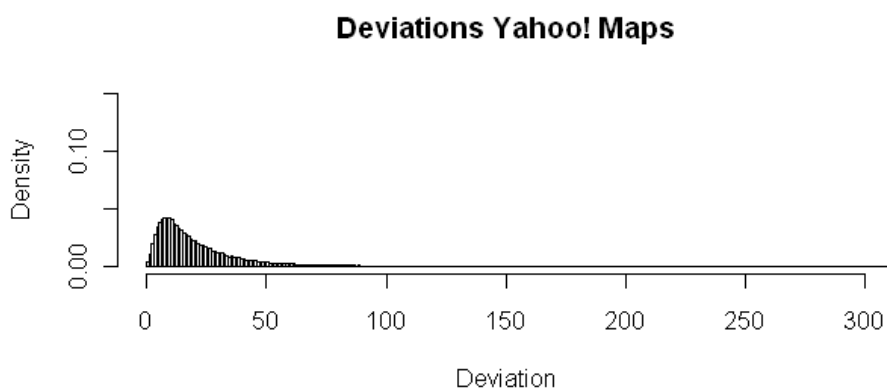


Fig. 51 Histogram of distances between Yahoo! Maps' geocoded objects and reference dataset

The density of deviations is highest with Google Maps, data followed by Yahoo! Maps and finally Bing Maps, where increased statistical density indicates greater overall spatial accuracy. In other words, Google Maps' dataset is spatially closest to the reference data, followed by Yahoo! Maps and finally Bing Maps, which show respectively broader ranges.

### 6.1.2.3 Boxplots on Error Distance

Fig. 52 to Fig. 54 show boxplots of deviations. The limits of the whiskers are defined as  $\pm 1.5$  times the interquartile range (IQR).<sup>64</sup> These boxplots indicate clearly gross errors of geocoding as outliers.

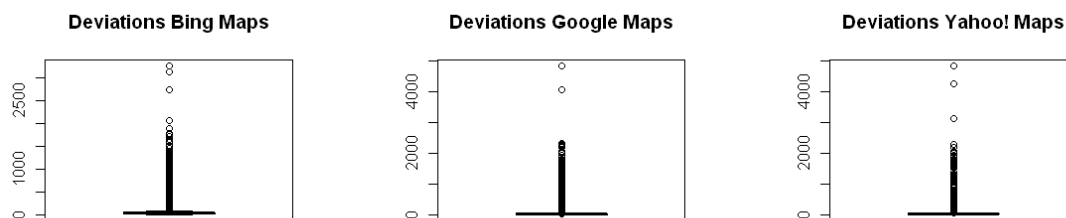


Fig. 52 Boxplot of deviations for Bing Maps

Fig. 53 Boxplot of deviations for Google Maps

Fig. 54 Boxplot of deviations for Yahoo! Maps

All three OWMSs show outliers. Table 20 shows the statistics of boxplots presented in Fig. 52 to Fig. 54. Google Maps shows by far the smallest values and can thus be considered by far the most accurate indicator of positional accuracy.

	Lower end of whisker	25th percentile	median	75th percentile	Upper end of whisker
Bing Maps	0.309	17.783	27.859	45.853	87.941
Google Maps	0.029	2.666	4.517	7.448	14.616
Yahoo! Maps	0.090	8.807	15.793	28.294	57.519

Table 20 Overview of boxplots statistics of distance evaluation

<sup>64</sup> [http://en.wikipedia.org/wiki/Interquartile\\_range](http://en.wikipedia.org/wiki/Interquartile_range) and <http://en.wikipedia.org/wiki/Boxplot> [online April 21, 2010]

## Quality Assurance of OpenAddresses

Fig. 55 to Fig. 57 show boxplots with identical ranges to show the whiskers more clearly.

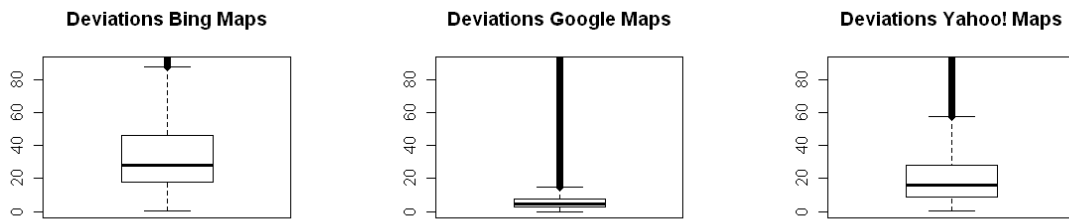


Fig. 55 Zoomed boxplot of deviations for Bing Maps

Fig. 56 Zoomed boxplot of deviations for Google Maps

Fig. 57 Zoomed boxplot of deviations for Yahoo! Maps

As an example of an outlier address, 'Kirchstrasse 7, 4716 Gänsbrunnen' is used. Based on the Google Maps geocoder, its location is presented below in Fig. 58, indicated by the red circle symbol in the image's upper right quadrant, while its true location is indicated with a yellow circle in the lower left quadrant. According to Google Maps the geocoding level is set to 'address', which is the best possible quality to achieve: Google Maps' address values match to the ones of the reference dataset exactly. However, Google Maps' location of this address is roughly 5km displaced.

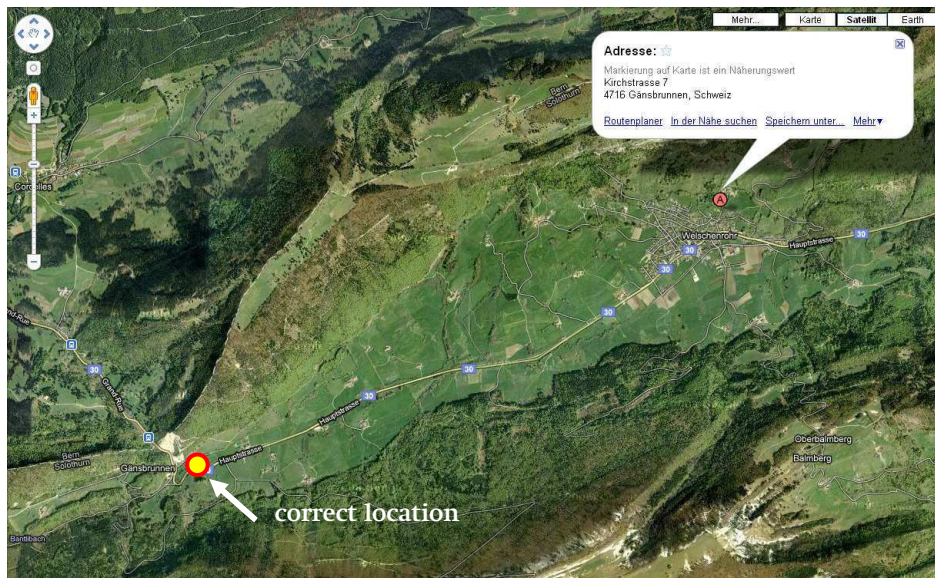
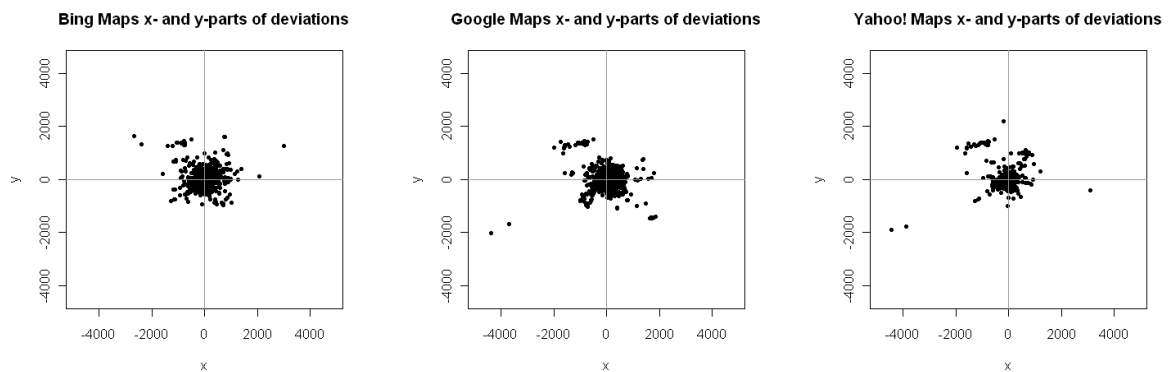


Fig. 58 Comparison of wrong location of sample address in Google Maps and correct location according to <http://www.sogis1.so.ch>

Due to the large size of the dataset such outliers were not investigated systematically. They are identified via boxplots (cf. Fig. 52 to Fig. 54) and the resulting computed values (cf. Table 20).

### 6.1.3 Further Statistical Analysis Based on x- and y-values

So far deviations have been analysed only as Euclidean distances (cf. Formula [2]). 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. Fig. 59 to Fig. 61).



**Fig. 59** Scatter plot of deviations split into x- and y-directions for Bing Maps

**Fig. 60** Scatter plot of deviations split into x- and y-directions for Google Maps

**Fig. 61** Scatter plot of deviations split into x- and y-direction 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 (cf. Table 19).

Also, each plot shows a striking cluster around  $(-1.000, 1.700)$ , possibly caused by the effect of displacement OWMS introduced in Fig. 58. Such an effect could appear in a certain region like a small village or hamlet.

## Quality Assurance of OpenAddresses

As elaborated above, outliers have an effect on the RMSE definition of the global dataset. ISO/TC 211:19138 (2006; p. 42) suggests the application of a threshold to determine the mean value of positional uncertainties excluding outliers.

As deviations, the positional uncertainties are calculated as follows:

$$[1] \quad e'_i = \begin{cases} e_i, & \text{if } e_i \leq e_{\max} \\ 0, & \text{if } e_i > e_{\max} \end{cases}$$

with

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

and

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

$x_{mi}$  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  $\bar{e}_{\text{excluding outliers}}$  is the RMSE based on all  $e'_i$ .

Because the range of deviations can vary greatly (cf. Fig. 35 and Table 4) 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 will be considered for further analysis.

Based on this limitation  $e_{\max}$  and consequently  $\bar{e}_{\text{excluding outliers}}$  are determined.

Hence  $\bar{e}_{\text{excluding outliers}}$  substitutes for RMSE as the indicator for positional accuracy.

Table 21 shows the thresholds in x- and y-directions for selecting addresses process further for each OWMS, along with the largest deviation within this subset and the resulting number of addresses, listed both as numbers and as percentages of the original count. Metric values are in meters.

	x 2.5%	x 97.5%	y 2.5%	y 97.5%	Max deviation <sup>65</sup>	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

Table 21 Determining limits in x- and y-directions using 95% Quantile

The analysis of the deviations' distribution in x- and y-directions for each OWMS is shown in Fig. 62 to Fig. 67.

---

<sup>65</sup> The values of Max deviations are the empiric ones from the dataset, not the theoretical ones from the analysis

# Quality Assurance of OpenAddresses

## Bing Maps

Distribution x-direction Bing Maps (Q95%)

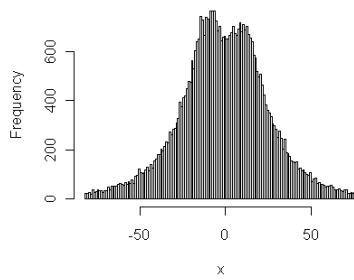


Fig. 62 Histogram of x-direction deviations for Bing Maps

Distribution y-direction Bing Maps (Q95%)

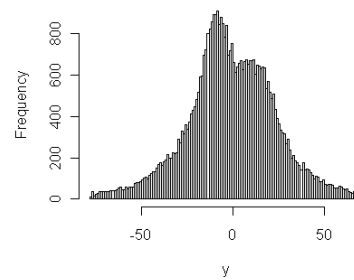


Fig. 63 Histogram of y-direction deviations for Bing Maps

## Google Maps

Distribution x-direction Google Maps (Q95%)

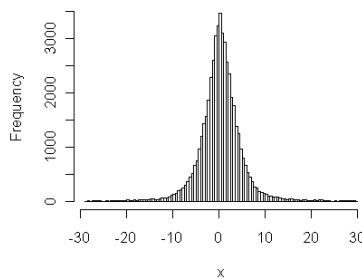


Fig. 64 Histogram of x-direction deviations for Google Maps

Distribution y-direction Google Maps (Q95%)

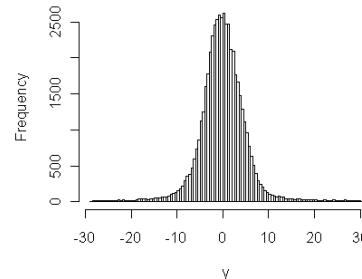


Fig. 65 Histogram of y-direction deviations for Google Maps

## Yahoo! Maps

Distribution x-direction Yahoo! Maps (Q95%)

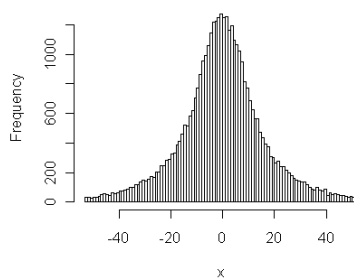


Fig. 66 Histogram of x-direction deviations for Yahoo! Maps

Distribution y-direction Yahoo! Maps (Q95%)

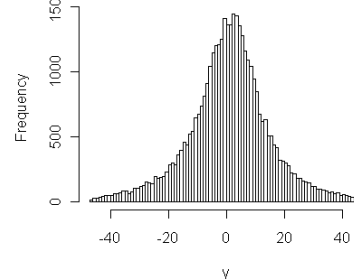
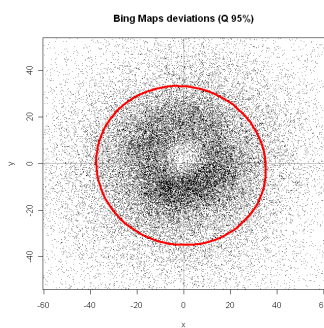


Fig. 67 Histogram of y-direction deviations for Yahoo! Maps

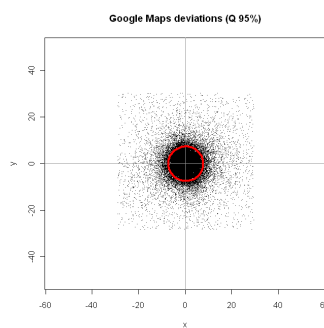
Fig. 62 to Fig. 67 all indicate symmetrical Gaussian distributions in both x- and y-directions, although in Fig. 62 and Fig. 63, towards the mean, i.e., 0, the frequency decreases. This is due to the geocoding algorithm: as shown in Fig. 35 Bing Maps uses an algorithm that aligns the interpolated locations

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 Fig. 62 and Fig. 63, 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 (Fig. 67).

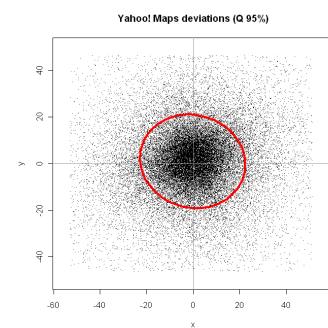
Redrawing scatterplots similar to Fig. 59 to Fig. 61 but with the subset described in Table 21 leads to the standard deviation ellipses as presented in Fig. 68 to Fig. 70.



**Fig. 68** Scatterplot of deviations in x- and y-directions for Bing Maps with standard deviation ellipse, Q95%



**Fig. 69** Scatterplot of deviations in x- and y-directions for Google Maps with standard deviation ellipse, Q95%



**Fig. 70** Scatterplot of deviations in x- and y-directions for Yahoo! Maps with standard deviation ellipse, Q95%

In Fig. 71, for comparison purposes, the standard deviation ellipses are drawn in a single plot computed using R according to the values described in Table 22. The figure shows that Bing Maps' ellipse is largest, while Google Maps' is the smallest and is almost circular.

## Quality Assurance of OpenAddresses

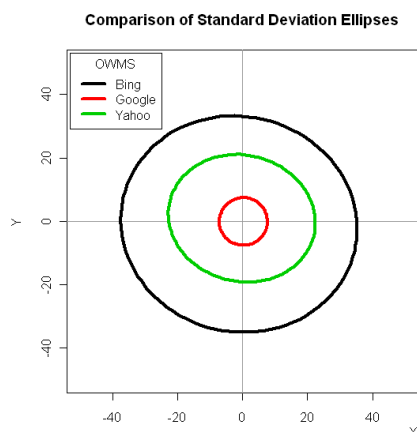


Fig. 71 Comparison of Standard Deviation Ellipses

As explained above, the influence of the geocoding algorithm applied by Bing Maps is also visible in Fig. 68: toward the centre of the standard deviation ellipse the point density becomes noticeably lighter.

	centre x	centre y	sigma x	sigma y	theta	Eccentricit y
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

Table 22 Statistics of Standard Deviation Ellipses based on x- and y-clusters

Table 22 gives an overview of the computed parameters of the three OWMS's standard deviation ellipses 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.<sup>66</sup>

The ellipses are all close to their origins. Those of Bing Maps and Yahoo! Maps are rotated circa 75 degrees in the same direction. Google Maps' ellipse is also oriented in that direction but with reversed axes. The half-length axes

<sup>66</sup> [http://bm2.genes.nig.ac.jp/RGM2/R\\_current/library/aspaces/man/calc\\_sde.html](http://bm2.genes.nig.ac.jp/RGM2/R_current/library/aspaces/man/calc_sde.html) [online April 23, 2010]

are different for all three: Google Maps' ellipse is nearly a circle with a radius of 7.5m, while Yahoo! Maps' and Bing Maps' have more unequal values (cf. Fig. 71). The rotation shows a slight isotropy – very likely caused by a non-random distribution of directions of street-segments in the Canton of Solothurn. Consulting Fig. 37, the shape of the canton and the distribution of the address-points indicate that there are potentially more streets running south-west to north-east than in other directions. Due to the lateral offset the parts of the error distances along this direction are smaller than those perpendicular to it. This illustrates well how the angles of the standard deviation ellipses match the direction of building locations and street directions.

As the next step, x- and y-values are tested for correlations. If they are not correlated the small differences between sigma x and sigma y values will result in a circular standard deviation ellipse. For this reason, empirical error distance values can be used as parameters to test against a defined threshold to evaluate the positional accuracy of OA's data. The correlation coefficients between x- and y-values are presented in Table 23.

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

Table 23 Correlation Coefficients of x- and y-values for each OWMS

The correlation coefficient for Google is nearly 0. Along with Fig. 69 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. Along with Fig. 68 and Fig. 70 this light correlation may be due to the eccentricity of the standard deviation ellipse. Based on the fact that deviations can vary strongly (cf. Fig. 35 and Table 4) no further analysis on the correlation of x- and y-values for Bing Maps and Yahoo! Maps are conducted. It is assumed that also for these two OWMS x- and y-values are independent - indicated by the small correlation coefficients - and therefore also a circle can be assumed as standard deviation ellipse. Hence empiric

## Quality Assurance of OpenAddresses

error distance values are used as parameters to test against a defined threshold to evaluate positional accuracy of OA data and no comparison in x- and y-direction of deviations is necessary.

The definition  $e_{\max}$  is derived from the computed values of the 95% Quantile in x- and y-direction for each OWMS (cf. Table 21). The resulting values are presented in Table 24 . These values are the result for the data quality subelement absolute or external accuracy introduced in Table 6.

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

Table 24  $e_{\max}$  and concluding RMSE /  $\bar{e}_{\text{excluding outliers}}$  for each OWMS

### 6.1.4 Conclusion

The difference between the location of an address in the real world and the location of the address that is determined based on interpolated navigation data provided by the OWMS, can vary largely. It also depends on its vicinity: highly populated areas have shorter street-line segments than less populated areas. Thus the interpolation accuracy is higher in highly populated areas. Based on the presented investigations it would be too ambitious to determine a threshold with a sub-metre accuracy to identify good locations versus outliers.

The purpose of the thresholds that are applied for quality assessment of OA data that is collected or modified is to give an indication of whether further investigation on the address is indicated or whether the dataset can be regarded as good quality. It must be recalled at this point that the quality assessment of OA data based on OWMS services is a comparison of a dataset that claims to have high spatial accuracy with a dataset that is of less (spatial) quality.

To evaluate reasonable estimators for threshold values for the quality assessment of positional accuracy in OA, boxplots for all three OWMS are

created based on the 95% quantile of x- and y-direction. This subset was already used in the analysis in 6.1.3. Within this subset again the 95% quantile is determined and defined as threshold for quality assessment of positional accuracy of OA data. The maximum distance of the 95% quantile in x- and y-directions defines the threshold to determine outliers.

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

Table 25 Threshold values for quality assessment of OA data for positional accuracy

Threshold values presented in Table 25 are except for Google Maps above the threshold that Ahlers and Boll (2008) defined. That does not mean that the quality of OA data is therefore bad. The threshold values serve as indicators for the quality evaluation. They can nevertheless assist in achieving a positional quality of OA data that fulfils the claim of OA that all addresses are located within the ground view of the building they belong to.

## **6.2 Evaluation of Quality Assessment for OpenAddresses**

### **6.2.1 Null-hypothesis for quality evaluation of OpenAddresses data**

The null-hypothesis is: An address is correctly collected and entered in OA.

Type I error<sup>67</sup> or false positives means that an address is classified as erroneous although it is correct. The consequences are that this address will be inspected and requests further resources to assess its quality although this is not necessary.

Type II error false negatives means that an address is classified as correct although it is erroneous. No further action is taken on this address and it is kept as is in the database. The consequences are that the overall database quality becomes worse.

---

<sup>67</sup> cf. [http://en.wikipedia.org/wiki/False\\_positives#Type\\_I\\_error](http://en.wikipedia.org/wiki/False_positives#Type_I_error) [online May 4, 2010]

Whenever possible errors of type II should be omitted because they are more dangerous and threatening the overall quality of the entire dataset.

### **6.2.2 Test-addresses for Quality Evaluation of OpenAddresses**

According to 5.4 and the results of 6.1.4 the data-collection of addresses within OA is implemented and tested. In order to get a certain degree of independence in the acquisition of addresses, students at the Institute of Geomatics Engineering of University of Applied Sciences Northwestern Switzerland were asked to collect some addresses that were divided into three test-classes.

The classification was applied with the following constraints:

- a) Digitise an address in OA at its correct location on the building (class c).
- b) Digitise an address in OA at close to its correct location outside the building (class f1).
- c) Digitise an address in OA at a location with a gross spatial error (class f2).

For all three types address values were entered correctly. The students are considered as familiar with the addresses they enter thus erroneous address values are considered as non-existent. The spatial distribution of test-addresses is shown in Fig. 72. The map was created in Manifold GIS.

Thresholds of Table 25 are used to define the null-hypothesis for the quality assessment of OA data collection.

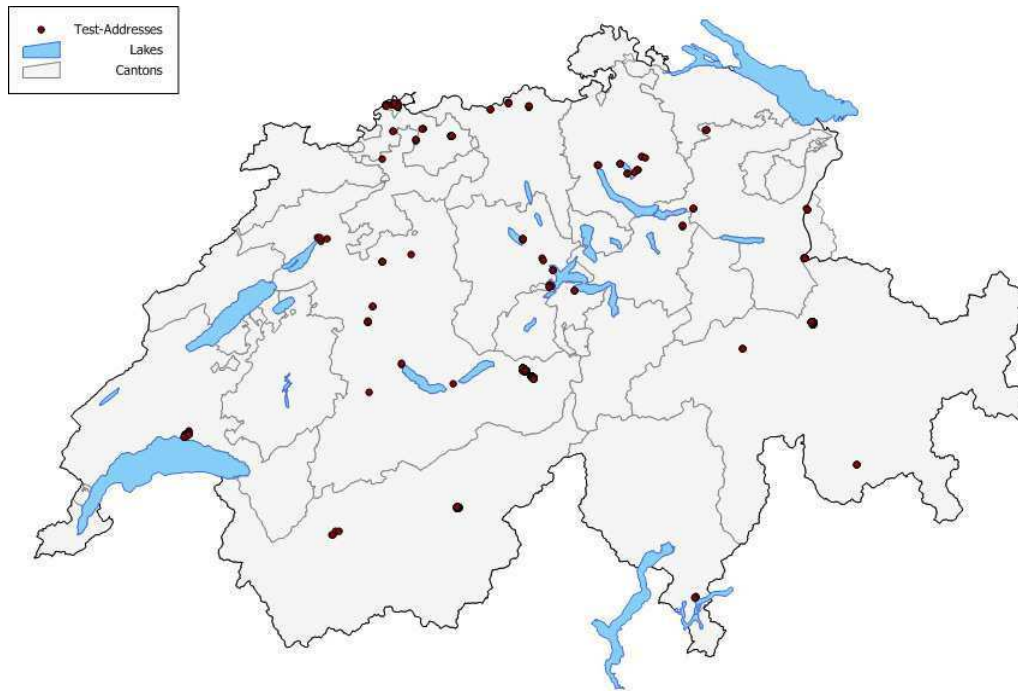


Fig. 72 Spatial distribution of test-addresses

Table 30 shows the total number of collected addresses per class:

	Class c	Class f1	Class f2
Number of addresses	172	118	123

Table 26 Number of test-addresses for quality evaluation of OA

### 6.2.3 Evaluation according to ISO/TC 211:19113 and ISO/TC 211:19114

Applying the measures defined in 4.3.2 leads to the following two statistics:

Non-quantitative attribute correctness is evaluated for all test-addresses independent of their test-class and presented in 6.2.3.1.

Absolute or external accuracy is computed for all test-addresses as error distance (cf. Formula [2] in Section 6.1.3) for each test-class and evaluated in Section 6.2.3.2.

### 6.2.3.1 Evaluation of Attribute Correctness

User entered values are compared binary to the address-values returned from OWMS server. Street name and house number were combined to 'address' information. This leads to the statistics in Table 27 to Table 29. The evaluation of the corresponding measure is assessed as pass or fail. This test is strict and also tests on case sensitivity of first letters of a string and on needless spaces.

#### 6.2.3.1.1 Bing Maps

	Total	Address		Zip-Code		City-Name		Address-level geocoding	
Evaluation <sup>68</sup>		P	F	P	F	P	F	P	F
Absolute	413	247	166	387	26	372	41	374	39
Relative	100	59.8%	40.2%	93.7%	6.3%	90.1%	9.9%	90.6%	9.4%

Table 27 Evaluation of non-quantitative attribute correctness for Bing Maps

In total 231 (55,9%) of all test-addresses showed identical address-values between user entrance and Bing Maps geocoder.

#### 6.2.3.1.2 Google Maps

	Total	Address		Zip-Code		City-Name		Address-level geocoding	
Evaluation		P	F	P	F	P	F	P	F
Absolute	413	254	159	389	24	375	38	398	15
Relative	100	61.5%	38.5%	94.2%	5.8%	90.8%	9.2%	96.4%	3.6%

Table 28 Evaluation of non-quantitative attribute correctness for Bing Maps

With Google Maps geocoding a total of 238 (57.6%) of all test-addresses showed identical address-values between user entrance and the OWMS geocoder.

---

<sup>68</sup> P = pass, F = fail

### 6.2.3.1.3 Yahoo! Maps

	Total	Address		Zip-Code		City-Name		Address-level geocoding	
		P	F	P	F	P	F	P	F
Absolute	413	285	128	410	3	379	34	404	9
Relative	100	69.0%	31.0%	99.3%	0.7%	91.8%	8.2%	97.8%	2.2%

Table 29 Evaluation of non-quantitative attribute correctness for Bing Maps

Yahoo! Maps geocoder returns for 265 (64.2%) of all test-addresses identical address-values compared to the ones the user has entered.

Overall Yahoo! Maps shows the best results for the address-value evaluation. Fortunately for all three OWMS the address-values returned from the geocoder match with the ones the user has entered for more than half of all test-addresses.

If only one of the three OWMS<sup>69</sup> is used as an indicator for correct address-values the number raises up to 318 (77%) which is around three quarters.

The results from 6.2.3.1.1, 6.2.3.1.2 and 6.2.3.1.3 show a better thematic accuracy than the ones elaborated in 6.1.1.

Generally speaking the test on attribute values is rather strict. Misspellings and other erroneously entered values are detected. With a combined approach only around 25% of addresses lead to an error of type I.

---

<sup>69</sup> i.e. either Bing Maps or Google Maps or Yahoo! Maps return identical address-values compared to the user entered values

### 6.2.3.1.4 Detected Problems with Address Comparison

There are some problems that cause a 'fail' value in the comparison between the user entered values and the ones returned from the OWMS geocoder. Some of these are briefly introduced.

The following address "Brüel 25 9496 Balzers" is an example for an address that causes problems. It is classified from Google Maps and Yahoo! Maps geocoders to address-level. Yet none of the three returns the address-values that a user from this place (native) enters (cf. Table 30).

	Address	Zip-Code	City-Name	Distance to user entered location [m]
Bing Maps	Balzersweg	5600	Ammerswil	102,655.0
Google Maps	25 Brüel	-	Mäls	15.7
Yahoo! Maps	25 Brüel	9496	Mäls	15.7

Table 30 Comparison of returned address-values for an address

Other problems are that occasionally the OWMS returns additional information on one of the address-information values such as the Canton in which a city is located: a search for 'Hergiswil' is returns 'Hergiswil (NW)' from Google Maps. This leads to a 'fail' value when comparing this returned value binarily to the user entered values.

House numbers that have an additional letter such as '20A' cause also problems and often only '20' is returned by the OWMS geocoder. This leads to a 'fail' in the binary comparison.

In some smaller villages like 'Maloja' Google Maps could not find a match at all. In other areas like 'Gamprin-Bendern' all three OWMS could not find a match to the user entered values. It seems that there are obviously regions or areas where data is missing.

Another problem is that locations of non-existent addresses can be computed by a street-geocoding algorithm. E.g. 'Froburgstrasse 6 4052 Basel'

does not exist, but 'Froburgstrasse 4' and 'Froburgstrasse 12' do. A street-geocoding algorithm like with Bing Maps and Yahoo! Maps will interpolate such an address entrance and return a location - although in the real world this address does not exist. Interestingly also Google Maps that uses the high quality Geopost dataset returns a value for such a non-existent address.

### 6.2.3.2 Evaluation of Positional accuracy

The positional accuracy is assessed based on the classes introduced in 6.1.4.

#### 6.2.3.2.1 Quality assessment with Class c

Class c contains addresses that are positioned on the building and should thus be considered as true locations. This means that they are supposed to be within Q95% according to the model of 6.1.4. Evaluating the error distances lead to the results in Table 31. The threshold values of Table 25 are applied.

	Bing Maps	Google Maps	Yahoo! Maps
Number of addresses classified as correct locations	119 69.2%	125 72.7%	129 75.0%
Number of addresses above Q95% threshold	53 30.8%	47 27.3%	43 25.0%
Number of addresses below outliers threshold	147 85.5%	144 83.7%	154 89.5%
Number of addresses classified as outliers	25 14.5%	28 16.3%	18 10.5%

Table 31 Evaluation test-addresses of class c on positional accuracy

If only one of the OWMS thresholds is taken into consideration between 69.2% and 75.0% of all correctly positioned addresses are declared as correct. It follows that in around 25% - 30% of addresses errors of type I occur. Between 10.5% and 16.3% of all addresses are erroneously judged as outliers if only the information of one OWMS assessment is considered.

## Quality Assurance of OpenAddresses

In order to achieve a more reliable assertion a combined OWMS assessment is considered. This means that information of all the three OWMS is taken into consideration. The assumption is that for only one of the OWMS the error distance must fall below the corresponding Q95-threshold and the geocoding type of the record must be on address level. This constraint *cn1* is formulated as:

$$cn1: \quad (d_{bm} \leq ts95_{bm} \wedge gcl_{bm}) \vee (d_{gm} \leq ts95_{gm} \wedge gcl_{gm}) \vee (d_{ym} \leq ts95_{ym} \wedge gcl_{ym})$$

with *cn*=constraint, *d*= error distance, *ts95*=threshold value Q95, *gcl*=geocoding level, *bm*=Bing Maps, *gm*=Google Maps and *ym*=Yahoo! Maps

This approach leads to 162 (94.2%) addresses that are judged as correct and thus the error of type I can be minimised to 5.8%.

Applying a more strict constraint *cn2* defined as

$$cn2: \quad (d_{bm} \leq ts95_{bm} \wedge gcl_{bm}) \wedge (d_{gm} \leq ts95_{gm} \wedge gcl_{gm}) \wedge (d_{ym} \leq ts95_{ym} \wedge gcl_{ym})$$

leads to 84 (48.8%) addresses that are classified as correctly positioned. This strict constraint is considered to be too pessimistic.

### 6.2.3.2.2 Quality assessment with Class f1

The same procedure as with test-class c (cf. Section 6.2.3.2.1) is performed on class f1.

	Bing Maps	Google Maps	Yahoo! Maps
Number of addresses classified as correct locations	53 44.9%	54 45.8%	92 78.0%
Number of addresses above Q95% threshold	65 55.1%	64 54.2%	26 22.0%
Number of addresses below outliers threshold	60 50.8%	86 72.9%	103 87.3%
Number of addresses classified as outliers	58 49.2%	32 27.1%	15 12.7%

Table 32 Evaluation test-addresses of class c on positional accuracy

The figures in Table 32 indicate a blurred basis for decision-making. Yahoo! Maps rates 78% as good locations while Bing Maps and Google Maps are around 45%. The percentage on the outlier-detection varies strongly among the three OWMS from 12,5% (Yahoo! Maps) up to 49,2% (Bing Maps).

Applying cn1 leads to 103 (87,3%) addresses that are judged as good locations. In other words for 87,3% of addresses with a slight erroneous position error of type II can occur. Applying cn2 lowers this figure down to 25 (21,2%) which is significantly better.

**6.2.3.2.3 Quality assessment with Class f2**

As with test-classes c (cf. Section 6.2.3.2.1) and f1 (cf. Section 6.2.3.2.2) the same evaluation procedure is performed with test-class f2.

	Bing Maps	Google Maps	Yahoo! Maps
Number of addresses classified as correct locations	15 12.2%	0 0.0%	14 11.4%
Number of addresses above Q95% threshold	108 87.8%	123 100%	109 88.6%
Number of addresses below outliers threshold	31 25.2%	10 8.1%	20 16.3%
Number of addresses classified as outliers	92 74.8%	113 91.9%	103 83.7%

Table 33 Evaluation test-addresses of class f2 on positional accuracy

Examining Table 33 it is striking that Google Maps detects nearly 92% of all addresses of test-class f2 correctly as outliers while Bing Maps and Yahoo! Maps achieve a rate of around 75% and 84%. If the threshold value of Q95 is applied Google Maps even scores 100%.

If the combined constraint cn3 is considered that any of the three OWMS error distances are above the outlier threshold 114 (92.7%) addresses are correctly identified as outliers. Thus the error of type II can be minimized to 7.3%

$$cn3: (d_{bm} > tsO_{bm}) \vee (d_{gm} > tsO_{gm}) \vee (d_{ym} > tsO_{ym})$$

with tsO=threshold outlier

A more strict constraint cn4 requires that all error distances are above the defined outlier threshold:

$$cn4: (d_{bm} > tsO_{bm}) \wedge (d_{gm} > tsO_{gm}) \wedge (d_{ym} > tsO_{ym})$$

Applying cn4 classifies 90 (73.2%) addresses as outliers.

### 6.2.3.2.4 Conclusio

Table 34 summarises the results of the quality assessment of classes c, f1 and f2.

	Class c	Class f1	Class f2
cn1	162 94.2%	103 87.3%	18 14.6%
cn2	84 48.8%	25 21.2%	0 0.0%
cn3	59 34.3	71 60.2%	114 92.7%
cn4	0 0.0%	5 4.2%	90 73.2%

Table 34 Applying quality assessment constraints on test-classes

Constraints cn1 and cn3 are compared to constraints cn2 and cn4 a little more relaxed. They lead to good results in the test-classes they are designed for (classes c and f2). One drawback with cn1 and cn3 is that objects in test-class f1 are not classified well; thus no reliable statement can be made on their data quality.

With cn1 nearly 15% of outliers are not detected and hence lead to error type II. This is considered as too dangerous.

With cn2 errors of type II are omitted for clear outliers (test-class f2) but the quota of error type I raises significantly from 5.8% (cn1) to 51.2%. This means that a little more than 50% of correctly located addresses are classified as not correctly positioned on the map. This is a high percentage. On the other hand cn2 avoids bad records in the database which is more important. Additionally the rate of error type II for addresses with minor positional errors (test-class f1) is lowered with cn2 down to 21.2%. This means that around one fifth of addresses with a slight positional error are considered as addresses with correct locations. This percentage is not as low as hoped but since these errors are minor the percentage is accepted.

## Quality Assurance of OpenAddresses

cn3 detects outliers well with a rate of 92.7%. Thus chances for an error of type II are minimised to 7.3%. With cn3 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.

cn4 allows for the exclusion of correctly located addresses in the category of outliers. This eliminates errors of type I. But cn4 also leads to a 26.8% risk of type II error. This risk is considered too high because it means that the quality of every fourth incorrectly located address could erroneously be judged as good.

Hence cn2 and cn3 are accepted as constraints for the classification of addresses in terms of positional accuracy within the process of quality evaluation in OA. cn2 and cn3 do not overlap. But they leave a certain "grey area" for which no predication can be made on the positional quality.

Constraints cn2 and cn3 are implemented additionally for the quality assessment of OA and lead to a new column on the website. Below in Fig. 73 the changes are implemented. Addresses of test-class c are mostly classified correctly. So are addresses of test-class f2. Two addresses of test-class f1 are not classified while a third one (at the end of listing in Fig. 73) is erroneously classified as address with correct location (error type II).

QA Ass	mapview	oid	street	house_nr	supplement	postal_code	city	Google				Bing				Yahoo				usr	date
								dist	addr	zip	city	addr_level	dist	addr	zip	city	addr_level	dist	addr		
OUT		538476	Seestrasse	10	-	8124	Maur	37.848	t	t	t	79.717	t	t	t	79.388	t	t	t	qaOA_f2	2010-05-05-15:37:22
OUT		538416	Homburgstrasse	21	-	4052	Basel	204.903	t	t	t	245.833	t	t	t	238.027	t	t	t	qaOA_f2	2010-05-05-15:35:40
+		538410	Homburgstrasse	6	-	4052	Basel	3.715	t	t	t	13.49	t	t	t	10.223	t	t	t	qaOA_e	2010-04-28-15:27:53
		538476	Seestrasse	8	-	8124	Maur	22.537	t	t	t	4.542	t	t	t	11.923	t	t	t	qaOA_f1	2010-04-28-15:25:44
+		538474	Seestrasse	6	-	8124	Maur	1.514	t	t	t	27.414	t	t	t	20.481	t	t	t	qaOA_e	2010-04-28-15:25:27
		538523	Bernoullistrasse	30	-	4056	Basel	22.667	f	t	t	32.505	f	t	t	10.502	f	t	t	qaOA_e	2010-04-28-15:20:07
		538524	Klingelbergstrasse	50	-	4056	Basel	18.693	f	t	t	25.302	f	t	t	28.566	f	t	t	qaOA_f1	2010-04-28-15:19:39
OUT		538525	Greifengasse	22	-	4058	Basel	39.04	f	t	t	83.545	f	t	t	75.633	f	t	t	qaOA_f2	2010-04-28-15:18:33
+		393070	Avenue du Théâtre	14	-	1005	Lausanne	5.92	t	t	t	14.955	t	t	t	4.822	t	t	t	qaOA_f1	2010-04-28-15:11:23

Fig. 73 Adapted quality assessment for OA

## 7 Summary, Conclusion and Future Work

### 7.1 Summary and Conclusion

The seven key findings of this thesis are:

1. ISO/TC 211 19100 family standards proved to be useful in the quality evaluation process of crowdsourced data.
2. OWMS can successfully be integrated into the real-time quality assessment for OpenAddresses data.
3. Statements on the correctness of attribute values of addresses are only reliable in around 75%. 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.
4. Positional accuracy is difficult to assess because error distances between true location and OWMS interpolated location can vary greatly. Nevertheless the statistical analysis in this thesis lead to a fairly robust indication on the positional correctness of a user entered address-location. Although small positional errors may not be detected gross errors are identified.
5. The web-based overview of real-time quality evaluation of OA data is useful and practical and allows for a real-time adjustment of erroneous data in the database via the graphical user interface of OA.
6. The visual impression of a static Google Maps map excerpt helps in evaluating whether the location of an address was defined as within the ground view of the building and thus it may help preventing records with small positional errors.
7. This thesis approves that a less accurate reference dataset can help in assessing a better dataset in terms of being an indicator especially for gross errors.

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 (cf. results found in Amelunxen (2009)). But since Switzerland has a high level of data quality in OWMS it can be assumed that rates of error of type I will increase compared to those of type II if an assessment in other countries is omitted.

### **7.2 Future Work**

During phase of this thesis OpenAddresses has undergone major changes and is no longer running on Google Maps environment. It has changed to MapFish<sup>70</sup> as portrayal component and uses Yahoo! Maps as backdrop map along with OpenStreetMap and other freely available WebMapServices (WMS). One of the inherent consequences is that the license type has changed and has been adapted to the one of OpenStreetMap in order to be compatible with the OSM project. This has lead to the deletion of all addresses in OA that had been captured based on Google Maps API. The new version is online since March 29, 2010 as beta release at [www.openaddresses.org](http://www.openaddresses.org).

The good news is that the entire work of this thesis was conceptually designed and implemented independent of the core mapping components of OA. Thus it can be implemented relatively easy and fast into the new environment.

The application of Google Maps static API has some major drawbacks since the daily amount is limited. For the future it must be decided on whether this feature shall be kept in the quality evaluation process or not or whether it shall be available only on demand.

---

<sup>70</sup> <http://www.mapfish.org> [online May 7, 2010]

## Quality Assurance of OpenAddresses

Since OA is now 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 Fig. 73 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 crowdsourcing 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.





## A Appendix

### A.1 Figures

Fig. 74 and Fig. 75 show the result of the batch geocoding of all addresses of the Canton of Solothurn via Bing Maps, Google Maps and Yahoo! Maps.



Fig. 74 Visual comparison of OWMS geocoding results in Laupersdorf produced with QGIS

## Quality Assurance of OpenAddresses



Fig. 75 Visual comparison of OWMS geocoding results in Olten produced with QGIS

## A.2 Listings of SQL statements on the evaluation of OWMS geocoding

The following code section presents applied SQL statements for the evaluation of user entered addresses in OpenAddresses. It is the formal representation of the four constraints introduced in 6.2.3.

```

////////////////////////////////////
/////
////////////////////////////////////Thematic Accuracy Evaluation
////////////////////////////////////
/////
//Bing (for Google Maps and Yahoo Maps accordingly)
select count(*),qaOA.bing_addr  from addresses as a, qaOA where (a.oid =
qaOA.oid and a.usr in ('qaOA_c', 'qaOA_f1', 'qaOA_f2')) group by 2;
select count(*),qaOA.bing_zip  from addresses as a, qaOA where (a.oid =
qaOA.oid and a.usr in ('qaOA_c', 'qaOA_f1', 'qaOA_f2')) group by 2;
select count(*),qaOA.bing_city  from addresses as a, qaOA where (a.oid =
qaOA.oid and a.usr in ('qaOA_c', 'qaOA_f1', 'qaOA_f2')) group by 2;
select count(*),qaOA.bing_precision  from addresses as a, qaOA where (a.oid =
qaOA.oid and a.usr in ('qaOA_c', 'qaOA_f1', 'qaOA_f2')) group by 2;

select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr in
('qaOA_c', 'qaOA_f1', 'qaOA_f2') and qaOA.bing_addr and qaOA.bing_zip and
qaOA.bing_city);
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr in
('qaOA_c', 'qaOA_f1', 'qaOA_f2') and qaOA.bing_addr and qaOA.bing_zip and
qaOA.bing_city and qaOA.bing_precision);

/////if one of the three OWMS is ok
select count(*) from addresses as a, qaOA
where ((a.oid = qaOA.oid and a.usr in ('qaOA_c', 'qaOA_f1', 'qaOA_f2')) and
((qaOA.bing_addr and qaOA.bing_zip and qaOA.bing_city and
qaOA.bing_precision) OR
(qaOA.google_addr and qaOA.google_zip and qaOA.google_city and
qaOA.google_precision) OR
(qaOA.yahoo_addr and qaOA.yahoo_zip and qaOA.yahoo_city and
qaOA.yahoo_precision)));

////////////////////////////////////
/////
////////////////////////////////////Positional Accuracy Evaluation
////////////////////////////////////
/////
////Class c / user = qaOA_c

//Bing
//within Q95
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_c' and bing_dist<=67.08)
//Outside Q95
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_c' and bing_dist>67.08)

//Google
//within Q95
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_c' and google_dist<=15.36)
//Outside Q95

```

## Quality Assurance of OpenAddresses

```
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_c' and google_dist>15.36)

//Yahoo
//within Q95
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_c' and yahoo_dist<=42.62)
//Outside Q95
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_c' and yahoo_dist>42.62)

////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
/
//combination inside Q95 and geocoding level = 'address' --> classified as
correct location (relaxed) as cn1
////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
/
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_c') and ((google_dist<=15.36 and google_precision) or bing_dist<=67.08
or yahoo_dist<=42.62)

//combination inside Q95 and geocoding level = 'address' --> classified as
correct location (strict) as cn2
////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
/
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_c') and ((google_dist<=15.36 and google_precision) and bing_dist<=67.08
and yahoo_dist<=42.62)

////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
/////Class f1 / user = qaOA_f1    according to Class c
////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////

////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
/////Class f2 / user = qaOA_f2    according to Class c
////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////

/////combination outlier --> classified as outlier cn3
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_f2') and (google_dist>40.81 or bing_dist>111.76 or yahoo_dist>68.41)
/////combination outlier --> classified as outlier cn4
select count(*) from addresses as a, qaOA where (a.oid = qaOA.oid and a.usr =
'qaOA_f2') and (google_dist>40.81 and bing_dist>111.76 and yahoo_dist>68.41)

////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////////
//combination of cn2 and cn 3 to check whether cn2 and cn3 do not overlap

select count(*) from addresses as a, qaOA
where (a.oid = qaOA.oid and a.usr = 'qaOA_c')
and ((google_dist<=15.36 and google_precision) and bing_dist<=67.08 and
yahoo_dist<=42.62)
and (google_dist>40.81 and bing_dist>111.76 and yahoo_dist>68.41)
```

**Listing 7 SQL statements applied in chapter 6.1**

### A.3 R Code Listings

The following section provides parts of the applied R commands that were used in chapter 6.1

```
#####
#####
#####
#####
#General Analysis
#####
#####
#####

#####
#####
##Bing Maps
#####
#####
#Read data in R
bing_dist=read.table("dist_sogis_bing.dat", head=TRUE)

#histogram of total range
hist(bing_dist$dist_to_reference, breaks=c(0:3500), xlab="Dist",
main="Deviations Bing Maps", freq=FALSE, xlim=c(0,3500))
#line
lines(density(bing_dist$dist_to_reference), col="red")

#histogram of specific range
hist(bing_dist$dist_to_reference, breaks=500, xlab="Deviation",
main="Deviations Bing Maps", freq=FALSE, xlim=c(0,300))

#Boxplot and its statistics
boxplot(bing_dist$dist_to_reference, main="Deviations Bing Maps")
boxplot(bing_dist$dist_to_reference, main="Deviations Bing Maps",ylim = c(0,
90))
boxplot.stats(bing_dist$dist_to_reference, do.conf = TRUE, do.out = FALSE)

#####
#####
##Google Maps
#####
#####
#Read data in R
google_dist=read.table("dist_sogis_google.dat", head=TRUE)

#histogram of total range
hist(google_dist$dist_to_reference, breaks=c(0:4900), xlab="Dist",
main="Deviations Google Maps", freq=FALSE, xlim=c(0,4900))
#line
lines(density(google_dist$dist_to_reference), col="red")

#histogram of specific range
hist(google_dist$dist_to_reference, breaks=500, xlab="Dist", main="Deviations
Google Maps", freq=FALSE, xlim=c(0,300))

#Boxplot and its statistics
boxplot(google_dist$dist_to_reference, main="Deviations Google Maps")
boxplot(google_dist$dist_to_reference, main="Deviations Google Maps",ylim =
c(0, 90))
boxplot.stats(google_dist$dist_to_reference, do.conf = TRUE, do.out = FALSE)
```

## Quality Assurance of OpenAddresses

```
#####  
#####  
##Yahoo Maps  
#####  
#####  
#Read data in R  
yahoo_dist=read.table("dist_sogis_yahoo.dat", head=TRUE)  
  
#histogram of total range  
hist(yahoo_dist$dist_to_reference, breaks=c(0:5000), xlab="Dist",  
main="Deviations Yahoo! Maps", freq=FALSE, xlim=c(0,5000))  
#line  
lines(density(yahoo_dist$dist_to_reference), col="red")  
  
#histogram of specific range  
hist(yahoo_dist$dist_to_reference, breaks=500, xlab="Dist", main="Deviations  
Yahoo! Maps", freq=FALSE, xlim=c(0,300))  
  
#Boxplot and its statistics  
boxplot(yahoo_dist$dist_to_reference, main="Deviations Yahoo! Maps")  
boxplot(yahoo_dist$dist_to_reference, main="Deviations Yahoo! Maps",ylim =  
c(0, 90))  
boxplot.stats(yahoo_dist$dist_to_reference, do.conf = TRUE, do.out = FALSE)  
  
#####  
#####  
#####  
#####  
#Percentile  
#####  
#####  
#####  
#####  
  
rng <- seq(0, 1.0, 0.01)  
  
b_quan <- quantile(bing_dist$dist_to_reference, probs = rng)  
g_quan <- quantile(google_dist$dist_to_reference, probs = rng)  
y_quan <- quantile(yahoo_dist$dist_to_reference, probs = rng)  
  
b_ser <- data.frame(col1=b_quan, col2=c(rng * 100))  
g_ser <- data.frame(col1=g_quan, col2=c(rng * 100))  
y_ser <- data.frame(col1=y_quan, col2=c(rng * 100))  
  
##Plot overview  
plot(c(0,100), c(0,5000), ylab="Deviataion [m]", xlab="Percentile", type="n",  
main="OWMS Comparison of Deviations")  
points(b_ser$col2,b_ser$col1, type="l", col="black", lwd=3)  
points(g_ser$col2,g_ser$col1, type="l", col="red", lwd=3)  
points(y_ser$col2,y_ser$col1, type="l", col="green", lwd=3)  
  
leg.txt <- c("Bing", "Google", "Yahoo")  
leg.col <- c("black", "red", "green")  
leg.lwd <- c(2,2,2)  
leg.lty <- c(1,1,1)  
legend(10,4500,leg.txt, col=leg.col, lwd=leg.lwd, lty=leg.lty, title="OWMS")  
  
##Plot focus percentile  
plot(c(0,100), c(0,5000), ylab="Deviataion [m]", xlab="Percentile", type="n",  
main="OWMS Comparison of Deviations", xlim=c(90,100), ylim=c(0,500))  
points(b_ser$col2,b_ser$col1, type="l", col="black", lwd=3)  
points(g_ser$col2,g_ser$col1, type="l", col="red", lwd=3)  
points(y_ser$col2,y_ser$col1, type="l", col="green", lwd=3)  
  
leg.txt <- c("Bing", "Google", "Yahoo")  
leg.col <- c("black", "red", "green")
```

```

leg.lwd <- c(2,2,2)
leg.lty <- c(1,1,1)
legend(91,400,leg.txt, col=leg.col, lwd=leg.lwd, lty=leg.lty, title="OWMS")

#####
#####
#####
#Analysis of x- and y-values individually
#####
#####
#####

#####
#####
##Bing Maps
#####
#####
#Read data in R
bing_x_y_sde=read.table("dist_deltas_sogis_bing_20100422_calc_sde.dat",
head=TRUE)

#Scatterplot
plot(bing_x_y_sde$x, bing_x_y_sde$y, asp = 1, main="Bing Maps x- and y-parts
of deviations", xlab="x", ylab="y", pch=20, xlim=c(-4500,4500), ylim=c(-
4500,4500))
abline(h=0, v=0, col = "gray60")
#Plot with focus to be able to see the ellipse later on
plot(bing_x_y_sde$x, bing_x_y_sde$y, asp = 1, main="Bing Maps x- and y-parts
of deviations", xlab="x", ylab="y", pch=20, xlim=c(-120,120), ylim=c(-
120,120))
abline(h=0, v=0, col = "gray60")

#Calc Standard Deviation Ellipse
#save SDE values in variable
b<-
calc_sde(id=1,filename="R.txt",calccentre=TRUE,points=bing_x_y_sde,weighted=FA
LSE)
str(b)

#Plot Standard Deviation Ellipse - save in file
png(filename="bing.png", width=512, height=512, pointsize=12)
plot_sde(plotSDEaxes=TRUE, points.col=1, points.pch=19, plotcentre=TRUE,
titlext="Standard Deviation Ellipse Bing Maps", xaxis="x", yaxis="y",
sde.col=2, sde.lwd=3)
dev.off()

#####
#####
##Google Maps
#####
#####
#Read data in R
google_x_y_sde=read.table("dist_deltas_sogis_google_20100422_calc_sde.dat",
head=TRUE)

#Scatterplot
plot(google_x_y_sde$x, google_x_y_sde$y, asp = 1, main="Google Maps x- and y-
parts of deviations", xlab="x", ylab="y", pch=20, xlim=c(-4500,4500), ylim=c(-
4500,4500))
abline(h=0, v=0, col = "gray60")
#Plot with focus to be able to see the ellipse later on

```

## Quality Assurance of OpenAddresses

```
plot(google_x_y_sde$x, google_x_y_sde$y, asp = 1, main="Google Maps x- and y-
parts of deviations", xlab="x", ylab="y", pch=20, xlim=c(-120,120), ylim=c(-
120,120))
abline(h=0, v=0, col = "gray60")

#Calc Standard Deviation Ellipse
#save SDE values in variable
gm<-
calc_sde(id=1,filename="R.txt",calccentre=TRUE,points=google_x_y_sde,weighted=
FALSE)
str(gm)

#Plot Standard Deviation Ellipse - save in file
png(filename="gm2.png", width=512, height=512, pointsize=12)
plot_sde(plotSDEaxes=TRUE, points.col=1, points.pch=19, plotcentre=TRUE,
titletxt="Standard Deviation Ellipse Google Maps", xaxis="x", yaxis="y",
sde.col=2, sde.lwd=3)
dev.off()

#####
####
##Yahoo! Maps
#####
####
#Read data in R
yahoo_x_y_sde=read.table("dist_deltas_sogis_yahoo_20100422_calc_sde.dat",
head=TRUE)

#Scatterplot
plot(yahoo_x_y_sde$x, yahoo_x_y_sde$y, asp = 1, main="Yahoo! Maps x- and y-
parts of deviations", xlab="x", ylab="y", pch=20, xlim=c(-4500,4500), ylim=c(-
4500,4500))
abline(h=0, v=0, col = "gray60")
#Plot with focus to be able to see the ellipse later on
plot(yahoo_x_y_sde$x, yahoo_x_y_sde$y, asp = 1, xlim=c(-120,120), ylim=c(-
120,120), main="Yahoo! Maps x- and y-parts of deviations", xlab="x", ylab="y",
pch=20)
abline(h=0, v=0, col = "gray60")

#Calc Standard Deviation Ellipse
#save SDE values in variable
y<-
calc_sde(id=3,filename="R.txt",calccentre=TRUE,points=yahoo_x_y_sde,weighted=F
ALSE)
str(y)

#Plot Standard Deviation Ellipse - save in file
png(filename="y.png", width=512, height=512, pointsize=12)
plot_sde(plotSDEaxes=TRUE, plotpoints=FALSE, points.col=1, points.pch=19,
plotcentre=TRUE, titletxt="Standard Deviation Ellipse Yahoo! Maps", xaxis="x",
yaxis="y", sde.col=2, sde.lwd=3)
dev.off()

#####
####
#####Analysis of x- and y-parts without outliers according to 95% quantil
#####
####
###Plot Ellipses in one chart for comparison purpose
#####*****
##Code from: http://www.math.mcmaster.ca/peter/s4m03/s4m03\_0102/ellipse.html
##[online May 4, 2010]
##and were partially adapted

myellipse <-
```

```

function(hlaxa = 1, hlaxb = 1, theta = 0, xc = 0, yc = 0, colr = 1, linewidth =
2, ltype = 1, npoints = 100)
{
  xp <- NULL
  yp <- NULL
  for(i in 0:npoints) {
    a <- (2 * pi * i)/npoints
    x <- hlaxa * cos(a)
    y <- hlaxb * sin(a)
    if(theta != 0) {
      alpha <- angle(x, y)
      rad <- sqrt(x^2 + y^2)
      x <- rad * cos(alpha + theta)
      y <- rad * sin(alpha + theta)
    }
    xp <- c(xp, x + xc)
    yp <- c(yp, y + yc)
  }
  lines(xp, yp, type = "l" , col = colr, lwd = linewidth, lty = ltype)
  invisible()
}
angle <-
function(x, y)
{
  if(x > 0) {
    atan(y/x)
  }
  else {
    if(x < 0 & y != 0) {
      atan(y/x) + sign(y) * pi
    }
    else {
      if(x < 0 & y == 0) {
        pi
      }
      else {
        if(y != 0) {
          (sign(y) * pi)/2
        }
        else {
          NA
        }
      }
    }
  }
}
#####
#####
##Bing Maps
#####
#####
#Read data in R
bing_x_y_sde_95=read.table("bing_95xy.dat", head=TRUE)

#Scatterplot
#Plot with focus to be able to see the ellipse later on
plot(bing_x_y_sde_95$x, bing_x_y_sde_95$y, asp = 1, main="Bing Maps deviations
(Q 95%)", xlab="x", ylab="y", pch=".", xlim=c(-50,50), ylim=c(-50,50))

#Calc Standard Deviation Ellipse
#save SDE values in variable
b95<-
calc_sde(id=1,filename="R.txt",calccentre=TRUE,points=bing_x_y_sde_95,weighted
=FALSE)

#Plot Standard Deviation Ellipse - save in file; one time with one time
without points

```

## Quality Assurance of OpenAddresses

```
x11() ##open new graphic window
plot(bing_x_y_sde_95$x, bing_x_y_sde_95$y, asp = 1, main="Bing Maps deviations
(Q 95%)", xlab="x", ylab="y", pch=".", xlim=c(-50,50), ylim=c(-50,50))
abline(h=0, v=0, col = "gray60")
elli_b95 <-
myellipse(b95$Sigma.x,b95$Sigma.y,as_radians(b95$ThetaCorr),b95$CENTRE.x,b95$C
ENTRE.y,2,4,1,100)

#####
####
##Google Maps
#####
####
#Read data in R
google_x_y_sde_95=read.table("google_95xy.dat", head=TRUE)

#Scatterplot
#Plot with focus to be able to see the ellipse later on
plot(google_x_y_sde_95$x, google_x_y_sde_95$y, asp = 1, main="Google Maps
deviations (Q 95%)", xlab="x", ylab="y", pch=1, xlim=c(-50,50), ylim=c(-
50,50))

#Calc Standard Deviation Ellipse
#save SDE values in variable
g95<-
calc_sde(id=1,filename="R.txt",calccentre=TRUE,points=google_x_y_sde_95,weight
ed=FALSE)

#Plot Standard Deviation Ellipse - save in file; one time with one time
without points
#png(filename="google95_test.png", width=512, height=512, pointsize=12)
x11() ##open new graphic window
plot(google_x_y_sde_95$x, google_x_y_sde_95$y, asp = 1, main="Google Maps
deviations (Q 95%)", xlab="x", ylab="y", pch=".", xlim=c(-50,50), ylim=c(-
50,50))
abline(h=0, v=0, col = "gray60")
elli_g95 <-
myellipse(g95$Sigma.x,g95$Sigma.y,as_radians(g95$ThetaCorr),g95$CENTRE.x,g95$C
ENTRE.y,2,4,1,100)

#####
####
##Yahoo! Maps
#####
####
#Read data in R
yahoo_x_y_sde_95=read.table("yahoo_95xy.dat", head=TRUE)

#Scatterplot
#Plot with focus to be able to see the ellipse later on
plot(yahoo_x_y_sde_95$x, yahoo_x_y_sde_95$y, asp = 1, main="Yahoo! Maps
deviations (Q 95%)", xlab="x", ylab="y", pch=1, xlim=c(-50,50), ylim=c(-
50,50))

#Calc Standard Deviation Ellipse
#save SDE values in variable
y95<-
calc_sde(id=1,filename="R.txt",calccentre=TRUE,points=yahoo_x_y_sde_95,weighte
d=FALSE)

#Plot Standard Deviation Ellipse - save in file; one time with one time
without points
#png(filename="google95_test.png", width=512, height=512, pointsize=12)
x11() ##open new graphic window
plot(yahoo_x_y_sde_95$x, yahoo_x_y_sde_95$y, asp = 1, main="Yahoo! Maps
deviations (Q 95%)", xlab="x", ylab="y", pch=".", xlim=c(-50,50), ylim=c(-
50,50))
```

```

abline(h=0, v=0, col = "gray60")
elli_y95 <-
myellipse(y95$Sigma.x,y95$Sigma.y,as_radians(y95$ThetaCorr),y95$CENTRE.x,y95$CENTRE.y,2,4,1,100)

#####
#####Comparison of Standard Deviation Ellipses
#####
##Plot axes
plot(c(-50,50), c(-50,50), ylab="Y", xlab="", type="n",
     main="Comparison of Standard Deviation Ellipses")
abline(h=0, v=0, col = "gray60")
##Bing Ellipse:
myellipse(33.9,36.6,as_radians(73.1),-1.14,-0.86,1,4,1,100)
##Google Ellipse:
myellipse(7.5,7.4,as_radians(76.6),0.27,-0.06,2,4,1,100)
##Yahoo Ellipse:
myellipse(19.9,22.8,as_radians(74.5),-0.23,0.96,3,4,1,100)
##add a title
##main="Comparison of Standard Deviation Ellipses", sub = "[Largest Ellipse:
Bing Maps \nSecond largest: Yahoo! Maps \n Smallest: Google Maps]")
##label for x-axis
mtext("X", 1, line=2, adj=1)

leg.txt <- c("Bing", "Google", "Yahoo")
leg.col <- c(1,2,3)
leg.lwd <- c(4,4,4)
leg.lty <- c(1,1,1)
legend(-53,53,leg.txt, col=leg.col, lwd=leg.lwd, lty=leg.lty, title="OWMS")

```

Listing 8 R commands applied in chapter 6.1

## **B Literature (Bibliography)**

- Acil, T. (2008). The Value of Spatial Information: The impact of modern spatial information technologies on the Australian economy. Prepared for the CRC for Spatial Information & ANZLIC – the Spatial Information Council. Melbourne.
- Aditya, T. (2008). "Participatory Mapping." *GIM International* **22**(9): 41 - 43.
- Agichtein, E., C. Castillo and D. Donato (2008). Finding high-quality content in social media. Proceedings of the international conference on Web search and web data mining, Palo Alto.
- Ahlers, D. and S. Boll (2008). Retrieving address-based locations from the web. Proceeding of the 2nd international workshop on Geographic information retrieval, Napa Valley, California, USA, ACM.
- Al Rahed, A., S. Coetzee and M. Rademeyer (2008). A data model for efficient address data representation - Lessons learnt from the Intiendo address matching tool. Free and Open Source Software for Geospatial (FOSS4G), Cape Town.
- Amelunxen, C. (2009). An Approach to Geocoding based on Volunteered Spatial Data, UNIGIS Salzburg. (MSc Thesis)
- Auer, M. and A. Zipf (2009). How do free and Open Geodata and Open Standards fit together? From Scepticism versus high Potential to real Applications. The First Open Source GIS UK Conference, Nottingham.
- Baumann, J. (2008). "Volunteered Geographic Information." *GEOconnexion International Magazine* **7**(10): 46 - 47.
- Boin, A. T. and G. J. Hunter (2006). Do spatial data consumers really understand data quality information? 7th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences, Lisbon.
- Bovard, R. (2009). Gestion collaborative du domaine communal, Haute Ecole d'Ingénierie et de Gestion du Canton de Vaud. (BSc Thesis)
- Brown, M. (2006). Hacking Google Maps and Google Earth. Indianapolis, Wiley.
- Cayo, M. R. and T. O. Talbot (2003). "Positional error in automated geocoding of residential addresses." *International Journal fo Health Geographics* **2**(10).

- Chrisman, N. (2006). Development in the Treatment of Spatial data Quality. Fundamentals of spatial data quality. R. Devillers and R. Jeansoulin. London, ISTE: 21 - 30.
- Coetzee, S. and A. Cooper (2007). The value of addresses to the economy, society and governance - a South African perspective. Annual URISA Conference, Washington DC, USA.
- Coetzee, S., A. K. Cooper, M. Lind, M. McCart Wells, S. W. Yurman, E. Wells, N. Griffiths and M. Nicholson (2008). Towards an international address standard. 10th Global Spatial Data Infrastructure Conference (GSDI-10), St Augustine, Trinidad.
- Coleman, D. J. and Y. Georgiadou (2010). "Volunteered Geographic Information. Why and What do individuals contribute?" GEOInformatics(3): 50 - 52.
- Coleman, D. J., Y. Georgiadou and J. Labonte (2009). "Volunteered Geographic Information: The Nature and Motivation of Producers." International Journal of Spatial Data Infrastructures Research 4: 332 - 358.
- Cooper, A. K. (2009). Geoinformation perspectives on innovation and economic growth. Economic Commission for Africa. Addis Ababa.
- Devillers, R. and R. Jeansoulin (2006). Spatial Data Quality: Concepts. Fundamentals of spatial data quality. R. Devillers and R. Jeansoulin. London, ISTE: 31 - 42.
- Elwood, S. (2007). "Grassroots groups as stakeholders in spatial data infrastructures: challenges and opportunities for local data development and sharing." International Journal of Geographic Information Science 22: 71 - 90.
- Elwood, S. (2008a). "Volunteered geographic information: future research directions motivated by critical, participatory, and feminist GIS." GeoJournal(72): 173 - 183.
- Elwood, S. (2008b). "Volunteered Geographic Information: key questions, concepts and methods to guide emerging research and practice." GeoJournal 72: 133 - 135.
- Elwood, S. (2009). "Geographic Information Science: new geovisualization technologies - emerging questions and linkages with GIScience research." Progress in Human Geography 33(2): 256 - 263.
- Erle, S., R. Gibson and J. Walsh (2005). Mapping Hacks. Sebasopol, CA, O'Reilly.

## Quality Assurance of OpenAddresses

- Fischer, F. (2008). "Collaborative mapping. How Wikinomics is Manifest in the Geo-Information Economy." *GEOInformatics* 11(2): 28 - 31.
- Fischer, F. (2009). "Volunteered Geographic Information - Baustein zukünftiger Geoinformationsinfrastrukturen?" *Wiener Schriften zur Geographie und Kartographie* 19: 148-153.
- Fisher, P., A. Comber and R. Wadsworth (2006). *Approaches to Uncertainty in Spatial Data. Fundamentals of spatial data quality.* R. Devillers and R. Jeansoulin. London, ISTE: 43 - 59.
- Flanagin, A. J. and M. J. Metzger (2008). "The credibility of volunteered geographic information." *GeoJournal*(72): 137 - 148.
- 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: 925 - 938.
- Gibson, R. and S. Erle (2006). *Google Maps Hacks.* Sebastopol, CA., O'Reilly.
- Goldberg, D. W. (2008). *A Geocoding Best Practices Guide,* Springfield, IL: North American Association of Central Cancer Registries.
- Goldberg, D. W., J. N. Swift and J. P. Wilson (2008a). *Geocoding best practices: Reference Data, Input Data and Feature Matching,* University of Southern California GIS Research Laboratory.
- Goldberg, D. W., J. P. Wilson, C. A. Knoblock, B. Ritz and M. G. Cockburn (2008b). "An effective and efficient approach for manually improving geocoded data." *International Journal of Health Geographics* 7(60).
- Goodchild, M. F. (2006). *Foreword. Fundamentals of spatial data quality.* R. Devillers and R. Jeansoulin. London, ISTE: 13 - 16.
- Goodchild, M. F. (2007). "Citizens as sensors: the world of volunteered geography." *GeoJournal*(69): 211 - 221.
- Goodchild, M. F. (2008a). "Citizens as sensors." *GIS Trends + Markets*(6): 27 - 29.
- Goodchild, M. F. (2008b). "Commentary: whither VGI?" *GeoJournal*(72): 239 - 244.
- Haklay, M. (2008). *How good is Volunteered Geographic Information? A comparative study of OpenStreetMap and Ordnance Survey datasets.*

- Hancock, C. (2010). "Address management for emergency services." *GEOconnexion International Magazine* 9(2): 20-21.
- Hanguët, J.-F. (2006). Data Quality Assessment and Documentation. *Fundamentals of spatial data quality*. R. Devillers and R. Jeansoulin. London, ISTE: 211 - 235.
- Harris, R., P. Sleight and R. Webber (2006). *Geodemographics, GIS and Neighbourhood Targeting*, John Wiley & Sons, Ltd.
- Harris, T. M., D. Weiner, T. A. Warner and R. Levin (1995). Pursuing Social Goals Through Participatory Geographic Information Systems. *Ground Truth. The Social Implications of Geographic Information Systems*. J. Pickles. New York, The Guildford Press: 196 - 221.
- Harvard University (2008). *The Public Health Disparities Geocoding Project Monograph*
- Howe, J. (2008). *Crowdsourcing. How the Power of the Crowd is Driving the Future of Business*. London, Random House Business Books.
- INSPIRE, I. f. S. I. i. E. (2009). "INSPIRE Data Specification on Addresses – Guidelines."
- ISO/TC 211:19112. (2003). "Geographic information – Spatial referencing by geographic identifiers."
- ISO/TC 211:19113. (2001). "Geographic Information - Quality Principles."
- ISO/TC 211:19114. (2001). "Geographic Information - Quality Evaluation Procedures."
- 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."
- ISO/TC 211:19138. (2006). "Text for TS 19138 Geographic Information - Data quality measures, as sent to ISO for publication." Retrieved March 25 2010, <http://www.isotc211.org/protdoc/211n2029/>.
- Jain, A. (2007). "Mechanisms for validation of volunteer data in open web map services." Retrieved March 11 2010, [http://www.ncgia.ucsb.edu/projects/vgi/docs/supp\\_docs/Jain\\_paper.pdf](http://www.ncgia.ucsb.edu/projects/vgi/docs/supp_docs/Jain_paper.pdf).
- Jakobsson, A. and J. Giversen. (2007). "Guidelines for Implementing the ISO 19100 Geographic Information Quality Standards in National Mapping and Cadastral Agencies." Retrieved March 3 2010, [http://www.eurogeographics.org/documents/Guidelines\\_ISO19100\\_Quality.pdf](http://www.eurogeographics.org/documents/Guidelines_ISO19100_Quality.pdf).

## Quality Assurance of OpenAddresses

- Jakobsson, A. and L. Tsoulos (2007). The Role of Quality in Spatial Data Infrastructures. 23rd International Cartographic Conference, Moscow.
- Mäs, S., W. Reinhardt, A. Kandawasvika and W. F. (2005). Concepts for quality assurance during mobile online data acquisition. AGILE Conference on Geographic Information Science.
- Maué, P. and S. Schade (2008). Quality of Geographic Information Patchworks. 11th AGILE International Conference on Geographic Information Science, Girona.
- Messina, J. P., A. M. Shortridge, R. E. Groop, P. Varnakovidia and M. J. Finn (2006). "Evaluating Michigan's community hospital access: spatial methods for decision support." International Journal of Health Geographics 5(42).
- Miller, F. P., A. F. Vandome and J. McBrewster, Eds. (2009a). Google Maps. Mauritius, Alphascript Publishing.
- Miller, F. P., A. F. Vandome and J. McBrewster, Eds. (2009b). Google Street View. Mauritius, Alphascript Publishing.
- North American Association of central cancer registries. (2002). "Using Geographic Information Systems Technology in the Collection, Analysis, and Presentation of Cancer Registry Data: A Handbook of Basic Practices." Retrieved March 17 2010, <http://www.naaccr.org/filesystem/pdf/GIS%20handbook%206-3-03.pdf>.
- Novak, J. and B. Voigt. (2006). "Mashing-up mashups: from collaborative mapping to community innovation toolkits." Retrieved May 10, 2010, [http://www.ifi.uzh.ch/im/fileadmin/user\\_upload/personen\\_downloads/MCIS06.MashingUpMashups.Study.FullPaper.pdf](http://www.ifi.uzh.ch/im/fileadmin/user_upload/personen_downloads/MCIS06.MashingUpMashups.Study.FullPaper.pdf).
- Oort, P. A. J. v. (2006). Spatial data quality: from description to application. Wageningen, Wageningen Universiteit. **PhD**: 140.
- Piper, J. (2008). "Methoden zur kleinräumigen Modellierung von Versorgungsdisparitäten in der vertragsärztlichen Versorgung." GIS Science(4): 21 - 27.
- Purvis, M., J. Sambells and C. Turner (2006). Beginning Google Maps Applications with PHP and Ajax. Berkeley, Apress.
- Ramm, F. and J. Topf (2006). OpenStreetMap. Die freie Weltkarte nutzen und mitgestalten. Berlin, Lehmanns Media.
- 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.
- Schweizerische Normen-Vereinigung. (2004). "Vermessung und Geoinformation - Gebäudeadressen - Struktur, Georeferenzierung, Darstellung und Datentransfer."
- Servigne, S., N. Lesage and T. Libourel (2006). Quality Components, Standards, and Metadata. *Fundamentals of spatial data quality*. R. Devillers and R. Jeansoulin. London, ISTE: 179 - 210.
- Sinclair, S. (2007). "Free Licensed Geo-data. Floods Prove Need for Mapping Investment." *GIM International* **21**(12): 26 - 27.
- SNV. (2004). "Vermessung und Geoinformation - Gebäudeadressen - Struktur, Georeferenzierung, Darstellung und Datentransfer."
- Stark, H.-J. (2008). Open Geodata - am Beispiel von OpenAddresses.ch. *Angewandte Geoinformatik 2008*. Edited by Strobl, J., Blaschke, T., Griesebner, G. Heidelberg, Herbert Wichmann: 142 - 151.
- Stark, H.-J. (2009). OpenAddresses - Free geocoded street addresses. *Applied Geoinformatics for Society and Environment*, Stuttgart.
- Stark, H.-J. (2010). Umfrage zur Motivation von Freiwilligen im Engagement in Open Geo-Data Projekten. *FOSSGIS 2010*, Osnabrück.
- 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* **32**: 1 - 5.
- swisstopo. (2005). "Gebäudeadressierung und Schreibweise von Strassennamen für die deutschsprachige Schweiz." Retrieved March 25 2010, <http://www.bfs.admin.ch/bfs/portal/de/index/news/publikationen.Document.68586.pdf>.
- Tapscott, D. and A. D. Williams (2008). *Wikinomics*. New York, Penguin Group.
- TeleAtlas, D. (2008). "Tele Atlas address points." *GIS Trends + Markets*(6).
- Wales, J. (2005). Quotation by creator of Wikipedia in Dec. 5, 2005 interview for podcast of "The Writing Show". Podcast found at <http://writingshow.com/podcasts/2006/01012006.html>. Written transcript can be found at <http://www.writingshow.com/articles/transcripts/2006/01012006.html>.

- 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).
- Zandbergen, P. A. and T. C. Hart (2009). "Geocoding Accuracy Considerations in Determining Residency Restrictions for Sex Offenders." *Criminal Justice Policy Review* **20**(1): 62-90.
- Zimmerman, D. L., F. Xiangming, S. Mazumdar and G. Rushton (2007). "Modeling the probability distribution of positional errors incurred by residential address geocoding." *International Journal of Health Geographics* **6**(1): 1 - 16.
- Zipf, A. (2009). "Nutzungspotenziale und Herausforderungen von "Volunteered Geography". Zur Kombination von GDI-Technologie und nutzergenerierten Geomassendaten." *Wiener Schriften zur Geographie und Kartographie* **19**: 121 - 128.