INAUGURAL - DISSERTATION

zur
Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen
Gesamtfakultät
der
Ruprecht-Karls-Universität
Heidelberg

Vorgelegt von
Master of Science Pascal Neis
aus: Hünstetten Wallbach

Tag der mündlichen Prüfung: 17. Juni 2014

Thema

Analysis of User-generated Geodata Quality for the Implementation of Disabled People Friendly Route Planning

Gutachter: Prof. Dr. Alexander Zipf
Jun.-Prof. Dr. Bernhard Höfle

Abstract

Traditional geodata sources are most commonly used for personal navigation purposes tailored to motorized traffic or pedestrian needs. Oftentimes commercial or administrative data providers are utilized for these purposes. In more recent years, volunteers of the *OpenStreetMap* (OSM) project started to collaboratively collect information, an approach similar to the widely known Wikipedia project. The data is stored in a database and freely available to everyone via the Internet. Unlike the volunteers of the Wikipedia project, contributors of OSM generate geographic information to build a world map that can be edited and utilized by interested users. The term most commonly used for this type of information is *Volunteered Geographic Information* (VGI) and it has several advantages in comparison to their authoritative counterparts. The data is cost effective, available under a free or open license and can contain a variety of objects and attributes that are not included in the portfolio of other data providers. People with disabilities, for instance, require very detailed, highly accurate geodata specifications such as the street surface or sidewalk condition. Commercial geodata providers do not offer this detail of information due to the high costs that arise during the collection and the maintenance of the data. However, concerns also exist within the research community and potential VGI users, about the credibility and data quality of the freely available and user-generated geodata sources.

The main goal of this dissertation is the quality assessment of the data provided by the OSM project, specifically with regards to the potential implementation as a data source for a route planning application tailored to disabled people. For this purpose, the first quality analysis investigates the completeness of the collected OSM street network in comparison to a commercial data provider in Germany. The relative comparison revealed that for selected regions, in which the OSM project shows high contribution rates, the data can be of equal or better quality than the proprietary dataset. The conducted research also highlighted that the number of active and local project contributors strongly influences different data quality aspects, such as data density and currentness. Similar to other online communities that depend on voluntary contributions, an investigation of the OSM member activities has shown that only a

fraction of the total number of registered members actively collects information in a meaningful way.

Beside these first quality analyses, two additional applications for the quality assessment and assurance of the OSM project are proposed in this work. The first tool, an initial prototype, is based on a comprehensive rule-based methodology. It was successfully tested in a case study and protected the OSM project by detecting different vandalism types, such as deleting large chunks of data. The second framework that was developed in the scope of this dissertation is able to automatically evaluate quality measures of the OSM dataset, such as the attribute accuracy. Thus, the tool facilitates the decision whether the quality of OSM data in a selected area of interest is sufficient for the desired purpose or application.

Additionally, a number of methods are proposed that evaluate the potential of user-generated geodata, and in particular OSM, for a disabled people friendly route planning application. A newly developed algorithm generates a sidewalk routing network tailored to people with disabilities. This newly created network proved to have several advantages over traditional routing networks and is highly adaptable. However, the results also showed that the quality, and in particular the availability of detailed sidewalk information of the utilized OSM dataset, can highly influence the proposed generation of the sidewalk network. Furthermore, two approaches to the assessment and evaluation of the feasibility of a wheelchair user friendly routing algorithm and its generated path are introduced. The first method computes a tailored, individual path based on specific user requirements, while the second method evaluates the generated path by providing a reliability factor based on the utilized data. Thereby the proposed reliability factor can give a direct feedback to the user if the required information is available and to what degree the generated path can be trusted.

The results of the conducted research in this dissertation highlight the potential of VGI data collections. The OSM project has the chance to be a public database for different types of spatial datasets including detailed information for disabled people. Nevertheless, it has been shown that it is still important to evaluate whether the OSM data is acceptable for each use case. There is no reliable estimation if a certain object or other detailed attribute information is included in OSM. Lastly, the introduced quality analyses methods and designed applications enable new possibilities for future research in different fields of geodata quality assessment and assurance of user-generated geoinformation and disabled people friendly route planning.

Kurzfassung

Häufig werden herkömmliche Geodatenquellen für die persönliche Navigation im Auto oder zu Fuß eingesetzt. Üblicherweise kommen dabei Daten von kommerziellen oder öffentlichen Anbietern zum Einsatz. In den letzten Jahren haben viele Freiwillige des *OpenStreetMap* (OSM) Projektes damit begonnen, in einem Wikipedia ähnlichen Ansatz, gemeinsam Informationen zu sammeln, sie in einer Datenbank zu speichern und über das Internet für jeden frei zugänglich zu machen. Anders als bei Wikipedia, sammeln die Mitwirkenden des OSM Projektes geographische Informationen, um eine frei verfügbare Karte, die Interessierte ändern und nutzen können, zu ermöglichen. Solche, durch freiwillige generierte Geodaten, auch *Volunteered Geographic Information* (VGI) genannt, besitzen einige Vorteile gegenüber ihrer autoritären Gegenstücke. Die Daten sind beispielsweise kosteneffektiv, unter einer freien Lizenz verfügbar und können eine Vielzahl von verschiedenen Objektarten und Attributen beinhalten, die nicht im Portfolio von herkömmlichen Geodatenanbietern zu finden sind. Menschen mit Einschränkungen haben beispielsweise sehr detaillierte Anforderungen an Geodaten. Sie benötigen genaue Informationen über die Beschaffenheit der Straßenoberfläche oder der Bürgersteige. Kommerzielle Geodatenanbieter können diese detaillierten Informationen aus Kostengründen, die beim Erheben oder bei der Wartung entstehen, nicht anbieten. Dennoch existieren Bedenken in der Forschungsgemeinschaft und bei potentiellen Nutzern über die Verlässlichkeit und die Qualität der frei verfügbaren und gemeinschaftlich zusammengetragen Geodaten.

Das Hauptziel dieser Arbeit ist die Qualitätsbewertung der Geodaten des OSM Projektes als Datenquelle für die potentielle Nutzung in einer adäquaten Routenplanung für Menschen mit Einschränkungen. Für diesen Zweck analysiert die erste Qualitätsuntersuchung die Vollständigkeit des zusammengetragenen Straßennetzwerkes des OSM Projektes für Deutschland im Gegensatz zu einem kommerziellen Anbieter. Der relative Vergleich bestätigt, dass in Regionen wo sich das OSM Projekt bereits gut entwickelt hat, die Daten vergleichbar oder besser als ein kommerzieller Datensatz sind. Die durchgeführten Untersuchungen zeigten auch, dass die Anzahl von aktiven und lokalen Mitwirkungen stark die Qualität, bezogen auf Aktualität und Datendichte, beeinflussen

kann. Eine Studie über die Mitwirkenden des OSM Projektes deckte auf, dass, ähnlich wie bei anderen Online-Projekten, lediglich ein kleiner Teil aller registrierten Mitglieder aktiv und in einer bedeutenden Art dem Projekt etwas beitragen.

Neben diesen und weiteren Analysen wurden zwei Anwendungen für die Bewertung und die Sicherung der Qualität des OSM Projektes vorgestellt. Die erste Anwendung ist ein Prototyp, für den eine umfangreiche regelbasierte Methodik entwickelt wurde. In einem Testzeitraum konnte der Prototyp erfolgreich das OSM Projekt vor unterschiedlichen Arten von Vandalismus schützen, beispielsweise vor dem flächenmäßigen Löschen von Daten. Das zweite entwickelte Framework im Rahmen dieser Dissertation kann für verschiedene Qualitätsuntersuchungen verwendet werden, wie zum Beispiel der Attributgenauigkeit. Dabei kann es bei der Entscheidungsfindung helfen, ob die Qualität eines OSM Datensatzes in einer ausgewählten Region für ein bestimmtes Vorhaben oder eine bestimmte Anwendung geeignet ist.

Des Weiteren wurden mehrere Methoden vorgestellt, die das Potential der Daten des OSM Projektes für die Umsetzung einer Routenplanung für Menschen mit Einschränkungen evaluieren. Ein neu entwickelter Algorithmus generiert einen maßgeschneiderten Bürgersteig Routing Graphen für eingeschränkte Menschen. Das vorgestellte Bürgersteignetzwerk bestätigt, dass es mehrere Vorteile gegenüber traditionellen Straßennetzwerken besitzt und dass es erweiter- und anpassbar ist. Die Ergebnisse zeigten dennoch, dass die Qualität des verwendeten OSM Datensatzes und die Verfügbarkeit der detaillierten Bürgersteiginformationen die Erstellung des vorgestellten Routing Graphen stark beeinflussen können. Weiterhin wurden zwei Ansätze für die eigentliche Routenplanung und die Bewertung der berechneten Route entwickelt. Die erste Methode ermittelt eine individuelle und den Bedürfnissen des Nutzers angepasste Route, während die zweite vorgestellte Methodik die ermittelte Route evaluiert und einen Verlässlichkeits-Faktor berechnet. Der vorgestellte Verlässlichkeits-Faktor gibt dem Nutzer ein direktes Feedback darüber, ob die erforderlichen Informationen verfügbar sind und zu welchen Grad der berechneten Route vertraut werden kann.

Die Ergebnisse von den durchgeführten wissenschaftlichen Untersuchungen dieser Arbeit zeigen das Potential der von Freiwilligen generierten Datensammlungen. Das OSM Projekt hat die Chance, eine öffentliche Datenbank für verschiedene Geodaten zu werden und detaillierte Informationen für Menschen mit Einschränkungen bereitzustellen. Nichtsdestotrotz wurde gezeigt, dass es wichtig ist, die Geodaten im Vorfeld und für den gewünschten Anwendungsfall zu evaluieren. Es gibt keine verlässlichen Aussagen darüber, ob ein bestimmtes Objekt oder die gewünschten Attributinformationen in OSM vorhanden sind. Letztendlich eröffnen die vorgestellten Qualitätsanalysen und

die entwickelten Anwendungen neue Möglichkeiten für zukünftige wissenschaftliche Untersuchungen in unterschiedlichen Feldern der Qualitätsbewertung und -sicherung von nutzergenerierten Geodaten und Routenplanern für Menschen mit Einschränkungen.

Contents

List of Figures

List of Tables

Part I.

Synopsis

1. Introduction

1.1. Motivation

Geographic information and its multi-propose applicability led to the development of a plethora of applications with a sheer unlimited potential for the future. The information is implemented in widely used Navigation Systems, maps and nowadays also in 3D applications. Overall it has become an important and sometimes essential part in our daily life. Authoritative geographic information was traditionally utilized for all types of platforms, services, tools, apps or printed products where location-based information plays an important role. The data created by administrative and commercial data providers also follows strict quality specifications and standards. People with special needs, however, who rely on a more specialized dataset, cannot utilize the provided proprietary geo-information and require highly detailed ground-truth data. A designated dataset tailored to these special needs can be costly, contain licensing restrictions and, most importantly, it can occur that a particular type of information is unavailable for the area of interest or the individual use case.

The Internet underwent a tremendous change in recent years, which lead to the introduction of term Web 2.0 (O'Reilly 2005), describing the Internet user's change in behavior from a passive consumer of information to an active contributor of content. Some of the well-known examples in the realm of Web 2.0 are Wikipedia, Youtube and Flickr, platforms that allow Internet users to collaboratively collect and share information such as videos or images. The main idea behind Wikipedia in its early stages in 2000 was the creation of a free online encyclopedia written by experts and reviewed by other participants. This approach has been enhanced in 2001, allowing anyone to contribute or edit information to the project in form of an article. This type of crowed sourced information is oftentimes also referred to as *User Generated Content* (UGC; Anderson 2007). The Wikipedia project demonstrates the potential of UGC with millions of volunteers that created a free Internet encyclopedia with more than 30 million articles in almost 290 languages. Based on similar ideas but different needs, several other UGC based online projects in form of blogs or forums were established. A

special type of UGC is *Volunteered Geographic Information* (VGI). The term, coined by Goodchild in 2007, describes collaboratively collected geographic information, provided by volunteers in a *World Wide Web* (WWW) repository (Goodchild 2007).

Different technological developments had a large impacted on the rise of the VGI phenomena. In 2000, *Global Positioning System* (GPS) technology was enabled to the public without "*Selective Availability*" and GPS enabled handhelds started to be available for reasonable prices. Additionally, new mobile phones were equipped with a GPS receiver to determine the phone's geographic position. Another important factor was the ubiquity of broadband Internet connections, at least in developed countries, allowing an increase in Internet users and faster access to content on the WWW.

Several VGI platforms such as Google Map Maker, TomTom's MapShare or Wikimapia were released between 2004 and 2008. However, in more recent years the *OpenStreetMap* (OSM) project has grown in popularity and is oftentimes cited as the most successful VGI. The idea behind the OSM project is comparable to Wikipedia: the creation of a central database with information that everyone can create, modify, correct, delete and in particular access. Due to the lack of official standards and quality assurance measures, projects such as OSM oftentimes rely on the large number of contributors to reduce errors in the data. Similar to Linus's Law, which Raymond (1999) described as a claim about open software development: "given enough eyeballs, all bugs are shallow". In contrast to the aforementioned other VGI projects such as Google or TomTom, the provided geographic dataset of the OSM project is freely available to Internet users. Based on the open approach to data contributions in OSM, every member can add any geographic object to the database or define new object tagging proposals. This approach has different assets and drawbacks. On the one hand, as already criticized by Brando and Bucher (2010) and Girres and Touya (2010), well defined object types or guidelines would improve the data quality and usability. On the other hand this would increasingly limit the possibility of data contributions tailored to a member's specific needs and as a consequence also limit the geographic datasource for applications with special requirements. For instance the Wheelmap.org project demonstrated in recent years that volunteers are willing and able to mark locations with wheelchair friendly environments or accessibility in OSM. Additionally, the humanitarian component in OSM played a major role during different types of natural disasters or political conflicts, during which a large number of people contributed information in so-called "crisis mapping" efforts (Roick and Heuser 2012). However, due to the fact that VGI data is mostly contributed by amateurs or by people with no special geography related education or training (Goodchild 2008, Haklay 2010), early concerns

about the credibility and quality of VGI arose (Flanagin and Metzger 2008). Haklay (2010) and Zielstra and Zipf (2010) revealed in first studies that OSM data could potentially be utilized for mapping applications at least in urban areas. Additionally both authors pointed out that OSM data was not an alternative or replacement for commercial, proprietary or administrative geodata products due to the lack of information in rural areas. They highlighted that OSM data in metropolitan regions can have very detailed and up-to-date information where contributors mapped objects such as streets, buildings or public transportation information. More importantly, they also started to collect very detailed information about street surfaces or sidewalks such as width or incline, data that cannot be found in proprietary datasets.

The local knowledge of an individual helps to find the shortest or fastest path in familiar places on a day to day basis. Routing applications can help to experience a similar situation in unfamiliar areas. Disabled people rely on very detailed information about potential obstacles in their neighborhood or in areas in which their daily life takes place. However, when visiting unknown areas mainstream routing applications, tailored to motorized traffic, do not provide the detailed information needed. Depending on the requirements of the user, information about sidewalks, steps, surface conditions, crossings or tactile paving could be essential and heavily improve the routing experience of a disabled person. Common authoritative data providers only produce network data for motorized vehicles and maybe pedestrian information for selected areas. Due to high personnel or data maintenance costs, the providers are not able to offer such detailed information. OSM on the other side can provide this detailed information for people with disabilities as long as contributors are willing to collect the data. Prior research focusing on the development of routing or navigation solutions for disabled people (Beale et al. 2006, Kasemsuppakorn and Karimi 2008, Kammoun et al. 2010), oftentimes created an individual, none interoperable and sometimes solely use case oriented network dataset. VGI projects and in particular OSM, has a high potential to be a public and central database for such desired spatial datasets with its corresponding attributes. In comparison to proprietary dataset providers, the OSM project has the significant advantage that contributors can simply add new objects to the database tailored to their needs. Furthermore, the active VGI community can lead to a high currentness of the collected geospatial information. On the other hand a VGI data consumer should always be aware about the credibility or heterogeneous quality of the collected data.

1.2. Research methods and objectives

The scope of this dissertation includes the assessment of user generated spatial data for the development of a route planning application tailored to people with disabilities. Figure 1.1 illustrates the principle workflow from spatial data production to utilization.

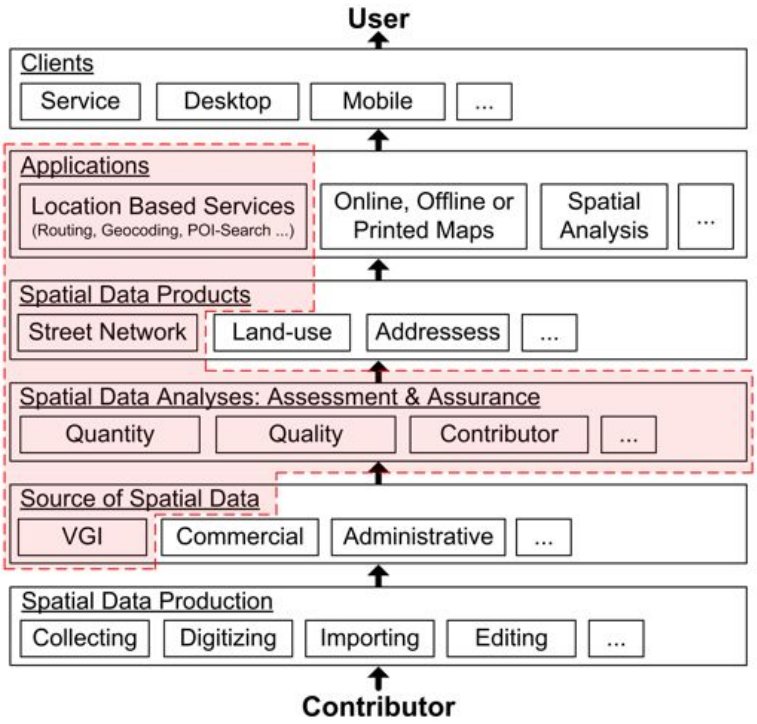


Figure 1.1.: Principle workflow from spatial data production to utilization. Scope of research highlighted in red.

Commonly, contributors produce spatial data by collecting GPS tracks, tracing satellite imagery or importing or conflate different data sources. It is important to distinguished for which purpose the data is being collected. The local knowledge (Goodchild 2007, Mooney and Corcoran 2013) of a contributor can play a major role, especially with regards to detailed information and attributes such as street names, turn restrictions or different types of traffic signs. Spatial data analyses are an important task after the data production and help to evaluate whether a specific VGI dataset meets the requirements of a particular application or final product. Depending on the type of the spatial dataset, different product types are feasible. For the purpose of this dissertation, a street network with special requirements is needed, that allows for the development of a routing or navigation application for people with disabilities. Ideally

the application can be utilized by different client types such as a desktop computer, mobile device or in any type of webpage. However, although the final development of an easily operational software application for navigation purposes is an interesting aspect, it is not part of this dissertation. Instead, Figure 1.1 depicts the main scope of the research (highlighted in red) within the workflow.

The following sections contain a summary of the most important methods that built the foundation of the conducted research. It starts with a brief introduction to VGI and why this particular data source was chosen for the research project. This is followed by different spatial data and contributor analyses that assess the quality of the VGI source. The last sections give a general overview about algorithms for the routing network generation and the path computation for disabled people.

1.2.1. Collaboratively collected geographic information

The development of a route planning application for wheelchair users or other disabled people heavily relies on geodatasets that include special attributes and details. As discussed in a prior section of this dissertation, these detailed datasets are usually not available from any authoritative data provider. In recent years, researchers performed extensive surveys to generate adequate datasets for their needs and research tasks. VGI data, however, as a new web phenomenon, could play a major role in this particular case. Only a few studies about the data quality of VGI have been conducted and published in scientific journals. Nevertheless, the general pattern that was determined in most of these prior studies was that VGI datasets and especially OSM has great potential to be a reliable data source in the near future. The studies also highlighted that further research is needed in the fields of VGI data quality analysis and assessment.

The OSM project was selected for this dissertation due to several reasons. First of all, the project has a highly active community. The total number of registrations has almost tripled between 2012 and 2013, generating more than 1.4 million registered members who try to create and distribute "free geographic data for the world" (OpenStreetMap 2014d). A general introduction to OSM can be found in Ramm et al. (2010), Bennett (2010) and in Chapter 12.2.2. Many variables can play a major role in the success of a VGI project. Besides a working infrastructure, the crowd of volunteers is the most important factor of an online VGI project. They contribute new data and keep existing data up to date. Another reason for the implementation of OSM in this dissertation is availability of the collected data. It is freely available, under certain licensing conditions, to any Internet user (OpenStreetMap 2014c). The data can be retrieved in different ways (e.g. via different *Application Programming Interfaces*

(APIs) or complete planet dump file) and datatypes (e.g. *Extensible Markup Language* (XML), *Protocolbuffer Binary Format* (PBF) or preprocessed shapefiles) for many different timeframes. Additionally, prior studies and projects have demonstrated that the collected information can be utilized for many applications such as different types of *Location based Services* (LBS; Schmitz et al. 2008, Neis and Zipf 2008), 3D applications (Schilling et al. 2009), disaster management (Neis et al. 2010) and, particularly important for this dissertation, in different types of routing or navigation services for wheelchair users (Holone et al. 2007, Müller et al. 2010).

Similarly to other online communities, the OSM project is referred to as a "do-ocracy". This term describes the organizational structure in which project members, who are responsible for different types of tasks, have to decide or choose how the work is being accomplished. This includes for instance the development of software-tools or APIs to access the project data. Another essential reason why OSM was utilized in this project is the openness and freedom that every contributor can map any object type or information of interest that has a relation to a geographic position in the real world. Other VGI platforms such as Google Map Maker or TomTom's MapShare only allow contributors to add objects that fit to the data standards or specifications of the platform owner. Next to the fact that these data sources are not publicly and freely available, these strict mapping and tagging limitations made other data sources inapplicable for the objectives of this dissertation. Additionally, it is more likely that a VGI dataset has been updated more frequently than its authoritative counterparts, due to the highly active community. VGI datasets can also be a very cost-effective approach due to the aforementioned aspects.

One of the main and fundamental objectives of this dissertation is to conduct several comprehensive quality and contributor analyses in the field of VGI. Therefore the first aim was to develop different software tools that retrieved, read, interpreted and finally analyzed the collaboratively collected geographic information and their contributors. Particularly, different file types, data types and APIs had to be considered during the development of the tools.

1.2.2. Spatial data quality analyses

Whether a specific routing application can be developed or not, oftentimes heavily depends on the quality of the spatial dataset at hand. Several research publications in the past two or more decades have been investigating the problem of spatial data quality (Veregin 1999, Van Oort 2006, Devillers et al. 2010). According to Devillers and Jeansoulin (2010) the different quality factors can be grouped into internal and

external factors. In most cases the evaluation of the external quality issue is based on a single question: "Can the dataset at hand be utilized for the particular use case and area of interest or not?". Chrisman (1983) and Veregin (1999) describe this quality assessment as the evaluation of the "fitness for use" of the dataset, whereas others refer to the notion of "fitness for purpose" (Devillers and Jeansoulin 2010). Both terms have in common that the dataset is tested from a consumer point of view. In contrast, internal quality evaluates the dataset from the producer's point of view, with factors such as completeness, logical consistency, positional accuracy, temporal accuracy and thematic accuracy. Additionally, the *International Organization of Standardization* (ISO) created a set a standards that define the quality attributes of geodata in ISO 19113 (principles for describing the geographic dataset quality) and ISO 19114 (procedural framework for evaluating geographic dataset quality). By the end of 2013, both ISO standards (19113 and 19114) have been aggregated to one single standard: ISO 19157:2013 (geographic information data quality). Veregin (1999) stresses, that the way in which data is produced directly affects its reliability and thus its general quality. In the case of VGI these factors can play a major role, due to open approach to data contributions that allows amateurs and non-professionals to add information to the platform (Goodchild 2007, Coleman et al. 2009). This can lead for instance to errors in completeness, i.e. overcompleteness (error of commission) or incompleteness (error of omission) in the geometry representation of a feature.

The second objective of this thesis is to examine and assess the quality of VGI data regarding the aforementioned data quality elements. The first step in this process focuses on a relative comparison between the collected VGI street network of OSM and an authoritative dataset to determine its completeness and thus its usability. Further, the analysis evaluates discrepancies in the development of urban and rural areas over a specific timeframe and internal logical consistency tests. Additionally, some novel methods are introduced that allow for the intrinsic spatial data quality analysis of OSM data based on the data's historical development and contribution behavior.

1.2.3. Contributor analyses

The quality of a created VGI object can be highly influenced by its producer. Therefore it is essential to evaluate who contributes geographic information to a VGI project. Once this information has been retrieved, it can be utilized for a number of different data quality assessments or inspections. There are two ways to gather additional information about VGI contributors, in form of extensive surveys or by investigating the contributions of a project member. The first scenario is oftentimes conducted through

extensive interviews or questionnaires and several analyses utilizing this approach, investigating contributor motivation (Budhathoki and Haythornthwaite 2013, Lin 2011) and gender dimensions (Stephens 2013, Steinmann et al. 2013), have been published. The results have shown that possible motivational factors for VGI contributors are: the idealist's approach, that geospatial information should be freely available to everyone, learning new technologies, self-expression, relaxation and recreation or just pure fun (Budhathoki 2010). Similar factors can also be found in other online communities and platforms such as Wikipedia, Youtube or Flickr. However, three different surveys (Budhathoki 2010, Stark 2010, Lechner 2011) have shown that the majority (97%) of the OSM project members are males. In addition 65% of the survey respondents were between the age of 20 and 40 (Budhathoki 2010, Lechner 2011). Furthermore 50% considered their profession as "computer science related" (Lechner 2011) or had some sort of GIS background (Budhathoki 2010). This contradicts the aforementioned assumption that VGI data is contributed by non-professionals without any specialized education (Goodchild 2007). One of the caveats of the conducted questionnaires is the limited number of interviewees or responses that are being received. However, as demonstrated in other studies about Wikipedia (Yasseri et al. 2012) or mobile phone users (Jo et al. 2012), the contributed information, such as Wikipedia articles, created by the members, can also be utilized for comprehensive contributor or user analyses. A first investigation of the OSM members (Budhathoki 2010) revealed that the community shows patterns of the so called "Participation Inequality" phenomenon as described by Nielsen (2006). His "90-9-1" rule stresses, that most projects that are based on an online community show a distinct pattern, where 90% of the members of the community only consume and never contribute any data, 9% create some minor content and only 1% of the members contribute the majority of the project data. Anthony et al. (2007) and Javanmardi et al. (2009) already stated that this rule can be applied to projects such as Wikipedia.

The third research objective of the dissertation is to develop a number of methods to analyze the contributor activity area of a VGI project member based on the collected data. Detailed contributor information about their editing behavior, home location, and activity area or activity timeframe can be extremely useful in the field of (intrinsic) quality assessment and assurance such as vandalism detection.

1.2.4. Generation of routing networks for disabled persons

The main goal of most routing applications is to generate the "best" path between two points based on an available graph represented by a street network. Car navigation

applications for instance need a sophisticated street network to achieve the best results. Usually this type of information is provided by commercial or administrative data providers that generate extensive network datasets. Usually the datasets follow certain data standards or provider related specifications, but mostly lack the detail of information that is needed for the aim of this dissertation. Chen and Walter (2009) demonstrated that the OSM street network data by default could not be utilized for routing applications. Schmitz et al. (2008) introduced the steps that are needed to create a routable network dataset from OSM data that is comparable to commercial data providers. In the case of the development of a route planning application for disabled people, a pedestrian network with additional requirements is essential. Especially information about sidewalk surface texture, street width or the position of steps is crucial (Matthews et al. 2003, Sobek and Miller 2006, Kasemsuppakorn and Karimi 2009). Similar to the ISO standards for the quality of spatial data, a standard specification by the German Institute for Standardization (*Deutsches Institut für Normung* (DIN)) provides a foundation of information for the accessibility requirements for disabled people in public transit infrastructure and buildings (DIN 18024-1 1998). Additionally the United States' *Americans with Disabilities Act* (ADA) standard for Accessible Design (ADA 2010) also sets minimum requirements for facilities and their environments to be accessible and useable by individuals with disabilities. Besides these standards, different research projects dedicated to routing specifications for disabled people, such as wheelchair users, blind, deaf or elderly people have been published (Sobek and Miller 2006, Kasemsuppakorn and Karimi 2009, Kammoun et al. 2010). The terminology used to describe the target user group of disabled people can vary. However, the aforementioned DIN 18024-1 standard also helps in this case to describe the individuals with disabilities in more detail:

- Wheelchair users
- Blind and visually impaired people
- Deaf and hearing impaired people
- Walking impaired people
- People with other handicaps
- Elderly people
- Children and people of short or tall statue

Different methods have been introduced regarding the creation of a tailored network for pedestrians that can build the foundation of a more sophisticated routing

graph. Some research projects traced the required sidewalk information from areal imagery (Beale et al. 2006, Kammoun et al. 2010), used pedestrian GPS traces (Kasemsuppakorn and Karimi 2013) or developed some binary image processing procedures (Gaisbauer and Frank 2008, Kim et al. 2009). Additionally, an adequate route planning for disabled people needs detailed information about traffic signals or crossings. Both intersections shown in Figure 1.2 would not include special considerations for people with disabilities if the basic geometric sidewalk network would be utilized. Due to the availability of additional attributes in the OSM dataset a more sophisticated and personalized route computation is possible.

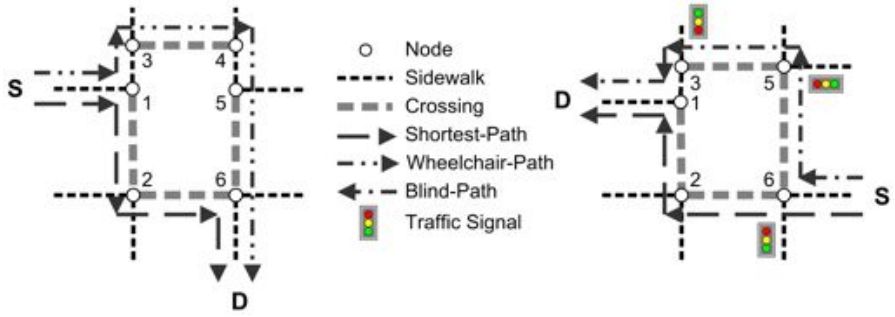


Figure 1.2.: Intersection examples for a wheelchair user (left) & for a blind person (right).

The left example in Figure 1.2 shows an intersection with four crossings. The general approach for a shortest-path X for pedestrians between a start S and destination D node would be represented by $X = \{S, 1, 2, 6, D\}$. The more detailed sidewalk information of Node 2 includes a curb height of 10 cm, thus this Node and the crossings between Nodes 1 & 2 and 2 & 6 are not passable for a wheelchair user. Consequently, the routing algorithm that considers these parameters, determines the "best" path for a wheelchair user as $X = \{S, 1, 3, 4, 5, 6, D\}$. The newly computed way for a wheelchair user is a bit longer, but does not contain any barriers or obstacles, which she/he cannot pass. The second example in Figure 1.2 (right) is a route that was generated for a blind person. The shortest path for a pedestrian would be $X = \{S, 6, 2, 1, D\}$. After including all relevant information about crossings, traffic lights and maybe acoustic signals at junctions, a routing algorithm for blind people would result in the path $X = \{S, 6, 5, 3, 1, D\}$, which is again longer but also contains two crossings with traffic lights, in comparison to the shortest-path which has only one "controlled" crossing. Both examples shown in Figure 1.2 only demonstrate a few new implementations that

would be applicable with the newly derived network for disabled people. In the context of VGI and in particular OSM, only a minor amount of studies focused on the usage of collaboratively collect geographic information for routing of wheelchair users (Holone et al. 2007, Müller et al. 2010).

For the implementation of a route planning application for disabled people the fourth objective focuses on the demonstration of how VGI geodata can be utilized for the generation of a tailored sidewalk network representation. Furthermore, the quality of the generated sidewalk routing graph, based on VGI data, will be tested and evaluated.

1.2.5. Graph-based route planning

The implementation of a network that represents the real world, allows the generation of a variety of routes based on specific criteria, such as shortest- or fastest route, most scenic or most ecological route. One of the main goals of this dissertation is to support disabled people by finding their "best" path based on their specific needs. However, a number of routing algorithms, such as Floyd–Warshall (Floyd 1962) or Bellman–Ford (Bellman 1958) exist for this purpose. The most common and utilized routing algorithm is the Dijkstra (2010) algorithm which was extended into the A* algorithm (Dechter and Judea 1985) with better performance due to the implementation of heuristics during the path computation. Besides the main routing algorithm, several techniques exist to speed-up time dependent route planning (Delling and Wagner 2009, Bauer et al. 2008). Nowadays in many cases the contraction hierarchies method (Geisberger et al. 2008), which creates a contracted versions of the routing graph in a preprocessing step, is applied to boost the route generation. However, due to the fact that disabled people require a multi-criteria network where route preferences can change dynamically based on the person's needs, these advanced speed-up methods for shortest-path computation are inapplicable. In recent years a number of studies dedicated to the development of a wheelchair friendly route planning have been published (Matthews et al. 2003, Beale et al. 2006) and showed that the routing process is much more complex than for more common car or pedestrian related purposes. Especially way properties, such as the width, incline or surface texture, can play a major role. Thus, not only the routing algorithm itself but in particular the weighting methods play a crucial role. Different weighting methods based on user surveys' (Matthews et al. 2003) or way properties (Kasemsuppakorn and Karimi 2009) have been introduced in the literature.

The fifth research objective of the dissertation focuses on the generation of a VGI sidewalk network for a wheelchair user route planning application. New methods that

implement a VGI dataset with a heterogeneous data quality needed to be developed for this purpose. This fifth and last objective of the dissertation can be seen as a synopsis of all previously introduced objectives. It combines a VGI dataset and evaluates its quality for a route planning application for disabled people and computes a reliability factor of the computed path that gives a direct feedback about the quality and quantity of the generated path based on the utilized VGI dataset.

1.3. Objectives and research questions

The main objective of this dissertation is to introduce the development of a routing application for people with disabilities based on collaboratively collected and freely available geographic information retrieved from the OSM project. Therefore, the first focus lies on the analysis of the spatial data quality and its contributor behavior to get meaningful results which help to improve the understanding of how VGI is being generated and which quantity and quality can be expected. Due to the new opportunities that arise with the availability of a detailed but sometimes questionable data source, new methods for the utilization of VGI in this special routing application are needed.

The dissertation's goal is to answer the following research questions:

1. How can the applicability of a VGI routing friendly street network dataset be evaluated?
2. Does the quality of VGI depend on factors such as contributor concentration, activity, population density or socio-economic parameters?
3. What methods are required to protect a VGI platform from vandalism?
4. Does the history of collected VGI data reflect different spatial data quality parameters?
5. How can VGI be used for the generation of a tailored routing network for disabled people?
6. Can VGI data quality parameters be utilized during a route computation and for the quality assessment of the computed path?

1.4. Dissertation outline and selected publications

The presented cumulative dissertation is divided into two sections: (I) Synopsis and (II) Publications. The Synopsis includes a brief introduction into the dissertation's objectives and research questions. Chapter 1.2 of the Synopsis introduces the utilized

methods implemented during the research process on a general and informative level. The core findings of each publicized research project are summarized and discussed in Chapter 2. It needs to be noted that some of the publications include additional results, which are not included in this chapter. The Synopsis ends with a conclusion of the dissertation (Chapter 3) and an outlook on future work (Chapter 4).

The second section of the dissertation contains eight peer-reviewed scientific journal publications. The first publication investigates the contributors of the OSM project. The quality of the OSM street network for Germany is analyzed in a comprehensive way in the second publication. Additionally, the third publication compares 12 selected world regions with regards to the amount of collected OSM data and the activity of the community. As a part of quality assurance, the fourth publication introduces a vandalism detection prototype for VGI projects. A novel and comprehensive framework for the quality assessments of VGI objects based on the history of the collected information is introduced in publication number five. The sixth publication introduces a newly developed algorithm that generates a disabled people friendly sidewalk network based on OSM data. The seventh publication proposes a new wheelchair routing method based on OSM data and the calculation of a reliability factor of the generated path. Finally, an extensive overview about the developments and future trends of VGI and in particular OSM research is presented in the eighth publication. It has to be clarified that Pascal Neis is the main author of seven out of the eight publications. Christopher Barron is the main author of the fifth publication and Pascal Neis contributed the main idea and design for the implementation of the comprehensive framework for intrinsic OSM data quality analysis.

The following list gives an overview of the selected publications that are the foundation of the cumulative dissertation:

1. Neis, P. and Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(2):146-165.
2. Neis, P., Zielstra, D., and Zipf A. (2012). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4(1):1-21.
3. Neis, P., Zielstra, D., and Zipf A. (2013). Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Future Internet*, 5(2):282-300.
4. Neis, P., Goetz, M., and Zipf, A. (2012) Towards Automatic Vandalism Detection

in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(3):315-332.

5. Barron, C., Neis, P., and Zipf, A. (2013). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, doi:10.1111/tgis.12073
6. Neis, P. and Zielstra, D. (2014). Generation of a Tailored Routing Network for Disabled People based on Collaboratively Collected Geodata. *Applied Geography*, 47:70–77.
7. Neis, P. (2014). Measuring the Reliability of Wheelchair User Route Planning based on Volunteered Geographic Information. *Transactions in GIS*, (accepted).
8. Neis, P. and Zielstra, D. (2014). Current Developments and Future Trends in VGI Research: The Case of OpenStreetMap. *Future Internet*, 6(1):76-106.

Figure 1.3 illustrates how the selected publications are related to the outline of the principle workflow of the dissertation from spatial data production to utilization (Figure 1.1).

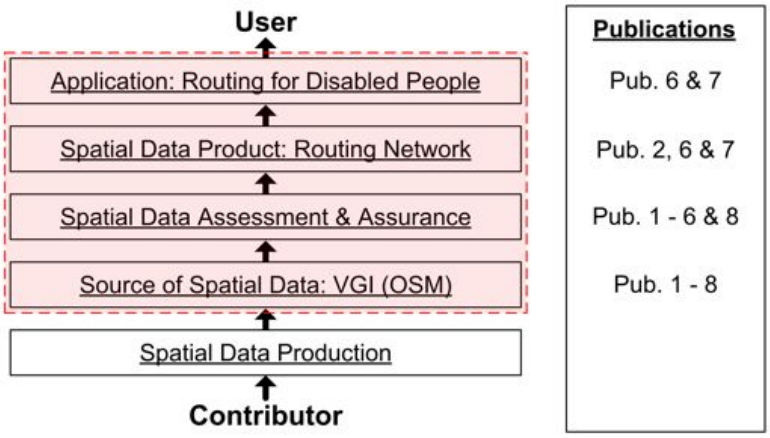


Figure 1.3.: Relation of published articles to simplified principle workflow from spatial data production to utilization.

1.5. Additional publications

Next to the aforementioned publications, a number of additional research projects were conducted and published during the preparation of this dissertation. Most of them tackled problems regarding the quality assessment of VGI data and are listed below.

1. Neis, P. (2014). Von Qualitätsuntersuchungen zu Nutzungspotentialen von gemeinsam zusammengetragenen Geodaten. *Kartographische Nachrichten - Journal of Cartography and Geographic Information*, (submitted).
2. Klonner, C., Barron, C., Neis, P., and Höfle, B. (2014). Updating Digital Elevation Models via Change Detection and Fusion of Human and Remote Sensor Data in Urban Environments. *International Journal of Digital Earth*, (accepted).
3. Hochmair, H.H., Zielstra, D., and Neis, P. (2014). Assessing the Completeness of Bicycle Trail and Designated Lane Features in OpenStreetMap for the United States. *Transactions in GIS*, (accepted).
4. Fan, H., Zipf, A., Fu, Q., and Neis, P. (2013). Quality assessment for building footprints data on OpenStreetMap. *International Journal of Geographical Information Science*, (accepted).
5. Zielstra, D., Hochmair, H.H., and Neis, P. (2013). Assessing the Effect of Data Imports on the Completeness of OpenStreetMap – A United States Case Study. *Transaction in GIS*, 17, 315–334.
6. Barron, C., Neis, P., and Zipf, A. (2013a). iOSMAnalyzer – ein umfassendes Werkzeug für intrinsische OSM Qualitätsuntersuchungen. In: *Proceedings of 25. AGIT Symposium für Angewandte Geoinformatik*. (July 3–5, 2012). Salzburg, Austria.
7. Barron, C., Neis, P., and Zipf, A. (2013b). Mappping the Crowd – zur Rolle der Mapper bei der Qualitätsanalyse von OpenStreetMap. In: *Proceedings of 25. AGIT Symposium für Angewandte Geoinformatik*. (July 3–5, 2012). Salzburg, Austria.
8. Barron, C., Neis, P., and Zipf, A. (2013c). Towards intrinsic Quality Analysis of OpenStreetMap Datasets. In: *Online Proceedings of the International Workshop on Action and Interaction in Volunteered Geographic Information (ACTIVITY) In: 16th AGILE Conference on Geographic Information Science*. (May 14–17, 2013). Leuven, Belgium.
9. Neis, P. and Stark, H.J. (2013). Crowdsourcing im Katastrophenfall – Am Beispiel OpenStreetMap. In: *Proceedings of 18. Münchner Fortbildungsseminar Geoinformationssysteme – Runder Tisch GIS e.V.* (Apr. 8–11, 2013). München, Germany.
10. Hochmair, H.H., Zielstra, D., and Neis, P. (2013). Assessing the Completeness of Bicycle Trails and Designated Lane Features in OpenStreetMap for the United

States and Europe. In: *Proceedings of the the Transportation Research Board - 92nd Annual Meeting*. (Jan. 13–17, 2013). Washington, DC, USA.

11. Neis, P. (2012). OpenStreetMap in D-A-CH 2012 – Welche Auswirkungen hat(te) der Lizenzwechsel auf den Datenbestand? In: *Proceedings of 24. AGIT Symposium für Angewandte Geoinformatik*. (July 4–6, 2012). Salzburg, Austria.
12. Helbich, M., Amelunxen, C., Neis, P., and Zipf, A. (2012). Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. In: *Proceedings of GI Forum 2012: Geovisualization, Society and Learning*. (July 4–6, 2012). Salzburg, Austria.
13. Roick, O., Neis, P., and Zipf, A. (2011). Volunteered Geographic Information – Datenqualität und Nutzungspotentiale am Beispiel von OpenStreetMap. In: *Symposium 2011 - Kommission „Angewandte Kartographie – Geovisualisierung" der Deutschen Gesellschaft für Kartographie (DGfK)*. (May 30–June 1, 2011). Königslutter, Germany.
14. Neis, P. and Walenciak, G. (2011). Zur Nutzung von TMC Verkehrsmeldeinformationen mit OpenStreetMap. In: *Proceedings of 23. AGIT Symposium für Angewandte Geoinformatik*. (July 6–8, 2011). Salzburg, Austria.
15. Neis, P. (2010). Qualität in Volunteered Geographic Information? – Am Beispiel Routing mit OpenStreetMap. In: *Proceedings of 7. GI/KuVS-Fachgespräch "Ortsbezogene Anwendungen und Dienste"*. (Sept. 23–24, 2010). Berlin, Germany.
16. Neis, P. and Bauer, M. (2010). Sprachunabhängige Routenanweisungen – Vorschlag zur Erweiterung der OGC OpenLS Route Service Spezifikation. In: *Proceedings of 7. GI/KuVS-Fachgespräch "Ortsbezogene Anwendungen und Dienste"*. (Sept. 23–24, 2010). Berlin, Germany.
17. Neis, P., Singler, P., and Zipf, A. (2010a). Collaborative Mapping and Emergency Routing for Disaster Logistics – Case Studies from the Haiti Earthquake and the UN Portal for Afrika. In: *Proceedings of the Geoinformatics Forum*. (July 6–9, 2010). Salzburg, Austria.
18. Neis, P., Zielstra, D., Zipf, A., and Struck, A. (2010b). Empirische Untersuchungen zur Datenqualität von OpenStreetMap – Erfahrungen aus zwei Jahren Betrieb mehrerer OSM-Online-Dienste. In: *Proceedings of 22. AGIT Symposium für Angewandte Geoinformatik*. (July 7–9, 2010). Salzburg, Austria.
19. Müller, A., Neis, P., and Zipf, A. (2010). Ein Routenplaner für Rollstuhlfahrer auf der Basis von OpenStreetMap-Daten. Konzeption, Realisierung und Perspek-

tiven. In: *Proceedings of 22. AGIT Symposium für Angewandte Geoinformatik.* (July 7–9, 2010). Salzburg, Austria.

20. Singler, P., Zipf, A., and Neis, P. (2010). Supporting Emergency Logistics for the United Nations Logistics Cluster through a Emergency Routing Portal using the UN Spatial Data Infrastructure for Transportation data for Africa. In: *Proceedings of International Symposium on GeoInformation for Disaster Management.* Torino, Italy.

2. Results and discussion

The aim of this chapter is to give a structured summary of the results of each investigation conducted for this cumulative dissertation. The first section provides detailed information about the OSM contributor activity analysis. This is followed by several OSM geodata quality analyses and assessments. The third section presents the newly developed method to generate a tailored routing network for disabled people. Lastly, the route planning application for wheelchair users is introduced with an additional method that attempts to measure the reliability of a computed path based on the quality of the utilized dataset.

2.1. Contributor growth, activity and distribution in a VGI project

In the initial research project the contributors of the OSM project were analyzed in a comprehensive way due to their importance as an essential element of an online community project. VGI projects do not only rely on volunteers to collect geo-information, but also expect the contributors to maintain the collected information to keep it as up-to-date and accurate as possible (Qian et al. 2009). One of the objectives of the first publication (Chapter 5) was to conduct research pertaining to whether the previously described participation inequality theory holds true for the members of the OSM project. Additionally, the first publication aims to get a better understanding of the home location, activity area and activity timeframe of the OSM contributors. Other important findings regarding contributor behavior can also be found in the third publication (Chapter 7).

The introduction of this dissertation already discussed that most online community projects follow a certain "Participation Inequality" (Nielsen 2006) pattern, where only a small amount of the community actively and regularly contributes some sort of data or information to the project. Wikipedia, for instance, had more than 16 million registered members at the beginning of 2012, of which almost 1.5 million made at least one edit and less than 1% (85,000) members made more than five changes. These numbers

did not change significantly in 2013, where Wikipedia had almost 20 million registered members, of which a total of 1.7 million members (9%) edited at least one article and only 125,000 (0,7%) had performed more than five changes (Wikipedia 2013). Similar patterns can also be found in OSM. In 2009 a first analysis showed that out of 120,000 registered OSM members only 33,400 made an edit to the database (Budhathoki 2010). For the first publication of this dissertation (Chapter 5) the contributor activity was analyzed in different ways. The results showed that for the end of 2011 of the 500,000 registered members only 192,000 edited at least one object. Furthermore, the number of project contributions showed that only 24,000 members, who represent 5% of all registered OSM project members, actively made changes in a more productive way.

The eighth publication (Chapter 12) also contains a registered members and active contributor analysis. Figure 2.1 illustrates the results of a newly created evolution and distribution analysis between 2005 and 2014, focusing on registered members and members who made several contributions to the project. The ratio of contributors who made more than 10 edits to the total number of registered members is depicted in Figure 2.2. After an increase of the relative contribution share between 2005 and 2010, the latest negative trend is mainly influenced by the large number of newly registered members in 2013 (Figure 2.1). In total the OSM project has gotten more newly registered members than active contributors in recent years.

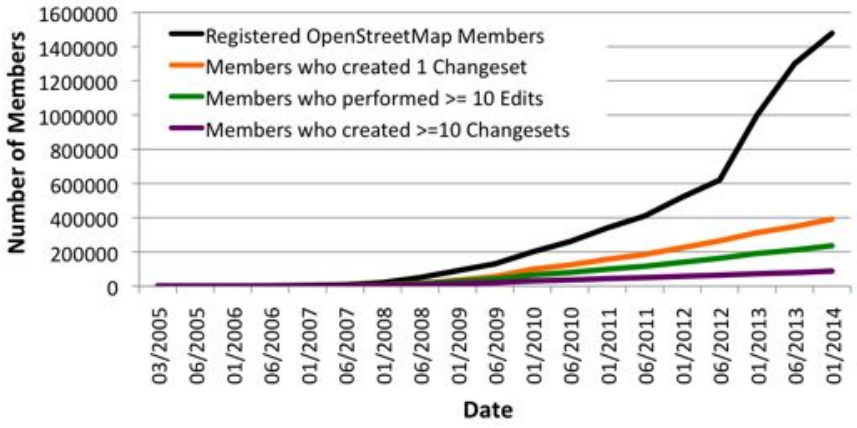


Figure 2.1.: Growth of OSM membership numbers.

The value of around 150 new daily active contributors taken from the second publication (Chapter 6) did not change between 2011 and 2014, despite a strong increase in new registered members. However, the number of daily and monthly active contributors increased from around 2,000 and 17,000 in 2011 to 2,500 and 20,000 in November

2013. A more detailed analysis conducted in the first publication, where members were divided into different groups based on their contributions, clearly showed that only a small amount of all project members actively contributed to the project. Additionally it was revealed that most contributors of the OSM project worked in the evening hours and that they do not prefer a special day of the week for their activities.

The activity timeframe of a contributor is also a critical aspect of the OSM project. Studies by Coleman et al. (2009) and Mooney and Corcoran (2013) questioned the long term motivation of contributors. Similar to the results gathered from the quantity of data contributed by registered members in general, only a small amount of members are long term contributors. Overall it has been proven that almost 70% of newly active contributors each year stopped contributing to the project after a few months.

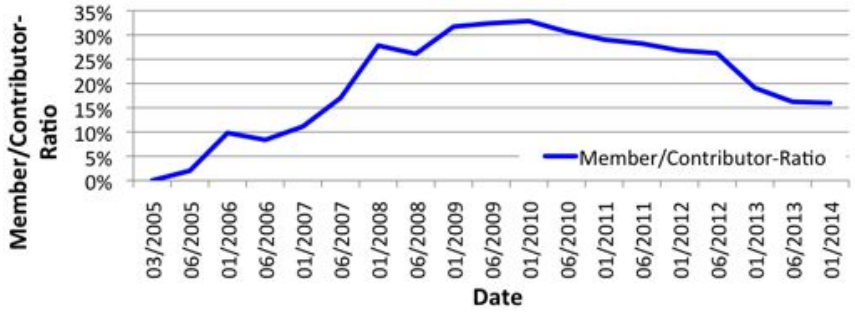


Figure 2.2.: Member/Contributor-ratio between 2005 and 2014.

The OSM project database does not provide information about the home location or the contributors' country of residency. To get a better overview where most contributors are located, four different approaches were tested to retrieve the home location/country of the OSM members:

1. The first Node that was created by the member. (Based on the assumption that the first object is located in close proximity to his or her residence).
2. The mass center of all changesets of each member. (Overlaying and merging all bounding boxes provided by the changesets allows the determination of a particular "activity" area. The center point of this area can be calculated to identify the country of residency of the OSM member).
3. All Nodes that were created by a member. (The country that shows the majority of created Nodes indicates the country/location of a member).
4. The center of member activity area. (Nearly all created Nodes of each member are used to generate areas which represent the activity area).

Approach number four is the most comprehensive and most accurate method. It uses a Delaunay triangulation (Lee and Schachter 1980) to create a polygon based on a mesh from all created Node objects of a contributor. This method also offers the opportunity to determine the area where the main data contribution efforts of the contributor took place. The estimation of the activity area of a contributor has several benefits. It can be used for detailed data quality analyses and to find other local contributors in the area of interest. The results of the activity area generation has shown that more than half of all active OSM contributors work in areas with a size between 10 and 50 km^2 or larger than 50 km^2. The majority of the less active members work within an area between 0.1 and 5 km^2. The results of the different methods also give insight about the distribution of the OSM contributors. The analysis conducted in 2011 was repeated in November 2013 and the results showed a similar number of active members for each continent. The majority of the contributors, who made at least one edit, are located in Europe (69%), while the remaining members (31%) are divided as follows: North America (13%), Asia (10%), South America (4%), Australia (2%), Africa (2%), and Oceania (1%).

Haklay (2010) and Schilling et al. (2009) stated that VGI projects oftentimes showed a better data quality in urban areas than in rural areas. Results have also shown that the number of contributors correlates with the population density of an area. The third publication (Chapter 7) investigated these findings and speculations in more detail. It was evaluated if factors, such as population density, and socio-economic parameters, such as income, can influence contributor concentration and its data contributions to OSM. The analysis was conducted for 12 world regions with at least one urban area for each continent. An additional objective of the study was to determine similarities or significant differences between the selected regions regarding their data growth and collection. Similarly to the results gathered from the OSM contributor distribution analysis, the results of the third publication clearly highlighted the differences between European and other urban areas regarding the number of contributors. Nearly all tested European areas contained a larger number of OSM contributors and higher data density. However, the number of OSM volunteers did not necessarily correlate with the population density in the tested areas. Due to these results, it has been tested if other socio-economic factors such as income can influence the number of OSM members. Overall the findings showed a correlation between income and the number of active OSM contributors, but they also highlighted that some regions with higher income could potentially inherit more contributors than currently available. Thus, the results did not completely confirm prior results gathered for England where “more affluent

areas and urban locations are better covered than deprived or rural locations" (Haklay and Ellul 2011). Furthermore, the previously introduced method to compute the activity area of a contributor has been utilized in this particular analysis to determine if data contributions in the different areas were made by local or external mappers. The results showed that particularly large numbers of external contributors can be found in regions with lower community member numbers. This is an important finding since this pattern contradicts in certain aspects the main idea behind VGI. Originally it was defined by Goodchild (2009) in which local-knowledge, -expertise or -activity (Goodchild 2007) of volunteers should be one of the main sources of information.

2.1.1. Discussion

The results of the "participation inequality" analyses clearly highlighted that the OSM project shows very similar patterns as other online community based projects when considering the member activity. The comparison with other prior findings from Budhathoki (2010) has shown that only a small amount of long-time contributors exist in OSM. Further research is needed to find potential reasons for the reduced workload over time by the contributors. Some speculations can be made about a general loss of interest for the project or that the preferred area of interest already shows a high data completeness. The latter can potentially be analyzed based on the amount of the collected information in the activity area of the contributor. Additionally, an extensive survey could provide some further reliable results. Several findings about the distribution of the contributors stressed that the main focus of the OSM project is based in Europe. Also, the results of the comparison analysis of selected urban areas worldwide showed that other factors besides the population density or income must have an impact on the activities of the contributors. It was hypothesized that differences in Internet access, culture, mentality, personal interests or acquaintance to the project due to language barriers could play a role. The access to such detailed information will be a challenge, due to the fact that the data is needed for several world regions and in a similar quality such as coverage and currentness. Other freely available open geodata can also possibly slow down the growth of OSM contributor numbers and data amounts. In France an active OSM community exists despite several open data initiatives. However, in the United States only a small number of active contributors add data to the project. Most of the aforementioned issues and speculative reasons for contributor decline can only be investigated by extensive surveys. On the other hand a more comprehensive investigation that includes more world regions could improve the results of the urban areas comparison. The analysis focusing on local and external

contributors revealed that at least in some areas major data contributions are made by members that might have never visited the area and collected information locally in person. Although these findings are a good first indicator additional research is needed to answer obvious questions such as: Do external or remote members provide a better, equal or worse data quality when contributing to the project? If contributors traced objects such as buildings or streets from up-to-date, high resolution satellite imagery, at least the geometric quality should have an adequate level. What about attributes such as the name of a feature though?

Several quality aspects have to be evaluated when VGI datasets are considered to be utilized in a project. Therefore the next section gives detailed insight into different VGI quality assessments and approaches to quality assurance.

2.2. Quality assessment and assurance of crowdsourced geodata

2.2.1. Road network evaluation

The earliest research projects about VGI data quality, reach back to 2008 and 2009, where the crowdsourced geodata of the OSM project were compared against administrative or commercial data providers (Haklay 2010, Zielstra and Zipf 2010). The studies that were conducted for the UK and Germany showed similar patterns, with highly detailed data densities in urban areas and a decline in detail richness in rural areas. A similar heterogeneity in data completeness was found in France. Girres and Touya (2010) also reported issues regarding low semantic and attribute accuracy due to different contributors, data sources or data imports in France. The main objective of the second publication of this cumulative dissertation (Chapter 6) was to analyze the spatial data quality of the OSM street network in Germany to evaluate whether the OSM dataset is suitable for routing and navigation applications. Some of the questions the study tried to answer were: How is the relative completeness of OSM in comparison to a proprietary dataset? How did the OSM street network develop in recent years? Can a prediction be made about how the street network will grow in the near future?

OSM data can be downloaded in different file types and for different time stamps. At the time when the second publications has been written, data dumps only existed for specific dates and in different API versions. Nowadays so-called full-history-dumps are available which contain the complete history of the OSM database in one API version. Thus, the analysis that was conducted proved to be a challenge due to different files

sizes and in particular different API versions, making the creation of detailed statistics about a timeframe of 5 years cumbersome. The spatial data quality of the selected OSM street network was tested for the following parameters: completeness, logical consistency and temporal accuracy. The completeness of the OSM street network was determined via a relative comparison with the TomTom Multinet dataset. It needs to be noted that the utilized proprietary dataset mostly contains street network data for motorized traffic navigation. However, the comprehensive analysis of the second publication showed that in June 2011 the crowdsourced OSM street network for car navigation in Germany was only 9% smaller than the one of the commercial provider. The total street network of OSM was almost 27% larger and in terms of pedestrian navigation related ways OSM was approximately 31% larger than the TomTom dataset. The detailed analysis of different time periods during the development of the road network revealed that from specific points in time several road categories did not improve any further. For the conducted analysis of Germany this pattern was interpreted as an indicator for street types which were "close to completion". The results also showed that OSM members most commonly start by contributing higher street types, such as motorways or carriageways, which is followed by lower street types such as residential or forest tracks. The attribute accuracy was analyzed by evaluating the completeness street names for the entire OSM road network of Germany. The result showed that almost 16% of the street network has neither a name nor a route number with a high concentration of unnamed ways and streets between villages or within residential areas. It was speculated that this could be the result of traced roads from satellite imagery in which non-local contributors did not have the knowledge to add a specific street name to an object. However, the analysis of the German OSM dataset also revealed that the implementation of satellite imagery such as Yahoo (until 2012) and Bing (since 2011) generally has a positive impact on OSM data contributions.

For a more detailed completeness analysis in relation to population density of an area, the street network dataset was divided by municipality and town boundaries of Germany. The results clearly showed a strong correlation between the completeness of the dataset and the population density. Less populated areas tend to miss way information accordingly. The analysis also indicated that in 2011 new street data was still being added to the OSM database for sparsely populated regions in Germany.

The temporal data quality of the dataset was evaluated by utilizing the timestamp, indicating when an object was created or last modified. The analyzed OSM dataset for Germany of 2011 showed that approximately one third of the data was created or updated during 2011 and 2010, and another third during 2009 and 2008.

Every routing application requires a graph for the path computation and not only for the objectives of this dissertation is it essential that the graph is topologically correct. OSM contributors attempt to collect topologically correct street or other objects, but OSM street datasets cannot be utilized for routing purposes without preparation. To evaluate the applicability of OSM, the entire street network for Germany was examined to detect topology errors such as unconnected, duplicate or overlapped way segments. The analysis was conducted with datasets for the years between 2007 and 2011 and the results showed that the network in its current state is not without faults. In 2011 most errors existed only in road categories that are not primarily important for routing purposes, such as service ways or paved/unpaved paths. Turn restrictions play another important role for routing applications and navigation systems. Due to different standards and specifications between the proprietary and freely available datasets, the comparison proved to be a challenge and preprocessing was mandatory to merge both data schemas into one consistent format. However, the analysis revealed that the reference dataset has five times more turn restrictions than the OSM dataset for Germany. This clearly highlighted one of the largest caveats of the OSM dataset and revealed that most contributors do not map this type of information.

2.2.1.1. Discussion

In the past few years preliminary quality statements and conclusions revealed that OSM data is sufficient to be used for limited mapping applications. The second publication implemented in this dissertation (Chapter 6) proved that, at least in countries in which the OSM project shows a stronger community, such as Germany, the data is of comparable quality to other datasets, provided by professional geodata vendors. The investigation also highlighted that essential attribute information such as turn restrictions or speed limits are still missing in the dataset and that this particular type of information is not growing at the same pace as the regular street data. The reason for this slow development could be based on the fact that turn restrictions are not rendered in the regular OSM map or that maybe some contributors do not entirely understand how to implement them correctly into the OSM database. The publication also stressed that OSM was missing about 9% of street network data related to car navigation in Germany. Based on the historical development of the road network, a prediction was made that the discrepancy between the data providers should disappear by the end of 2012. In December 2013 the length computations were repeated and the results still showed a difference of 5% between the commercial dataset provider and OSM. One possible reason for this remaining gap could be the OSM license change in

2012 and its corresponding deprecation of data that was contributed by members that did not agree to the new license terms. Many contributors started to recollect data that was deleted after the license change, instead of collecting missing streets in the database.

When including smaller paths and ways for non-motorized traffic in the updated 2013 analysis, the OSM total street network is more than 55% larger in comparison to the commercial TomTom dataset[1] (1,288,374 km). However, the findings of the publication about turn restrictions in OSM, which are of critical importance for car navigation applications, showed that it will take several years for OSM to catch up with the proprietary dataset provider. The results of the updated 2013 analysis revealed at least a strong increase in turn restrictions in the OSM Germany dataset to almost 70,000. This means that the number of restrictions has doubled in comparison to the 2011 dataset.

The findings of the OSM street network analysis for Germany also demonstrated that the completeness strongly correlates to the population density. The results that were found in the comparison analysis of different world regions (Chapter 7) do not support these findings for all tested areas. As previously discussed in Chapter 2.1.1, this means that other aspects must influence the number of contributors, the data density and other spatial data quality parameters.

2.2.2. Intrinsic quality analysis

A commonly applied way of analyzing and assessing the spatial data quality of OSM data in recent years was the relative comparison with authoritative reference datasets (Haklay 2010, Zielstra and Zipf 2010, Chapter 12.3.1.1). Sometimes ground truth reference datasets for quality analyses of a VGI dataset are not applicable due to lack of availability, accessibility, costs or licensing restrictions. However, other analyses for testing or evaluating the spatial data quality without a specific reference dataset are feasible. Thus the main objective of the fifth publication (Chapter 9) was to investigate how OSM data can be assessed without a reference dataset. Therefore novel methods, indicators and visualizations were required for spatial data analysis. The presented intrinsic approach for the quality assessment of OSM is solely based on the data's history. Thereby it captures the data's inherent quality (Batini and Scannapieco 2006) and includes intrinsic parameters such as accuracy, objectivity, believability and reputation (Wang and Strong 1996). The idea behind this approach is to use the

[1] TomTom Multinet - http://www.tomtom.com/en_gb/licensing/products/maps/multinet/#tab:tab2 (visited on 15 December 2013)

entire temporal dimension of a specific object, respectively all objects of a required type, to analyze and assess different aspects about the data quality. Additionally, several other internal quality analyses are possible that evaluate whether the feature attribute accuracy of a POI or if the attributes of an address feature are complete.

Overall 25 methods from existing OSM, VGI or other data quality analyses were taken into account, improved and merged into an expandable framework, named *iOSM-Analyzer*. The designed and implemented framework can create several reports, based on different categories. It can be used to create arbitrary OSM data analyses for any part of the world. The main focus during the development was to evaluate if a dataset of interest is suitable or not for a variety of different use cases. The novel *iOSMAnalyzer* framework contains intrinsic quality indicator analyses for the following six categories: "General Information on the Study Area", "User Information and Behavior", "Routing and Navigation", "Geocoding", "Points of Interest-Search" and "Map-Applications". For instance, general information about a study area can be gathered by investigating the growth of the different OSM features, currentness of the collected data or number of (active) contributors. The implemented framework utilizes the entire OSM full history database dump file during the process. This is a major difference to previously conducted research projects regarding the quality analyses of OSM data where in most cases data extracts with a specific timestamp were utilized.

Three cities, San Francisco (USA) Madrid (Spain) and Yaoundé (Cameroon), were chosen to demonstrate how the framework can be applied. The results of the examples approved that the completeness of a street network can be assessed by evaluating the historical growth of the different road categories (Chapter 6.3). The newly developed method also showed how important information for geocoding applications, such as address information, can be analyzed without a reference dataset. Additionally, based on the results gathered for the three example cities, positive and negative effects of data imports could be detected. As already proven by other studies (Girres and Touya 2010, Haklay et al. 2010), the examples in the analysis also revealed that a high number of active contributors leads to an adequate and up-to-date OSM dataset.

2.2.2.1. Discussion

The proposed framework was developed by applying a wide range of different spatial data quality methods of recent years. However, the implemented framework cannot determine absolute statements for instance about the completeness of a specific feature. In several cases only relative indicators and approximate assessments of the spatial data quality and quantity are possible. Thus, the results of the analysis can only facilitate

the decision whether a selected dataset is usable or not in a predefined region.

Some of the implemented methods require more comprehensive investigations by utilizing additional reference or ground truth data. For instance, most commonly the currentness of the collected geodata is determined by the latest object modification date. This assumption can be problematic in cases where the object is still up-to-date, but has not been changed in the past few years and no update was needed. In this case a more sophisticated approach would be to implement adjacent features in the object's close proximity during the evaluation. This probabilistic approach utilized the latest modifications of surrounding features to adjust their currentness (Exel et al. 2010).

Several data imports can affect the results of the introduced method. Unfortunately some OSM contributors do not mark their data imports with the correct attributes. They can also use different attributes to specify the source or how the data was collected. All conducted quality analyses would benefit if contributors would attach a "source" attribute to the imported data. For instance, contributors imported several license-conform datasets, such as streets, buildings and trees, for the city Yaoundé (Cameroon) into the OSM database. Thus, the city has a quite high OSM data density (Figure 2.3(a)). In contrast, the visualization of the contributors of the street network illustrates that most ways were only modified by merely 5 contributors (each contributor represented by a color in Figure 2.3(b)). This example shows the importance of considering the complete dataset and the number of active contributors in an area for a sophisticated analysis.

(a) (b)

Figure 2.3.: The city of Yaoundé (Cameroon) rendered as default OSM map (a) and as a visualization which illustrated the different contributor who last edited a road segment (b) (date: 22 December 2013).

The *iOSMAnalyzer* framework utilized the OSM full history dump file during the investigation of the specific area of interest. This does not allow us to make statements about the reputation of a particular contributor when considering the entire OSM database. The main focus during the development of the framework was on data and not on contributor behavior. However, as previously discussed, the quality of the dataset highly depends on the contributors actions. Additional methods are needed to determine a contributor's mapping experience or mapping quality. To achieve this type of analysis other OSM datasets next to the full history dump file could be utilized. For instance, the changeset dump file is not integrated in the current framework and contains summary information about the edits of each contributor.

2.2.3. Impact of local and external contributors

Several of the methods implemented in the *iOSMAnalyzer* framework showed that the contributor activity has an impact on OSM data. This finding is particularly important for the interpretation of some intrinsic quality aspects. Haklay et al. (2010) revealed that the positional data quality increases with the number of contributors in an area. The third publication of this dissertation (Chapter 7) demonstrated that, at least in the selected regions, the number of newly created objects highly correlates with the number of contributors in that area. Also, the temporal data quality is better in regions with a high number of active contributors. License-conform open geodata was imported to the OSM database in several cases, examples can be found in the United States or in France (OpenStreetMap 2014a). Although this approach allows project members to fill the map with information where oftentimes previously no or only limited data collections could be found, problems can occur in the long term of a VGI project. Research has shown that the community is not willing to update, nor focuses on improving the completeness of the originally imported data, at least in the US (Zielstra et al. 2013). As already mentioned in Chapter 2.1 the local community always highly influences the quality of the collected information. Therefore it is fundamental to consider the number of active contributors and the total numbers of contributors in the area of interest. The data density can be a first indicator about the quality in a particular region. Due to the aforementioned data imports this picture can be biased. It is possible that the area of interest contains a high amount of OSM data that was not initially collected by individuals. Additionally, external members could contribute larger amounts of collected data too. The assessment of the data contributed by external OSM members in comparison to local members was not part of the third publication (Chapter 7). However, the publication clearly showed that

members, who did not have their main activity area in or around the city of Istanbul (Turkey), contributed a large amount of OSM data. An additional example can be found in Homs (Syria) where external contributors also collected most of the data. An extensive visual comparison of the completeness between the OSM Standard Mapnik Map, Google Maps and Google Satellite for Homs (Figure 2.4(a)) and Istanbul (Figure 2.4(b)) revealed that the OSM street network has a good or partially more complete geometric coverage. The visual assessment was accomplished with the help of the Map compare tool (Geofabrik 2014).

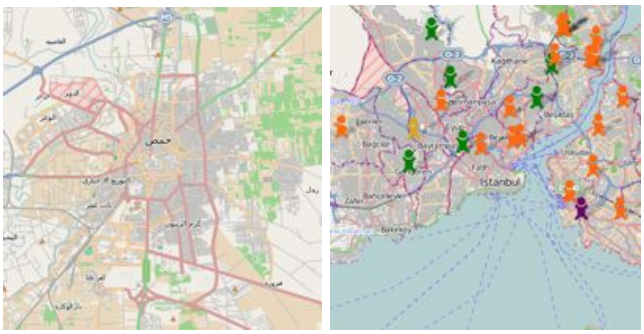
(a)

(b)

Figure 2.4.: Visual comparison of the completeness between the OSM standard Mapnik map, Google Maps and Google Satellite for Homs (a) and Istanbul (b) (source: map compare tool – date: 20 December 2013).

The presented approach to determine the activity area polygon of a contributor, introduced in Chapter 5.4, was implemented in a webpage (Neis 2014b). It visualizes those contributors on a map who have their main activity polygon in the area of interest

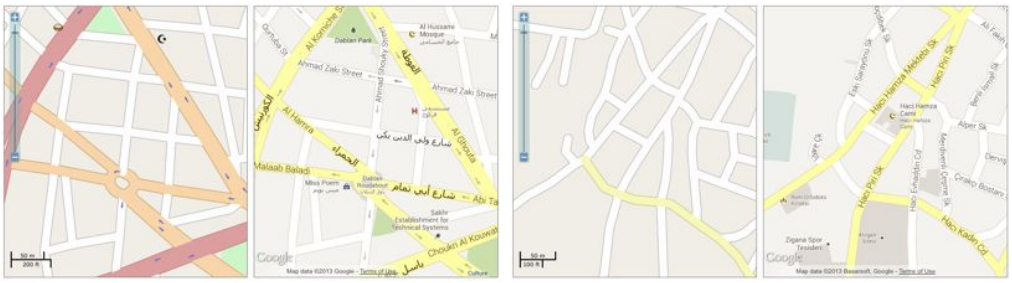
for every part of the world. Homs (Figure 2.5(a)) and Istanbul (Figure 2.5(b)) solely show small numbers or no local contributors at all. Each colored symbol in the map (orange, green, gold, purple) represents one contributor.



(a) (b)

Figure 2.5.: Overview of active OSM Contributors in Homs (a) and Istanbul (b) (date: 22 December 2013).

Further investigations revealed that in many cases only the geometric information of the street network is available and oftentimes the name of the street is missing in both cities (Figure 2.6). Several randomly selected samples had missing minor and major street names. The results clearly show that if an area solely relies on external data contributors, the metadata will be incomplete. Nevertheless, in the cases of political crisis or natural disasters, it is oftentimes better to have any type of map information than having no map at all.



(a) (b)

Figure 2.6.: Comparison of street names between OSM (left) and Google Maps (right) for sample areas in Homs (a) and Istanbul (b) (date: 22 December 2013).

2.2.3.1. Discussion

The necessity of a more detailed analysis of local and external community member contributions has been discussed in Chapter 2.1.1. The evaluation of the previously introduced example cities demonstrated that external contributors partially only map the geometrical representation of objects and do not add important information such as street names. However, more extensive investigations or analyses that include reference datasets could improve the findings that were presented. One analysis could evaluate whether areas with large external community member contributions show similar patterns as regions that experienced large data imports. Zielstra et al. (2013) showed in their study that the community did not focus on improving imported data in the US. Mooney and Corcoran (2012) also mentioned that most contributors only edit or update their own collected and contributed objects. Still, large, regional external mapper contributions and the effect of data imports leave many unanswered questions: Who updates or adds new details in an area which was initially covered by an import or by the work of an external contributor? Maybe there is a chance that the members update their own initially created data as stated by Mooney and Corcoran (2012), but what happens if the member does not contribute to the OSM project for a long term? The availability of satellite imagery provided by companies such as Microsoft, generated larger amounts of objects such as buildings (Goetz and Zipf 2012) in the OSM database and allowed contributors to collect information of remote areas. Furthermore, they enable visual data comparisons and validations. The question that remains is: What happens to the project if Microsoft does no longer provide their aerial imagery service or withdraws the right to trace from their imagery?

The aforementioned comparison of selected urban regions in the world highlights the disparity between local and external contributors. It also confirmed that, despite similar population densities, the data density development is not identical in each region. It needs to be noted that the analyses did not consider any type of buildings, local facilities or infrastructure. In a more extended analysis these parameters could be compared between the different regions, a country or even a continent in combination with the population density.

2.2.4. Quality assurance based on a vandalism detection tool

Similar to other UGC based projects such as Wikipedia, one of the main caveats of the OSM project are different vandalism types that can occur. Oftentimes data vandalism detection is closely related to data validation (to some extent). While data validation

incorporates different methodologies for data quality assurance, vandalism focuses on the damage properties made without permission by the owner. In the case of UGC projects it is also possible that new contributors accidently make changes that are harming the project's main goal. It is necessary to counteract such patterns due to the fact that the projects applicability and reliability are heavily affected. The fourth publication (Chapter 8) attempts to answer the following question: How can the OSM project protect itself against data vandalism? Therefore, the main objective of the publication was the investigation and detection of vandalism events in OSM, combined with the development of a rule-based system for automated vandalism detection.

The analysis of the detected OSM vandalism cases showed no particular geographical pattern. Only a slight concentration in larger cities could be found. In one third of vandalism cases, users created some fictional data or modified some existing data, e.g., non-regular geometrical modification. More than 40% of the vandals deleted existing data. In almost 80% of the cases, a new project member vandalized the OSM data by using the default OSM Potlatch Editor. Compared to the Wikipedia project, where a vandalism event is usually reverted within minutes (Kittur and Kraut 2008), 63% of the events in the OSM project were reverted within 24 hours and 76.5% within 48 hours. Some outliers could be determined, which needed more than 5 days up to a maximum of 29 days.

A prototypical implementation of a rule-based decision system, named *OSMPatrol*, was developed based on an investigation of past vandalism incidents, the current OSM database and its contributions/contributors, as well as related Wikipedia vandalism detection tools. Thereby the system considers the contributors' individual project reputation, as well as the performed mapping action. It is crucial that both parameters are evaluated independently from each other. This allows the detected vandalism cases to be filtered in a later step. To be able to detect a vandalism case as fast as possible, the OSM minutely "Diff"- files were utilized. These files contain all information about every single change that was made to the OSM database every minute. It has to be noted, that the files only contain the latest object version and not any specific information about the actual type of change. Thus, *OSMPatrol* requires an OSM database to be able to compare the former and the newly created or updated OSM object. This enables the possibility to evaluate whether the geometry or the attributes of the object has changed. Furthermore it answers questions such as: Who is the former object-owner or what is the edit date of the former object? The optional attribute information can be tested against a created information table that consists of well-known OSM Map Features accepted and widely used by the community (OpenStreetMap 2014b).

This allows the determination of additional (semantic) information accuracy. Also, the attribute information of the former and the latest object will be compared to analyze which information has been modified. The geometry of the different object versions will also be compared with each other to detect if an object has been moved more than, e.g., 11 m.

The contributor plays a major in VGI projects. One assumption that can be made is that new project members are more prone to errors, mistakes or vandalism in comparison to more experienced contributors. Therefore a similar database, as implemented by Neis (2014a), was integrated in the prototype and enabled the determination of individual project member reputations. This member reputation contains information such as how many objects the contributor has modified, when she/he started to contribute to the project or what her/his favorite or most commonly used OSM attributes are.

The prototype was tested for one week in August 2013. Despite the high number of false positives in the experimental results, the tool was able to find one real vandalism case per day. Every second case of the detected vandalism events were reverted by other OSM members within two days. The results also showed that almost one third of the 9,200 users, who were detected as possible vandals, were new members of the project. Additionally, only 1,000 contributors committed almost 85% of the detected vandalism cases. The latter value highlights the importance of using and maintaining the introduced OSM members black and white list of the vandalism detection tool. The list could significantly improve the proposed approach. Overall, the introduced rule-based system (*OSMPatrol*) was able to detect vandalism types committed by new contributors, "illegal" imports or automated mass edits.

2.2.4.1. Discussion

The developed prototypical implementation of the vandalism detection tool demonstrated some first and promising results. One of the major challenges was the processing of the minutely "Diff"-files to detect vandalism cases. This initial aim could not be entirely realized, based on the utilized hardware configuration of the server. Furthermore, a couple of issues and ideas became apparent. The implemented user reputation needs some more attention. For instance the project-membership time span should be computed based on the combination of the days since the member has registered to OSM and the days the member was actually actively contributing to the project. Furthermore, some rules have to be adjusted. For instance the version number of an OSM object is used to evaluate its "quality".

This is based on the assumption that a high version number represents more potential feature reviewers, because the version is incremented with each change of the object. Thus, the prototype will assess an edit on an object with a high version number more likely as some kind of vandalism than a change on an object with a low version number. However, the aforementioned rule should be adjusted to a combination of the version number and the number of distinct editors of an object. This would probably represent an adequate indicator, instead of using only the general version number.

The presented version of the vandalism detection prototype did not consider further investigations of the neighborhood or surrounding area of a newly created or edited object. For instance, if a contributor changed a living street to a motorway in a residential area, it is obviously some type of vandalism. Based on the large variety of different object types, several rules and checks have to be implemented. These additional checks, and in particular time consuming geometrical tests including the surrounding neighborhood of a feature, affects the processing time in a negative way. Next to changes or modifications to attributes, another major issue was the detection of errors or changes to existing object names. Generally, it is probably impossible to properly detect such vandalism cases without the integration of a dictionary which might provide some improvements. However, different languages could still play a major role and cause several errors.

Wikipedia usually saves the IP address of a user who edited an article. This type of information of a contributor is not available for the OSM project (except for OSM server administrators). To be able to detect cases of vandalism, it might be a beneficial indicator to investigate if the geolocation of the IP suits the region in which the edit has been executed. Thus, it is possible to investigate if a contributor changes an object that is far away or one which is next to her/his access point. However, since the IP addresses of the contributors are only available to the administrators of the OSM server infrastructure, the vandalism detection tool would have to be integrated into the main project infrastructure and cannot be run by non-authorized individuals.

The number of detected false positive vandalism cases was fairly large in the tested phase. Although several vandalism events could be spotted with this automated process it needs to be noted that a manual review of each case is still necessary and preferable. The two separated parameters (project reputation of the contributor and performed mapping action) enable a well-defined way to filter the results of the tool for individual user requirements. A minor disadvantage is that the results of *OSMPatrol* are partially hard to interpret. A possible solution for this issue might be to provide a webpage or program that supports the user in the review process of the detected edits in an easier

and more comfortable way. Particularly the illustration of the objects latest and former geometry version or the comparison of the attributes would enhance the interpretation and validation of the vandalism edits.

2.3. Generation of a sidewalk routing network

Commercial or administrative geo-information is widely used to generate an adequate routing network for car, pedestrian or bicycle navigation applications or devices. The level of detail that is needed to receive similar results for people with disabilities differs significantly from the traditional data. Due to the high costs that arise when collecting this detail of information i.e. personnel, equipment, time or data maintenance, authoritative data providers usually do not provide this type of information.

Most research projects in the realm of pedestrian friendly routing networks in recent years based their analysis on networks that were oftentimes created via extensive surveys or by tracing the pedestrian information needed from satellite imagery (Kasemsuppakorn and Karimi 2008, Kasemsuppakorn and Karimi 2009). Other studies demonstrated that a sidewalk routing network could be derived through binary image processing methods (Gaisbauer and Frank 2008, Kim et al. 2009). In more recent years, the OSM contributors have been collecting this detail of information for more advanced applications in selected regions. Thus, the aim of sixth publication (Chapter 10) is to present an approach for the generation of a tailored routing network for disabled people based on the freely available data of the OSM project. In a first step a comprehensive investigation was conducted to evaluate which specific parameters are needed for a disabled friendly routing network. Thereby different findings for wheelchair users, blind and visually impaired people, elderly people or children have been summarized based on a variety of publications. In a second step the selected parameters were matched with the corresponding OSM attributes that can be found in the database of the project.

The method presented for the final generation of a sidewalk routing network for disabled people consists of several processing steps. It has been prototypically implemented and tested for specific central regions of all European capital cities. The results showed that 36 out of the 50 selected and tested regions provided less than 1% of the needed OSM sidewalk information. Further, eleven tested cities had less than 10% of relevant information and only the city centers of Berlin (Germany), London (United Kingdom) and Riga (Latvia) provided more than 30% of important information. For Berlin the sidewalk information was spread over the selected area of interest, whereas

the results for Riga showed that the majority of the information was only concentrated in one city district. For London most required sidewalk parameters could be found along the main streets in the city center. The efficiency of the sidewalk algorithm and its generated network was tested for a number of randomly selected routes with different start and end points in the predefined areas. The results of the test clearly showed an improvement in pedestrian friendliness in comparison to regularly created routing networks. Particularly the number of utilized footway sections and the additional sidewalk information during the route computations proved to be significant. Next to the assessment of the efficiency of the presented algorithm, the quality and quantity of crucial attributes were evaluated. The list of introduced and required parameters for disabled people friendly routing network is quite detailed and comprehensive. Thus, several cities did not provide any of the information. Figure 2.7 illustrates the generated sidewalk network (black dotted lines) with its crossings (red dotted lines) on a satellite imagery background. Overall this example of an additional test area in Bonn (Germany) shows multiple crossings, but it is also obvious that some sidewalk information is missing in the OSM data.

Figure 2.7.: Generated OSM sidewalk network (satellite imagery by Google Maps).

In OSM a sidewalk is mapped by adding a number of attributes to the corresponding street segment. The disadvantage of this approach is that the sidewalk information will not be rendered in the default OSM maps. This could potentially be one of the reasons why many sidewalks in the tested areas were mapped as separate footway objects. However, the presented approach revealed that a disabled people friendly sidewalk routing network can be generated based on VGI data. Furthermore, the generated network is highly adaptable and can be utilized for different use cases such as online or offline and printed maps or personal navigation assistants for people with disabilities.

2.3.1. Discussion

During the development of the generated sidewalk network, based on the suggested algorithm, several issues regarding the quality of the OSM data arose. Simple errors, such as duplicate nodes or ways in the database, can cause the network generation to be erroneous. Another common error occurs when ways in the dataset do not share a common node at a junction, which results in an overlap between roads. Next to the detection of these types of topology errors it is important to conduct additional tests to validate the generated sidewalk network. While the tagging quality in OSM plays a major role to retrieve particular attributes that are important for a tailored routing network, the geometrical quality cannot be neglected during this process. The final step of the generation process should be a validation and verification of the created network. During this step the entire network should be tested for duplicate ways and intersections of sidewalks or crossings. Figure 2.8 depicts an example error found at a complex junction with a turning lane and incorrect tagging in the OSM data which results in an intersection of two sidewalks of different streets (ID 1 & ID 2). Besides the described scenario, it is also possible that sidewalks intersect with neighboring streets.

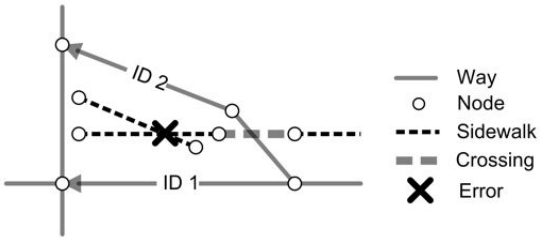


Figure 2.8.: Intersection between two sidewalks of different streets.

However, a particular set of specifications or standardization would be preferable for the mapping process of the OSM project. For instance, some contributors use a point as a decimal mark while others prefer to use a comma. Others again switch between centimeter and meter or attach the units directly to the object attribute. Of course more strict mapping standards would contradict the important bottom-up tagging approach, but it would most likely improve the data quality. It has to be noted, that the data of the OSM project sometimes renders to be useless, due to the fact that it is not interpretable. For instance, the attribute value of the "incline" tag oftentimes contains an "up" or "down" annotation. This information is for the generation of a routing network for disabled people not helpful, because it misses a precise value of the incline of a street. However, the following section demonstrates

how this newly introduced sidewalk network can be utilized as an adequate foundation for more advanced routing applications.

2.4. Route planning for wheelchair users

Most of the commonly used routing applications for cars or pedestrians calculate the fastest or shortest path between a start and a destination point. For a sophisticated wheelchair routing application, the algorithm that generates the "best" path is more complex. Nearly all published research projects in recent years in the realm of wheelchair friendly navigation applications used multi-criteria networks to determine the path for different wheelchair types and individual user requirements. Thereby different properties of each way segment, such as surface or width, can play a major role.

The seventh publication (Chapter 11) included in this dissertation introduced two novel approaches for the assessment and evaluation of the feasibility of a wheelchair user friendly routing algorithm and its generated path. The first method computes a tailored route based on specific individual user requirements. The second method evaluates the generated path by providing a reliability factor based on the quality of the applied dataset. Several studies about the development of a wheelchair user route planning application in the past concluded that the following parameters are essential for each segment of the route network: length, width, slope, sidewalk surface, steps, sidewalk conditions and sidewalk traffic (Matthews et al. 2003, Beale et al. 2006, Kasemsuppakorn and Karimi 2009). All of these parameters can be found in the previously introduced OSM sidewalk routing network (Chapter 10), with the exception of sidewalk traffic which was not included due to obvious data collection constraints for this variable. Next to the quality of the applied routing network, an adequate weighting method was introduced and played a major role during the route generation.

The newly presented approach only utilizes way segments that coincide with the users' individual needs. For instance, if a user specifies that the calculated path should only include segments with a width of more than 100 cm, only way segments with the matching criteria will be included. This guarantees that the determined route will only contain streets, which are passable for each individual wheelchair user. This step differs from other proposed methods that utilize all segments of the routing network during the path determination. The computation of the final route can also be influenced by the quality and quantity of the utilized OSM geodata. Thus, a second method computes a *Reliability Factor* (RF) for the generated route. Each individual user requirement and

its corresponding parameter of the newly presented weighting method will be utilized to evaluate the fitness of the determined path. It needs to be noted, that the RF is influenced by the attribute availability during the route computation. However, due to the heterogeneous pattern of the OSM data quality, it is necessary to provide additional information about the quality of the generated path and to what degree the generated path can be trusted. The presented approach is not limited to OSM datasets only. It can be implemented in any type of routing application in which not only the geometric representation is crucial, but also other related details and attributes about the object or way segments.

2.4.1. Discussion

The developed methods have some advantages and disadvantages compared to regular pedestrian routing or prior developed wheelchair algorithms by Matthews et al. (2003), Beale et al. (2006) or Kasemsuppakorn and Karimi (2009). Based on the user requirements, which can be defined before each computation, an individual route for the wheelchair user will be generated. In particular the different number of attributes related to a way segment, which are implemented during the path finding process, can be a limitation of the introduced algorithm. If multiple way conditions are not suitable for wheelchair routing, such as sloped curbs that are too high or poor surface conditions, the algorithm may not be able to compute a route between the start and destination point based on the declared requirements. Although these particular cases might be exceptions based on the routing graph, larger, wheelchair user unfriendly detours could become a more common concern as shown in tested case studies (cf. Chapter 11.5). Further research is needed here, for instance in form of an extensive survey, to determine the maximum detour length a wheelchair user is willing to travel. Another option to avoid these detours could be the determination of a route with lower restrictions or expectations to the way conditions, e.g. if the wheelchair user relies on help from a passerby to pass a barrier. Matthews et al. (2003) and Beale et al. (2006) have shown that the integration of an additional parameter that controls the importance of the length of the calculated path can be feasible for routing applications tailored to wheelchair users or disabled people. However, the parameter of the aforementioned studies acts as a weighting parameter and not as a distance limit that could be implemented with the cost function of the proposed routing algorithm.

The results discussed in Chapter 2.3 already highlighted that many cities have a good geometric representation of the regular street network in OSM but oftentimes lack the detail of information that is needed for the specific use case of disabled people. Missing

attribute information or incorrectly tagged information can impact the generation of a path. Figure 2.9 illustrates an incident in which the attribute information for a sidewalk (wrong slope curb height) leads to a detour in which the wheelchair user has to cross the street to use the sidewalk on the opposite side of the street. The dotted line indicates the optimal path that should be used when travelling from North to South, while the generated route in blue leads the wheelchair user to her or his destination via a detour by using an additional crossing and the opposite side of the street. This example proves how small errors in the dataset can lead to incorrect results.

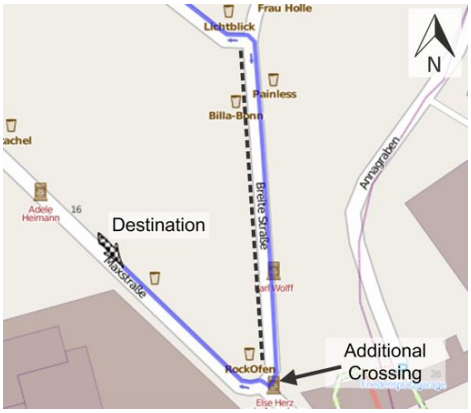


Figure 2.9.: Example of incorrectly-tagged sidewalk information in OSM, resulting in a detour for the wheelchair user.

Although the quality of the utilized dataset of the OSM project can be problematic at times, no other freely available large area data alternatives are available for the development of a disabled people friendly network or wheelchair user routing application. The current situation in OSM regarding this special type of information is similar to the situation some years ago when the project was lacking general routing information until first applications sparked the interest of the community to add this type of information. Maybe a similar development can be seen in the near future with applications that require detailed attributes for route planning for disabled people.

3. Conclusion

With the development of the Web 2.0 phenomena, many community based online projects arose that collect a variety of data. One of these UGC/VGI related projects is OSM with a large number of volunteers that contributed a collection of geographic objects with different levels of detail and specific attribute information. Nevertheless, collaboratively contributed spatial data provided by volunteers raises many questions about quality assessment and assurance, credibility, reliability, trust and utilization. However, the collected data provides a plethora of opportunities for new application scenarios.

Therefore, the main objective of this dissertation was to contribute new insights in the realm of the aforementioned quality assessment, assurance and the utilization of VGI for the implementation of a route planning application for people with disabilities. For this goal, a number of newly developed methods and tools were introduced supporting new findings in the field of VGI research, routing networks and wayfinding for disabled people.

The conducted contributor activity analysis provides an excellent foundation to understand how the community is structured and contributing to the OSM project (Chapter 5). The collected German street network of the OSM project has been analyzed in a comprehensive investigation to demonstrate how user-generated geodata is comparable to commercial datasets in different aspects of spatial data quality (Chapter 6). OSM shows a heterogeneous data quality pattern as highlighted by the contributor and data analysis. A comparison of different selected urban areas was conducted to determine similarities or significant differences regarding their data growth and collection efforts (Chapter 7). The development of a comprehensive rule-based prototypical tool that allows for automatic vandalism detection in the OSM project has been discussed (Chapter 8). Additionally, due to the limited access of reference datasets for quality assessments, a framework has been designed to enable intrinsic data quality analyses for the evaluation of an OSM dataset (Chapter 9). For the implementation of a route planning for disabled people, a newly developed algorithm was introduced that generates a sidewalk routing network from a dataset provided by the OSM project (Chapter

10). This novel sidewalk representation builds the foundation for the development of a new approach to wheelchair-user friendly route computation and evaluation (Chapter 11). Furthermore, a comprehensive overview and detailed discussion of recent VGI related research projects and a state of the art literature review with regards to the OSM project was conducted (Chapter 12). The latter also contains a discussion of potential future VGI and OSM trends in research and development.

The findings of this dissertation can be summarized as followed:

- The analyses of the OSM contributors highlighted that the OSM project follows a general participation inequality pattern, which means that only a minority of all registered members actively contribute to the project. The investigation also showed that the majority of OSM contributors are located in Europe.
- The comprehensive investigation of the German street network evolution and the comparison with a reference dataset revealed that collaboratively collected OSM data is becoming comparable in quality and quantity to commercial datasets as well as temporal and geometric accuracy. However, missing turn restrictions, which are essential for car navigation, revealed one of the major caveats of the VGI project.
- The comparison of selected world regions showed significantly different results in data collection efforts and local OSM community sizes. Especially the number of OSM members can differ largely between European and other world regions in OSM resulting in smaller amounts of collected geodata. Further investigations also highlighted that socio-economic factors, such as income, can have an impact on the number of active contributors and the data provided in the analyzed areas.
- Based on the investigation of different vandalism cases, the current OSM database and its contributions and a variety of other Web 2.0 vandalism detection tools, a comprehensive rule-based system for automatic vandalism detection was developed. The introduced prototype can provide useful information about the vandalism types and their impact on the OSM project data.
- The introduced framework to evaluate the quality of an OSM dataset contains a broad range of more than 25 different methods and indicators. The applied intrinsic approach for quality analyses allows for an evaluation of the OSM dataset without the need for a ground truth reference dataset for any part of the world. The results of the framework can facilitate the decision whether the quality of OSM data in a selected area of a user's choice is sufficient for her or his use case or not.

- A newly developed algorithm introduced in this dissertation can generate a disabled people friendly sidewalk routing network based on OSM geodata. This new representation of a routing graph can be utilized in numerous applications and maps dedicated to people with disabilities.
- The new approach to advanced route planning for wheelchair users enables tailored way computations for individual and personal requirements provided by the user and the calculation of a reliability factor of the computed path. The utilization of both methods is not limited to an OSM dataset and can be utilized for any type of routing or navigation purpose that is based on other commercial or administrative datasets.

4. Future work

A comprehensive list of potential future trends in VGI and in particular OSM research and development are discussed in Chapter 12.4. Additionally, each chapter of this dissertation, representing a published journal article, contains ideas for possible future research (Chapter 5-12). The following section gives a summary of potential questions that need to be addressed in the future.

The contributors play a major role for the future of the OSM project. Since OSM data does not solely rely on data collection, but also on maintaining the data to keep it as accurate and up-to-date as possible. At least every third contributor (Chapter 12.3.2.1), of all active contributors that ever added information to the project, will continue to contribute. Future research could reveal what type of information long-term members contribute. Do they collect new data in different areas or do they start collecting more detailed information such as trees, sidewalk and surface information near their home location? On the other hand it would be interesting to know what demotivates contributors and makes them stop contributing data to the project. However, for the introduced vandalism detection tool (Chapter 8) and the intrinsic OSM quality framework (Chapter 9) it is necessary to have more additional in-depth analyses regarding their mapping experience, quality of contributions, project reputation and behavior (cf. Chapter 12.4).

The highest concentration of active OSM contributors can be found in Germany. Thus, the road network completeness shows good results, sometimes exceeding commercial providers for some particular areas (Chapter 6). Further work is needed to investigate the evolution of the dataset in other countries that do not have a broad contributor base. Is there any similar pattern detectable in the data or contributor density evolution over the years? The comparison of the selected world regions (Chapter 7) also highlighted that several regions rely on the contributions of non-local members. This raises questions about the general quality of the objects that were mapped by external contributors. Do external members provide a better, equal or worse data quality when contributing to the project? For instance, do they only trace street objects to fill an empty map (as shown in Chapter 2.2.3) or do they also contribute more advanced

objects from other open data sources such as buildings with their addresses? And finally who maintains objects that are mapped by non-local contributors?

It was shown that there is not enough evidence to generally correlate population density to the number of OSM members (Chapter 2.1 and Chapter 7). Further analyses revealed that socio-economic factors, such as income, could have an impact on the number of active contributors and the data density. However, questions remain about potential reasons why some regions in different countries or continents only show small OSM communities and not similar success as in Europe. Possible effects could be differences in freely available geodata in various countries, Internet access, culture, mentality, personal interests or acquaintance to the project due to language barriers could play a role too. Other influential indicators could most likely only be determined by conducting an extensive survey.

The developed vandalism detection prototype tends to detect more vandalism cases than there actually are in reality. One of the main ideas behind the prototype was similar to a network firewall, preferring the detection of too many rather than the detection of too few cases of vandalism. For future research, it would be important to enhance the number and variety of filters to sort the gathered results (Chapter 8.5 & 8.6). Additionally, the API of the developed prototype should be extended to visualize and present the detected vandalism cases via a well-defined interface. One possible (and desirable) application would enable users to register as a patrol guard for a distinct area. This way, a user can define a distinct region and/or distinct attributes and as soon as *OSMPatrol* detects a vandalism type that suites to a patrol's preference pattern, she or he is informed via e-mail. However, this would require volunteers that are willing to act as contribution reviewers.

The proposed algorithm for the generation of a disabled people friendly sidewalk network (Chapter 10) provides room for new research projects based on the current findings, such as the combination with a multi modal routing graph that implements sidewalk and public transportation network information. It is also feasible that the new representation is utilized in OSM 3D city models (Goetz 2012) or routing applications for blind people (Kammoun et al. 2010). Furthermore, the resulting network of the algorithm could be improved by applying additional information. For instance, during the generation of the sidewalk network it could be useful to consider building information or barriers such as street lamps or road signs in the middle of a sidewalk, which is also available in the OSM project database, to position the sidewalks correctly between the road and a row of houses, similar to the work introduced by Ballester et al. (2011).

It has already been discussed (Chapter 2.4.1) that for specific cases the limitations defined by the user or poor road and pavement quality can lead to detours or unsuccessful path generations by the proposed wheelchair routing method. Therefore further research is needed to determine the maximum detour length a wheelchair user is willing to travel with the guarantee that the calculated route does not contain any impassable way segments.

Overall the utilized OSM dataset proved to be a valuable source for the generation of a sidewalk network and for wheelchair route planning as long as the detailed tags for disabled people are included. Future research focusing on the timelines of the data needs to be conducted to insure that the OSM dataset is undergoing certain update processes to maintain and improve the currently available geodata. However, one of the main questions that remains is: Do contributors collect this detailed information worldwide although it is not being rendered in the OSM standard maps? As an alternative, required attribute information such as the incline of a road could be improved by the combination of the 2D way geometry from OSM together with a *Digital Elevation Model* (DEM) (cf. Beale et al. 2006).

References

ADA (2010). *Americans with Disabilities Act (ADA) Standards for Accessible Design*. URL: `http://www.ada.gov/2010ADAstandards_index.htm` (visited on 01/01/2014).

Anderson, P. (2007). What is Web 2.0? Ideas, Technologies and Implications for Education. In: *JISC*.

Anthony, D., Smith, S. W., and Williamson, T. (2007). The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia. In: *Dartmouth Computer Science Technical Report TR2007-606*. Hanover, NH: Dartmouth College.

Ballester, M.G., Pérez, M.R., and Stuiver, H.J. (2011). Automatic Pedestrian Network Generation. In: *Proceedings of the 14th AGILE International Conference on Geographic Information Science*. (Apr. 18–21, 2011). Utrecht, The Netherlands.

Batini, C. and Scannapieco, M. (2006). *Sharing Imperfect Data*. Berlin, Germany: Springer.

Bauer, R., Delling, D., Sanders, P., Schieferdecker, D., Schultes, D., and Wagner, D. (2008). Combining Hierarchical and Goal-Directed Speed-Up Techniques for Dijkstra's Algorithm. In: *Proceedings of International Workshop on Experimental Algorithms (WEA 2008)*. (May 30–June 2, 2008). Provincetown, Massachusetts, USA.

Beale, L., Field, K., Briggs, D., Picton, P., and Matthews, H. (2006). Mapping for Wheelchair Users: Route Navigation in Urban Spaces. *The Cartographic Journal*, 43(1), 68–81.

Bellman, R.E. (1958). On a Routing Problem. *Quarterly of Applied Mathematics*, 16(1), 87–90.

Bennett, J. (2010). *OpenStreetMap: Be Your Own Cartographer*. 1st. Birmingham, UK: Packt Publishing.

Brando, C. and Bucher, B. (2010). Quality in User-Generated Spatial Content: A Matter of Specifications. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

Budhathoki, N. (2010). Participants' Motivations to Contribute to Geographic Information in an Online Community. Ph.D. Dissertation. University of Illinois, Urbana-Champaign, Urbana, IL, USA.

Budhathoki, N. and Haythornthwaite, C. (2013). Motivation for Open Collaboration: Crowd and Community Models and The Case of OpenStreetMap. *American Behavioral Scientist*, 57, 548–575.

Chen, H. and Walter, V. (2009). Quality Inspection and Quality Improvement of Large Spatial Datasets. In: *Proceedings of the GSDI 11 World Conference: Spatial Data*

Infrastructure Convergence: Building SDI Bridges to Address Global Challenges. (June 15–19, 2009). Rotterdam, The Netherlands.

Chrisman, N.R. (1983). The Role of Quality Information in the Long-term Functioning of a GIS. In: *Proceedings of AUTOCART06*. Falls Church, VA.

Coleman, D., Georgiadou, Y., and Labonte, Y. (2009). Volunteered Geographic Information: The Nature and Motivation of Producers. *International Journal of Spatial Data Infrastructures Research*, 4, 332–358.

Dechter, R. and Judea, P. (1985). Generalized Best-first Search Strategies and the Optimality of A*. *Journal of the ACM*, 32(3), 505–536.

Delling, D. and Wagner, D. (2009). Time-dependent Route Planning. In: *Robust and Online Large-Scale Optimization*. Berlin/Heidelberg, Germany: Springer207–230.

Devillers, R. and Jeansoulin, R. (2010). Spatial Data Quality: Concepts. In: *Fundamentals of Spatial Data Quality*. London, UK: ISTEpp. 31–42.

Devillers, R., Stein, A., Bédard, Y., Chrisman, N., Fisher, P., and Shi, W. (2010). Thirty Years of Research on Spatial Data Quality: Achievements, Failures, and Opportunities. *Transactions in GIS*, 14(4), 387–400.

Dijkstra, E.W. (2010). A Note on two Problems in Connexion with Graphs. *Numerische Mathematik*, 1(1), 269–27.

DIN 18024-1 (1998). *Standard specification by the German Institute for Standardization (Deutsches Institut für Normung (DIN)) 18024-1: "Barrierefreies Bauen - Teil 1: Straßen, Plätze, Wege, öffentliche Verkehrs- und Grünanlagen sowie Spielplätze; Planungsgrundlagen" / Barrier-Free Design - Part 1: Streets, Places, Roads and Recreational Areas; Planning Basics*. URL: `http://nullbarriere.de/din18024-1.htm` (visited on 01/03/2014).

Exel, M. van, Dias, E., and Fruijtier, S. (2010). The Impact of Crowdsourcing on Spatial Data Quality Indicators. In: *Proceedings of the Sixth International Conference on Geographic Information Science (GIScience 2010) Workshop on the Role of Volunteered Geographic Information in Advancing Science*. (Sept. 14–17, 2010). Zurich, Switzerland.

Flanagin, A. J. and Metzger, M. J. (2008). The Credibility of Volunteered Geographic Information. *GeoJournal*, 72, 137–148.

Floyd, R.W. (1962). Algorithm 97: Shortest Path. *Communications of the ACM*, 5(6), 345–370.

Gaisbauer, C. and Frank, A.U. (2008). Wayfinding Model for Pedestrian Navigation. In: *Proceedings of the 11th AGILE International Conference on Geographic Information Science 2008*. (May 5–8, 2008). Rome, Italy.

Geisberger, R., Sanders, P., Schultes, D., and Delling, D. (2008). Contraction Hierarchies: Faster and Simpler Hierarchical Routing in Road Networks. In: *Proceedings of the 7th Workshop on Experimental Algorithms, Volume 5038 of Lecture Notes in Computer Science*. Springer.

Geofabrik (2014). *Map Compare | Geofabrik Tools*. URL: `http://tools.geofabrik.de/mc/` (visited on 01/02/2014).

Girres, J.F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS*, 14, 435–459.

Goetz, M. (2012). Using Crowd-sourced Indoor Geodata for the Creation of a Three-dimensional Indoor Routing Web Application. *Future Internet*, 4, 575–591.

Goetz, M. and Zipf, A. (2012). Towards Defining a Framework for the Automatic Derivation of 3D CityGML Models from Volunteered Geographic Information. *International Journal of 3-D Information Modeling*, 1, 1–16.

Goodchild, M.F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211–221.

Goodchild, M.F. (2008). Spatial Accuracy 2.0. In: *Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. (June 25–27, 2008). Shanghai, China.

Goodchild, M.F. (2009). NeoGeography and the Nature of Geographic Expertise. *Journal of Location Based Services*, 3, 82–96.

Haklay, M. (2010). How good is OpenStreetMap Information: A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. *Environment and Planning B*, 37, 682–703.

Haklay, M., Basiouka, S., Antoniou, V., and Ather, A. (2010). How many Volunteers does it take to Map an Area well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47, 315–322.

Haklay, M. and Ellul, C. (2011). *Completeness in Volunteered Geographical Information - The Evolution of OpenStreetMap Coverage in England (2008–2009)*. URL: `http://povesham.wordpress.com/2010/08/13/completeness-in-volunteered-geographical-information-%E2%80%93-the-evolution-of-openstreetmap-coverage-2008-2009/`.

Holone, H., Misund, G., and Holmstedt, H. (2007). Users Are Doing It For Themselves: Pedestrian Navigation with User Generated Content. In: *Proceedings of the 2007 International Conference on Next Generation Mobile Applications, Services and Technologies*. (Sept. 12–14, 2007). Cardiff, Wales, UK.

Javanmardi, S., Ganjisaffar, Y., Lopes, C., and Baldi, P. (2009). User Contribution and Trust in Wikipedia. In: *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing*. (Nov. 11–14, 2009). Washington, DC, USA.

Jo, H.H., Karsai, M., Kertész, J., and Kaski, K. (2012). Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*. DOI: 10.1088/1367-2630/14/1/013055.

Kammoun, S., Dramas, F., Oriola, B., and Jouffrais, C. (2010). Route Selection Algorithm for Blind Pedestrian. In: *Proceedings of the International Conference on Control, Automation and Systems*. (Oct. 27–30, 2010). KINTEX, Gyeonggi-do, Korea.

Kasemsuppakorn, P. and Karimi, H.A. (2008). Data Requirements and Spatial Database for Personalized Wheelchair Navigation. In: *Proceedings of the 2nd International Convention on Rehabilitation Engineering and Assistive Technology*. (May 13–15, 2008). Bangkok, Thailand.

Kasemsuppakorn, P. and Karimi, H.A. (2009). Personalised Routing for Wheelchair Navigation. *Journal of Location Based Services*, 3(1), 24–54.

Kasemsuppakorn, P. and Karimi, H.A. (2013). A Pedestrian Network Construction Algorithm Based on Multiple GPS Traces. *Transportation Research Part C: Emerging Technologies*, 26, 285–300.

Kim, J., Park, S.Y., Bang, Y., and Yu, K. (2009). Automatic Derivation Of A Pedestrian Network Based On Existing Spatial Datasets. In: *Proceedings of the ASPRS /MAPPS 2009 Fall Conference*. (Nov. 16–19, 2009). San Antonio, Texas.

Kittur, A. and Kraut, R.E. (2008). Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination. In: *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. (Nov. 8–12, 2008). San Diego, CA, USA.

Lechner, M. (2011). Nutzungspotentiale crowdsource-erhobener Geodaten auf verschiedenen Skalen. Ph.D. Dissertation. University Freiburg, Freiburg, Germany.

Lee, D.T. and Schachter, B.J. (1980). Two Algorithms for Constructing Delaunay Triangulation. *International Journal of Computer and Information Sciences*, 9, 219–242.

Lin, Y. (2011). A Qualitative Enquiry into OpenStreetMap Making. *New Review of Hypermedia and Multimedia*, 17, 53–71.

Matthews, H., Beale, L., Picton, P., and Briggs, D. (2003). Modelling Access with GIS in Urban Systems (MAGUS): Capturing the Experiences of Wheelchair Users. *Area*, 35(1), 34–45.

Mooney, P. and Corcoran, P. (2012). How Social is OpenStreetMap? In: *Proceedings of the 15th AGILE International Conference on Geographic Information Science*. (Apr. 24, 2010–Apr. 27, 2012). Avignon, France.

Mooney, P. and Corcoran, P. (2013). Has OpenStreetMap a role in Digital Earth Applications? *International Journal of Digital Earth*. DOI: 10.1080/17538947.2013.781688.

Müller, A., Neis, P., and Zipf, A. (2010). Ein Routenplaner für Rollstuhlfahrer auf der Basis von OpenStreetMap-Daten. Konzeption, Realisierung und Perspektiven. In: *Proceedings of 22. AGIT Symposium für Angewandte Geoinformatik*. (July 7–9, 2010). Salzburg, Austria.

Neis, P. (2014a). *How Did You Contribute to OpenStreetMap?* URL: http://hdyc.neis-one.org (visited on 01/02/2014).

Neis, P. (2014b). *Overview of OpenStreetMap Contributors*. URL: http://resultmaps.neis-one.org/oooc (visited on 01/05/2014).

Neis, P., Singler, P., and Zipf, A. (2010). Collaborative Mapping and Emergency Routing for Disaster Logistics – Case Studies from the Haiti Earthquake and the UN Portal for Afrika. In: *Proceedings of the Geoinformatics Forum*. (July 6–9, 2010). Salzburg, Austria.

Neis, P. and Zipf, A. (2008). OpenRouteService.org is Three Times "Open": Combining OpenSource, OpenLS and OpenStreetMaps. In: *Proceedings of the GIS Research UK 16th Annual conference GISRUK 2008*. (Apr. 2–4, 2008). Manchester, UK.

Nielsen, J. (2006). *Participation Inequality: Encouraging More Users to Contribute*. URL: http://www.nngroup.com/articles/participation-inequality/ (visited on 09/23/2013).

OpenStreetMap (2014a). *Import – OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Import (visited on 01/01/2014).

OpenStreetMap (2014b). *Map Features – OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Map_Features (visited on 01/04/2014).

OpenStreetMap (2014c). *Open Database License – OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Open_Database_License (visited on 01/01/2014).

OpenStreetMap (2014d). *OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Main_Page (visited on 01/01/2014).

O'Reilly, T. (2005). *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. URL: http://oreilly.com/web2/archive/what-is-web-20.html (visited on 09/21/2013).

Qian, X., Di, L., Li, D., Li, P., Shi, L., and Cai, L. (2009). Data Cleaning Approaches in Web 2.0 VGI Application. In: *Proceedings of the 17th International Conference on Geoinformatics*. (Aug. 12–14, 2009). Fairfax, VA, USA.

Ramm, F., Topf, J., and Chilton, S. (2010). Cambridge, UK: UIT.

Raymond, E.S. (1999). The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. In: Beijing, China: O'Reilly.

Roick, O. and Heuser, S. (2012). Location Based Social Networks – Definition, Current State of the Art and Research Agenda. *Transaction in GIS*, 17, 763–784.

Schilling, A., Over, M., Neubauer, S., Neis, P., Walenciak, G., and Zipf, A. (2009). Interoperable Location Based Services for 3D Cities on the Web Using User Generated Content from OpenStreetMap. In: *Proceedings of the 27th Urban Data Management Symposium*. (June 24–26, 2009). Ljubljana, Slovenia.

Schmitz, S., Neis, P., and Zipf, A. (2008). New Applications Based on Collaborative Geodata - the Case of Routing. In: *Proceedings of XXVIII INCA International Congress on Collaborative Mapping and Space Technology*. (Nov. 4–6, 2008). Gandhinagar, Gujarat, India.

Sobek, A. and Miller, H. (2006). U-Access: A Web-based System for Routing Pedestrians of Differing Abilities. *Journal of Geographical Systems*, 8(3), 269–287.

Stark, H.J. (2010). Umfrage zur Motivation von Freiwilligen im Engagement in Open Geo-Data Projekten. In: *Proceedings of FOSSGIS Anwenderkonferenz für Freie und Open Source Software für Geoinformationssysteme*. (Mar. 2–5, 2010). Osnabrück, Germany.

Steinmann, R., Häusler, E., Klettner, S., Schmidt, M., and Lin, Y. (2013). Gender Dimensions in UGC and VGI - A Desk-based Study. In: *Proceedings of Angewandte Geoinformatik 2013*. (July 3–5, 2013). Salzburg, Asutria.

Stephens, M. (2013). Gender and the GeoWeb: Divisions in the Production of User-generated Cartographic Information. *GeoJournal*, 1–16.

Van Oort, P.A.J. (2006). Spatial Data Quality: From Description to Application. Ph.D. Thesis. Wageningen University.

Veregin, H. (1999). Data Quality Parameters. In: *Geographical Information Systems: Principles and Technical Issues*. New York, USA: John Wiley and Son177–189.

Wang, R. and Strong, D. (1996). Beyond Accuracy: What Data Quality means to Data Consumers. *Journal of Management Information Systems*, 12, 5–33.

Wikipedia (2013). *Wikipedia:Statistics*. URL: `http://en.wikipedia.org/wiki/Wikipedia:Statistics` (visited on 10/05/2013).

Yasseri, T., Sumi, R., and Kertész, J. (2012). Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis. *PLoS ONE*. DOI: 10.1371/journal.pone.0030091.

Zielstra, D., Hochmair, H.H., and Neis, P. (2013). Assessing the Effect of Data Imports on the Completeness of OpenStreetMap – A United States Case Study. *Transaction in GIS*, 17, 315–334.

Zielstra, D. and Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

Part II.

Publications

5. Analyzing the Contributor Activity of a Volunteered Geographic Information Project – The Case of OpenStreetMap

Authors

Pascal Neis and Alexander Zipf

Journal

ISPRS International Journal of Geo-Information

Status

Published: 27 July 2012 / Accepted: 17 July 2012 / Revised: 17 July 2012 / Submitted: 13 June 2012

Reference

Neis, P. and Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(2):146-165.

Contribution statement

Pascal Neis conducted all analyses for this study and wrote the majority of the manuscript. Alexander Zipf supported this publication through continuing discussions about the methods and the results of the study. Furthermore, extensive proof-reading by the co-author led to substantial improvements to the manuscript.

.....

Prof. Dr. Alexander Zipf

Abstract

The *OpenStreetMap* (OSM) project, founded in 2004, has gathered an exceptional amount of interest in recent years and counts as one of the most impressive sources of *Volunteered Geographic Information* (VGI) on the Internet. In total, more than half a million members had registered for the project by the end of 2011. However, while this number of contributors seems impressive, questions remain about the individual contributions that have been made by the project members. This research article contains several studies regarding the contributions by the community of the project. The results show that only 38% (192,000) of the registered members carried out at least one edit in the OSM database and that only 5% (24,000) of all members actively contributed to the project in a more productive way. The majority of the members are located in Europe (72%) and each member has an activity area whose size may range from one soccer field up to more than 50 km^2. In addition to several more analyses conducted for this article, predictions will be made about how this newly acquired knowledge can be used for future research.

Keywords: volunteered geographic information (VGI); OpenStreetMap; contributions; community; activity.

5.1. Introduction

The *World Wide Web* (WWW) has evolved significantly from its early stages in the 1990s, sometimes referred to as Web 1.0, to a sophisticated source of information. At the beginning of the 21st century, the term "Web 2.0" was first introduced (Knorr 2003). However, the term experienced its real attention after a publication by O'Reilly in 2005, entitled "What Is Web 2.0?" (O'Reilly 2005).

The change in terminology is based on a shift in the usage of the Web, which is no longer characterized by the consumption of predefined content. In fact the term Web 2.0 relates to a "new" platform where users can customize their own applications on the WWW to meet their own design, ideas, and functionality and, most importantly, can create their own data or edit existing data.

The online encyclopedia "Wikipedia," established in 2001, is based precisely on this phenomenon. The newly created information is referred to as "user-generated content" or "user-created content" (Wunsch-Vincent and Vickery 2007). The voluntary users, who are spread all over the world, share their knowledge on various topics on one particular online platform. Other websites that are based on a similar approach

allow users to share their videos (YouTube) and photos (Flickr, Panoramio) with others. Similar efforts are the foundation of geodata platforms such as *OpenStreetMap* (OSM), Tagzania, Wayfaring.com, the People's Map, and Platial or The People's Atlas, where volunteers, amateurs, or professionals gather information and upload it to a central database available on the Internet (Coleman et al. 2009). However, unlike other platforms that rely on user contributions such as Wikipedia and Flickr the collected information is not about a particular topic or image; instead, it contains more specific details about elements such as streets, points of interest, or buildings, which always include a geographic reference. The literature describes this particular type of data as *Volunteered Geographic Information* (VGI; Goodchild 2007), while others describe the process as "crowdsourcing geospatial data" (Heipke 2010).

The OSM project has developed into one of the largest sources of VGI in recent years. Hundreds of thousands of members are contributing to the project worldwide. Different applications based on spatial data provided by the OSM project have been developed. Besides the creation of different maps for hikers[1], skiers[2] and public transportation networks[3] the information also shows potential for more advanced applications such as *Location-based Services* (LBS; Neis and Zipf 2008) or a Web 3D Service (Over et al. 2010). Also, the implementation of OSM data for indoor areas has been discussed (Goetz and Zipf 2011).

With the change of the licensing model by Google Maps in early 2012 (Google 2011) and the potential costs that can arise, more and more businesses are moving to the free option offered by the OSM project. The location-based social network FourSquare (Foursquare 2012) and the Nestoria Property Search (Nestoria 2012) are two major examples that have changed their services to the OSM platform. Also professional spatial data providers and companies have seen the potential in user-generated information and have created their own platforms in the past few years such as Google Map Maker (Google 2012), TomTom Map Share (TomTom 2012) and Nokia Map Creator (Nokia 2012), which allow customers and users to edit their own data to the provided maps. However, the collected information on these platforms, including the changes provided by volunteers, is the property of the platform operator and will not be freely available to other users.

These developments show that over the past few years the success of the VGI approach to data collaboration and sharing is undeniable. However, questions remain

[1] http://www.wanderreitkarte.de (visited on 27 March 2012)
[2] http://openpistemap.org (visited on 27 March 2012)
[3] http://www.öpnvkarte.de (visited on 27 March 2012)

about the motivation of the members to participate in projects such as OSM. According to different research results, there are a variety of factors that can play a major role (Budhathoki 2010, Lin 2011). Possible factors might be the unique ethos or that geospatial information should be freely available to everyone. For others, learning new technologies, self-expression, relaxation and recreation or just pure fun can play a major role. However, these particular motivational factors are certainly not unique to VGI-related projects, but also to other online communities and platforms such as Wikipedia.

One major caveat of VGI platforms that has been identified is the very small percentage of members, e.g., in Wikipedia that actively contribute to the project (Javanmardi et al. 2009). During the writing process of this article, Wikipedia had a worldwide community of more than 16 million registered members, of which a total of 1.5 million members had at least made one change to an article (Wikipedia 2012). Less than 85,000 members had made more than five changes, which represents less than 1% of the registered members. These results correspond closely to what has been termed "Participation Inequality" and a general "90-9-1" rule that can be applied to community-based projects (Nielsen 2006). The numbers represent the 90% of an online community who consume the provided information but do not contribute to the project, the 9% who contribute occasionally, and the only 1% who create or edit most of the content and can be considered active members. Similar results were found in a previous study about Wikipedia, indicating that only a small percentage of the members actively contributed to the project (Anthony et al. 2007). The research conducted provides information pertaining to whether the above-mention-ed participation inequality theory holds true for the members of the OSM project. A first analysis focusing on participants' characteristics and motivations in OSM in April 2009 had shown that of 120,000 registered users only 33,440 contributors made at least one edit to the database (Budhathoki 2010). Similar to previous work, the members are split into several groups based on their contributions to the project to provide a better overview (Javanmardi et al. 2009). Owing to the different methods that were applied during the research process, it is possible to provide statements about the origin of a member and her/his main area of activity.

The remainder of the paper is structured as follows: The next section gives an introduction to the VGI OpenStreetMap project. The third section of the paper compares registered vs. active contributors of the OSM project, while sections four through six focus on the determination of the location, the activity area, and the active time frame of a member. The last section summarizes the results of this paper and presents some future research suggestions. It needs to be noted that almost all analyses in this article

are based on the full history dump file of the OSM project dated 8 December 2011. In the compressed format the file has a size of 30 GB while the uncompressed file size increases to 500 GB. All programs that were applied to perform the conducted analyses were developed in Java with the implementation of a variety of open source libraries.

5.2. The OpenStreetMap project

The OSM project, founded in 2004 at the University College London, has the goal to create a free database with geographic information of the entire world, and detailed introductions to the project have been published (Haklay and Weber 2008, Bennett 2010, Ramm et al. 2010). A plethora of spatial data such as roads, buildings, land use areas, or points of interest is entered into the project's database. Similar to other community-based projects on the Internet, any user can start contributing to the project and editing data after a short online registration. This simple approach allowed the project to gather more than 640,000 registered members by June 2012 (OpenStreetMap 2012b).

The contribution of new data to the project can be accomplished in different ways. The most classical, yet still most common, approach is to record data using a GPS receiver and edit the collected information using one of the various freely available editors. The user provides additional information about the collected data by adding attributes and stores the final results in the OSM database. Several companies such as Yahoo (up to 2011) and Microsoft Bing support the project by providing various aerial images to the community, which allows the OSM members to digitize data such as streets from the images (OpenStreetMap 2012d, OpenStreetMap 2012a). However, this process has its advantages and disadvantages. While this method is a very simple way to add new data, the disadvantage is that aerial imagery is oftentimes outdated or not properly geo-referenced. More importantly, it is not possible to get any metadata information such as the road or street names from an aerial image. Another common practice in recent years has been the import of other freely available data into OSM. For instance, the complete TIGER/Line dataset of the United States and donated data from *Automotive Navigation Data* (AND) for the Netherlands were imported into the OSM database. The database server plays a major role and contains the membership administration, the GPX tracks, and, of course, all spatial data of the project.

For this article, it is important to note what type of data provided by the OSM project has been used and how to retrieve it. The major OSM components that were utilized for the analyses are shown in Figure 5.1.

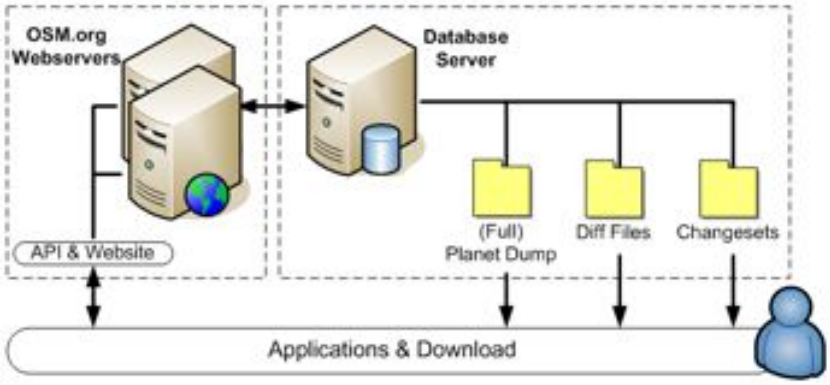


Figure 5.1.: How to retrieve OpenStreetMap (OSM) data.

There are different methods for retrieving raw data from the project. One way is to download "dump files" which are updated on a weekly basis and include the latest versions of the objects of the database. Additionally, once every quarter, a complete database dump file with all available versions of the objects is released. If a user is only interested in changes that were made to the database, OSM provides "diff files" that contain the latest changes to the database by minute, hour, and day. Any modification made by a member to an object in OSM is stored in a "changeset" file. This particular information can also be downloaded as a weekly dump file. Most of the information provided is stored in XML format and sometimes in binary format, which allows for significantly faster processing of the data. Additionally, there are various third-party applications and web pages that provide maps for GPS devices or shapefiles based on OSM data.

The geographic information in the OSM database, such as roads, land use information, or buildings, is stored by using three object types: Nodes, Ways and Relations. A "Node" in the database contains location information of a point in the form of latitude and longitude coordinates. Lines such as roads and polygons are stored as "Ways," and "Relations" define logical or geographic relationships between the objects. Each object contains additional information such as a version number, an ID, the name of the editor, the date when it was created or last modified, and, of course, further attributes, so-called Tag/Value pairs.

Anonymous changes to the database are no longer supported; however, any Internet user who registers for the project can add information to the map and change existing data. This open approach to collaborative data collection creates questions about the quality of the spatial data. Studies regarding this topic have been conducted

and published in recent years (Neis et al. 2012, Mooney and Corcoran 2012a). The OSM data collection shows an overall very heterogeneous quality, i.e., the quality and completeness of the database varies highly from country to country. For urban areas in Europe, especially in the United Kingdom, Germany, Austria, and Switzerland, the OSM data proves to have a similar completeness to commercial or governmental data providers. However, rural areas show a lower data concentration in the OSM database with the exception of the USA, where an opposite pattern, i.e., better coverage in rural areas and less completeness in urban areas, could be determined (Haklay 2010, Zielstra and Zipf 2010, Girres and Touya 2010, Zielstra and Hochmair 2011).

5.3. Registered members vs. active contributors

It is often stated that the OSM project has hundreds of thousands of members who help to collect or improve the data of the project. As outlined in the introduction to this article, this pattern seems to contradict that of most other online portals that are based on user contributions. The direct extraction of information about the members of the OSM project, such as a list of all members, or registration information, is not possible. Thus, a different approach needs to be considered to be able to analyze OSM contributors' actions.

Based on the full history dump file[4] of 8 December 2011, and the changeset dump file[5] of 7 December 2011, it was possible to create a list of all members who made changes to the database. The registration date of each member can be retrieved from the corresponding user's website. The collected information for the OSM dataset is shown in Figure 5.2. The increase of registered members since the beginning of the project is represented by the black line, while the red line represents the number of the members who have at least created a changeset, and the orange line represents the number of members who have edited at least one object (Node, Way or Relation). Finally, the green line represents the number of members who have created at least one object in the database. The results in Figure 5.2 show that in December 2011 the OSM project had approximately 505,000 registered members.

However, comparing the number of registered members with any of the other retrieved information reveals a strong difference in growth over the past few years. At the end of 2011 almost 43% (213,000) of the members created a changeset but only 38% (193,000) of all members edited (created, modified or deleted) at least one object

[4] http://planet.openstreetmap.org
[5] http://planet.openstreetmap.org

type (Node, Way or Relation) and only 33% (169,000) of all members created an object in the database. It must be noted that, in the past, if a member logged into the OSM online editor, a changeset was created independent from the fact of actual changes to the database being made or not. This particular error causes the small difference between the number of members who created a changeset or edited at least one object. Lastly, 62% (312,000) of the members of the project did not actively contribute any information.

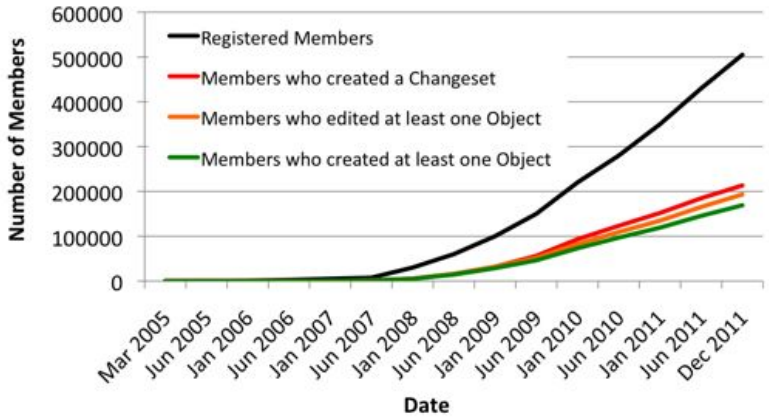


Figure 5.2.: Registered OSM members vs. OSM members with at least one edit (2005–2011).

Considering these first results, the question remains whether some of the new members will become more active in the near future. Based on the information retrieved from the database, it was possible to determine the time that elapsed between the date of registration and the first edit to the database or creation of an OSM object by a user. The results shown in Figure 5.3 indicate, similar to Figure 5.2, that slightly less than two-thirds of all members have never created an OSM object. This large number can partially be explained by the widespread misconception that users need to be registered to retrieve OSM data. Thus, users register for the project but do not actively contribute any information. It can also be determined that in most cases the OSM member made his or her first edit to an OSM object on the same day as the registration (about 30% of all members). Based on this information there is no evidence for an increased activity in the near future for a large number of OSM members.

In mid-2011, around 150 new members began to actively contribute to the project (Neis et al. 2012). Based on the newly conducted analysis presented in this article, an average of about 600 new members registered each day in January 2012, of which

200 began to contribute actively to the project. These numbers correspond to the 30% threshold shown in Figure 5.3.

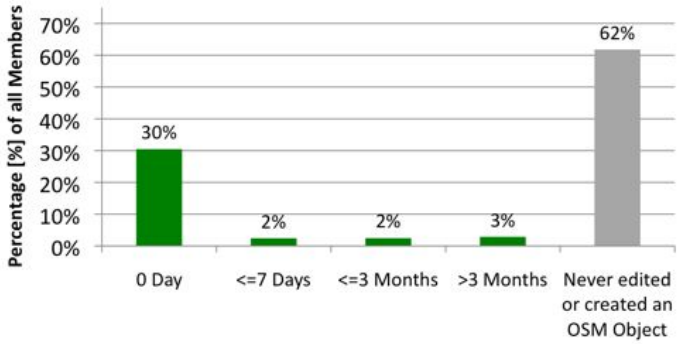


Figure 5.3.: Days between registration and the first created OSM object (2005–2011).

Table 5.1 shows the total number of objects that were retrieved from the OSM dump file of December 2011. As mentioned before, all Nodes, Ways, and Relations were collected by at least 193,000 different members. Approximately 98% of the Nodes in the database were collected by almost 14,500 members, 98% of the Ways by 17,400, and 98% of the Relations by only 5,400 members.

Table 5.1.: Statistics of the OSM database (December 2011).

Object	Overall	Visible	Versions	Total Number of Contributors	Contributors of 98% Data
Node	1.47 billion	1.29 billion	2.01 billion	182,000	14,600
Way	129 million	117 million	228 million	156,000	17,400
Relation	1.7 million	1.2 million	5.5 million	33,500	5,400

To give a better overview of the number of members, their work and their activity with the project, a diagram was created based on their created objects (Figure 5.4). Four particular member groups could be determined after investigating and visualizing the skewed distribution of the values (due to the large number of users who did not contribute any data) and applying different bin sizes. Crucial to the group assignment of a member was the number of Nodes that were created by the member. The results gathered showed that approximately 24,000 members created more than 1,000 Nodes, representing 5% of the 505,000 registered members. This group of members is referred to as “Senior Mappers”. About 73,000 members, who correspond to 14% of the total number of members, created at least 10 and fewer than 1,000 Nodes, and these members

may be referred to as “Junior Mappers”. Nearly 96,000 members created fewer than 10 Nodes, which makes them the least active, but also the largest member group, with 19%. Members falling into this class are referred to as “Nonrecurring Mappers”. The largest group without any action in the OSM project is represented by the remaining 312,000 members (62%). Thus, the remainder of the analyses will focus on Groups 1–3 with a total of 193,000 members.

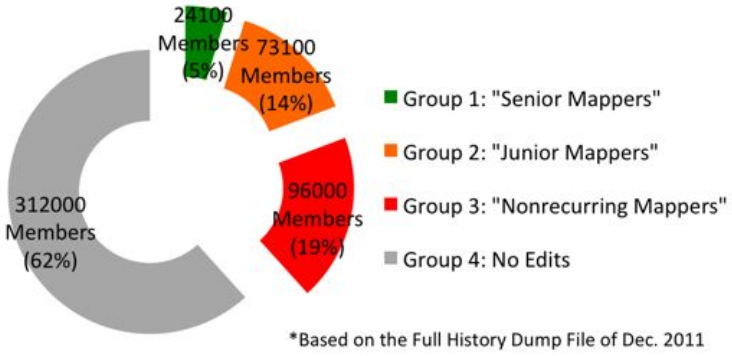


Figure 5.4.: Distribution of registered members based on their node contributions*.

In order to show the exact activity of each group per weekday and time of day, all changesets of the OSM project were investigated based on a changeset file retrieved on 7 December 2011. As previously described, a changeset contains information about who has made an edit at what time. Also it describes with its coordinates a rectangular area in which the changes by the member have been made. Of the approximately 10 million changesets provided by the database, 89% were created by Group 1 (“Senior Mappers”), which represents only 5% of all members. The “Junior Mappers” group generated 9%, while the “Nonrecurring Mappers” generated only 2% of the changesets. Figure 5.5 provides more detailed information about the weekday on which most changesets were created. Almost all weekdays show similar changeset values with the exception of Sunday, which has a slightly larger value. In addition to the distribution of changesets per weekday, more detailed information could be gathered by analyzing the changesets by the time of day. Therefore, the timestamps, provided by OSM in *Coordinated Universal Time* (UTC), of all changesets were evaluated. The results, shown in Figure 5.6, highlight that the majority of changesets were created during the afternoon and evening hours. Ideally, the changeset information would be evenly distributed throughout the daylight hours based on the worldwide community character of the project. However, currently this is not the case, and the results support the

aforementioned focus of the project in the European time zones.

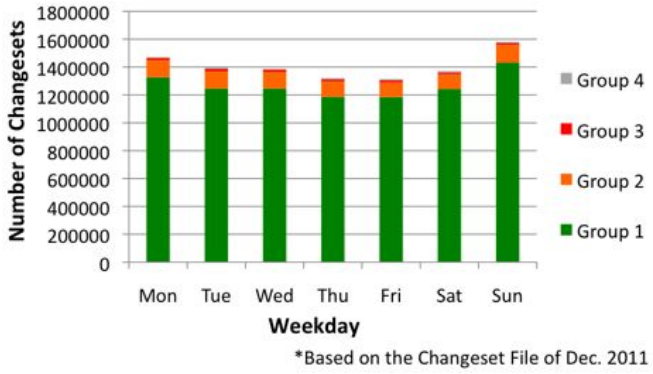


Figure 5.5.: Changesets per weekday*.

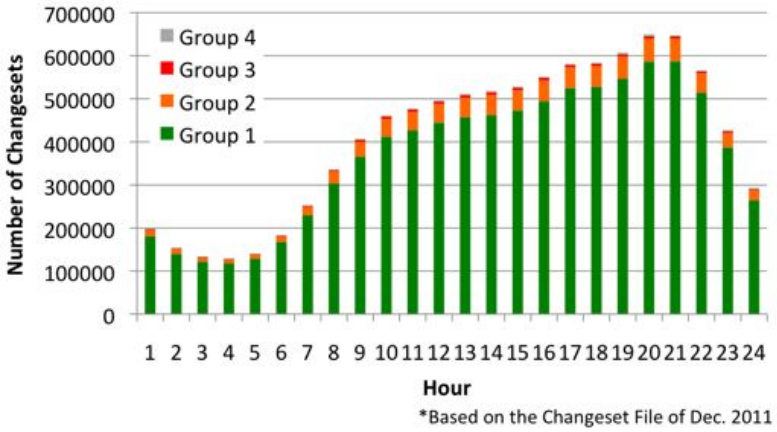
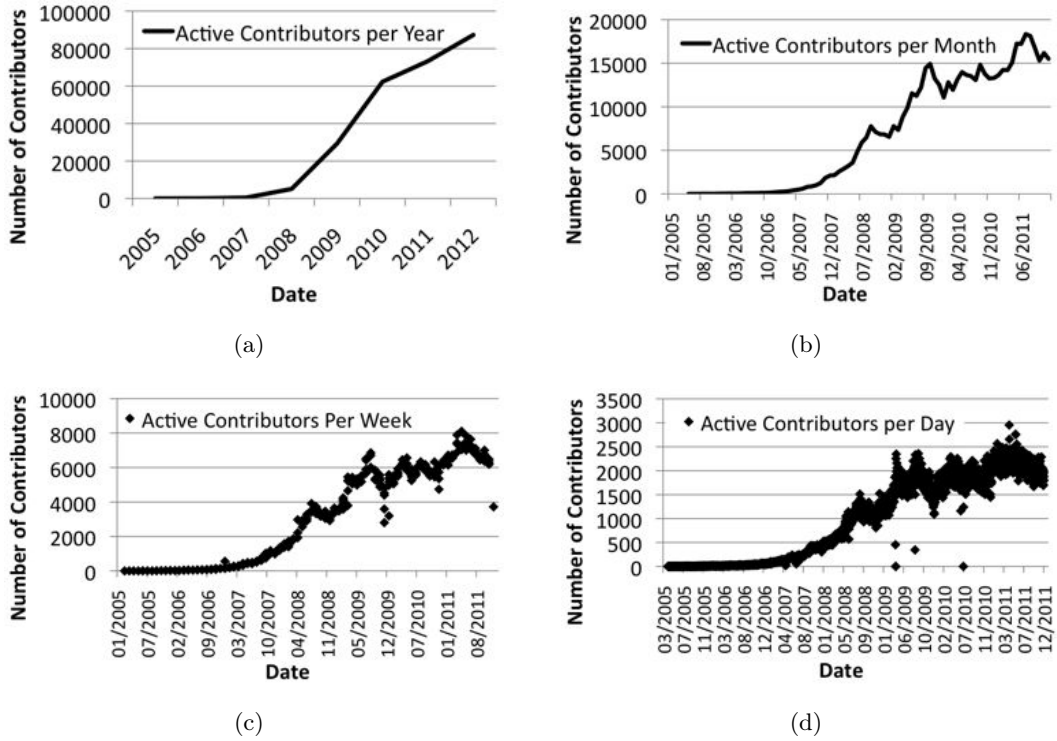


Figure 5.6.: Changesets per hour*.

Further information can be gathered by analyzing the number of members per year, month, week, and day who make changes to the database. At the beginning of 2008, about 10% of the 30,000 registered members of the OSM project added new data every month (Ramm and Stark 2008). A year later, this value decreased to almost 8%, although the total number of project members increased to 200,000 members. In 2010, only about 4% of the members collected new data per month (OpenStreetMap 2012c). This negative trend continued in 2011, when the number dropped to about 3%.

Figure 5.7(a–d) shows the corresponding figures for the years 2005 to late 2011. In 2011, almost 87,000 different users made at least one change to the database, corresponding to approximately 17% of the total number of members. The monthly analysis

showed that at the end of 2011, between 16,000 and 18,000 active members (representing approximately 3% of all members) contributed to the project. The weekly number of members with at least one contribution fluctuated between 6,000 and 7,000, representing only 1% of the total community. The daily member numbers were between 1,800 and 2,200, representing a percentage of active members far below 1%.



(a) (b) (c) (d)

Figure 5.7.: Number of active contributors per (a) year, (b) month, (c) week and (d) day.

An analysis that we conducted in January 2012 based on OSM "Diff" files showed that, in total, all members generated about 1.2 million Nodes, 130,000 Ways, and 1,500 Relations per day, with about 2,100 active members for each day of the month. This means that each member created on average about 570 Nodes, 60 Ways, and 0.7 Relations.

5.4. Member location

In addition to the information gathered that was based on the contributions of OSM members, further tests were conducted with a focus on member locations and activity

areas. The OSM database does not provide specific information about the member's country of residency. However, if this information could be retrieved in a different manner, it could give data about how many active and inactive members each country hosts. The quality of the dataset relates closely to the number of members in an area or country that add or improve the data (Haklay et al. 2010). Four different approaches to retrieve member location information from the OSM database were applied:

1. The first Node that was created by the member determines the country/location of the member.
2. The mass center of all bounding boxes provided by the changesets of each member is determined to retrieve the country/location of a member.
3. All Nodes that were created by a member are taken into consideration and the country that shows the majority of the Nodes indicates the country/location of a member.
4. The center of the activity area polygon of a member provides the country/location information of the member.

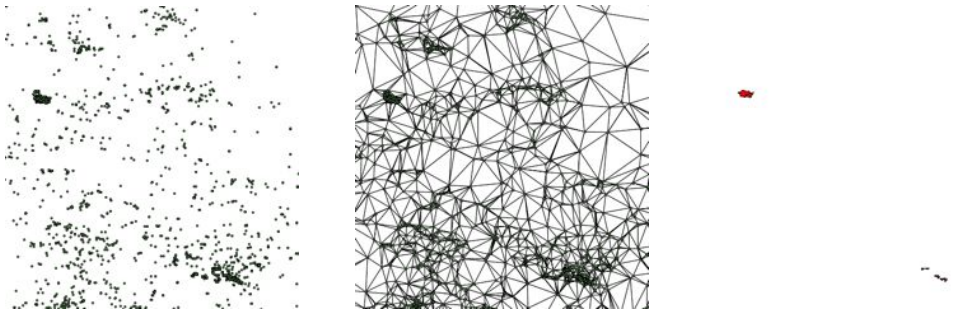
The first approach is based on the assumption that the first Node that a member creates is located in close proximity to his or her residence or mapping base. Usually, new members create their first new objects in areas that they are very familiar with and where local knowledge is very detailed.

The second method relies on the previously introduced changeset information. For each OSM member analyzed there were a certain number of changeset files available. Overlaying and blending these files allows the visualization of a particular area, which is covered in most of the overlaying files. Subsequently, the center point of this area can be calculated for the identification of the location or country of the OSM member.

The third approach identifies the location of an OSM member by analyzing the created Nodes. The country in which the member created most of its Nodes was used as the origin of the member.

The fourth and last method is the most comprehensive and most accurate approach to determining an activity area polygon of the member. The polygon represents an area in which the member focuses his or her activities when collecting information for the OSM project. To create the polygon for each member, all created Nodes of a member were meshed using a Delaunay triangulation (Lee and Schachter 1980) and a flipping algorithm (Berg et al. 2008). This creates a triangle mesh from all Nodes. Subsequently, all triangle edges and their points were removed from the network where

the edge lengths were longer than 1km. Figure 5.8 shows the processing steps for determining the polygons.



(a) (b) (c)

Figure 5.8.: Member activity area creation: (a) nodes of contributor, (b) triangulation, (c) edge-distance-filtering (final activity areas result).

It is important to note that for this particular method to determine the activity area polygon of a member, only Nodes that a member created were included, no edited Nodes or deleted Nodes were considered. Initial calculations that included all Nodes showed some irregularities, which were based on a software error in the OSM editors in the past (before 2011). This error increased the version number of a Node although the object was not changed in any way by any user directly, but because the Node would fall into the range of a certain changeset. Thus, the database would count a change to a Node, although the member did not actually edit the data. It is important to consider these errors when conducting similar studies to (Mooney and Corcoran 2012a, Haklay et al. 2010, Mooney and Corcoran 2012b, Mooney and Corcoran 2011), in which the versions of an OSM object should be based on real changes and not primarily on the number of editors and the absolute version number.

The results of the different methods that were applied showed that, based on the first approach by analyzing the first created Node of each member, a total of 167,000 members could be assigned to a particular country. Determining the center of all overlapping changeset areas allowed about 192,000 members to be assigned to a country. The analysis of all created Nodes by a member and the countries in which they were located helped to determine a country for almost 167,000 members. The difference between the number of members for which a country of origin could be determined when applying the two methods is either caused by Nodes that cannot be assigned to a particular country (e.g., Nodes in international waters), or the fact that the member did not create any Nodes at all. The most computationally intense, but also most ac-

curate method was able to generate an activity polygon in several countries for 123,000 members. In this case, the difference between the member numbers can be caused by insufficient amounts of Nodes to create a polygon.

Figure 5.9 shows the distribution of OSM members by country based on the results gathered from the different methods (country borders taken from the OSM project (OpenStreetMap 2011)). If an activity area polygon could be determined, the location of the member was chosen based on the center of the polygon. If this method did not provide the information needed, the country in which the member created the most Nodes was chosen. If this approach did not provide enough information, either the midpoint of the overlapping changesets or the member's first created Node determined the country of origin. In total 192,000 members could be associated with a country in which they showed their major OSM contributions.

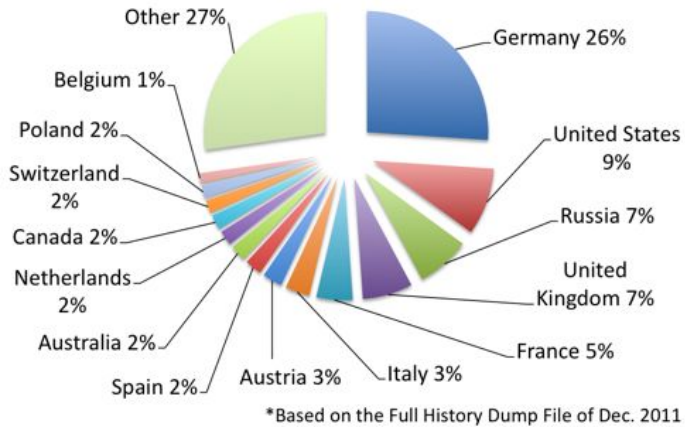
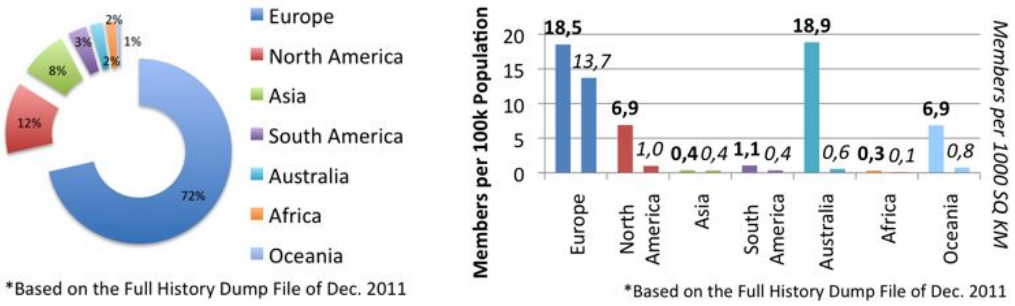


Figure 5.9.: Contributors per country*.

The result of the distribution analysis of OSM members highlights the concentration of the project in European countries at the end 2011. About 26% of the total members who have contributed to the project are working on the German dataset. In 2009, almost 50% of all changes in the database were made within Germany (Ramm 2009). Nearly 30% of all active OSM members collected information in Germany in mid-2011 (Neis et al. 2012). This value decreased to 26% in January 2012 (OSMStats 2012). Taking the aforementioned groups of contributors into consideration (Figure 5.4) the results showed that, with variations between 1 and 3%, all groups were represented in similar ways in each country, as shown in Figure 5.9. The comparison of the daily active member values for each country from the middle of November 2011 to the middle of December 2011 showed not significantly large differences of up to 3%. Figure 5.10(a)

shows the distribution of the 192,000 members by continents. Almost three-quarters of the total members of the project are from Europe. Prior research using a different approach to determine user origin has shown similar results (Budhathoki 2010).



(a) (b)

Figure 5.10.: (a) Contributors per continent* and (b) ratio of members to population per continent*.

Figure 5.10(b) illustrates the ratio of OSM members to the population of the continent based on values provided by (Population 2011). For these particular results, Australia surprisingly shows a similar magnitude as Europe. Considering the relationship between the number of members and the total area of the continent, Australia shows a very low value. However, these values could be based on varying population density factors. Overall, Europe shows the closest relation between number of members per 100,000 inhabitants and number of members per 1,000 km^2.

In addition to the studies that focus on the different countries of origin of the OSM members, a further analysis was conducted to evaluate the number of countries in which the different members created at least one Node. Figure 5.11 shows the distribution of the different contributor groups by the number of countries in which they have been collecting information. The results show that more than half of the members of the "Senior Mapper" group (Group 1) have contributed information about more than one country to the OSM project. These additional contributions may be based on several reasons, such as moving to another country, vacation, a business trip, or digitizing data from aerial photographs of foreign countries. Overall, approximately 86% of the members of Group 1 were active in up to four different countries, roughly 11% in five to ten countries and slightly less than 3% in more than ten different countries. In Group 2 ("Junior Mappers"), almost 86% contributed in one and 11% in two countries. Nearly 98% of the members of Group 3 ("Nonrecurring Mappers") were only active in

one country.

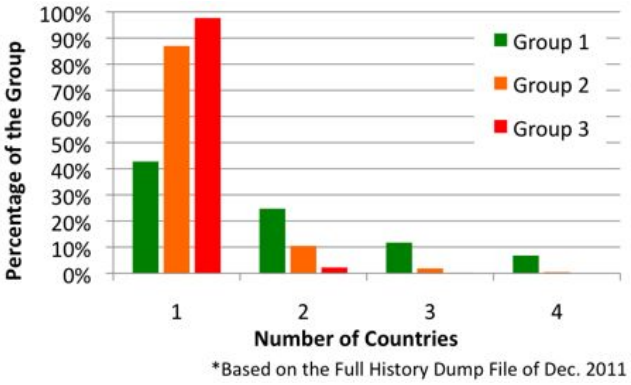


Figure 5.11.: Number of countries per OSM contributor group*.

5.5. Activity area of a member

To determine the country of origin of a member, a method was applied that determines a polygon representing the activity area of a member. It is based on the aforementioned triangulation of the Nodes. In this particular case a maximum value of one million Nodes of all created Nodes of a member were included in the analysis. Nodes that represented a boundary in the database, such as state or city limits, were excluded in a prior step. During the triangulation process, a minimum edge length between 10 and 500 m was adapted based on the number of Nodes created by the member. Thus the number of generated triangles was reduced to limit the consumption of resources during the calculation process. Also, the activity polygon of active OSM members is large and therefore does not require a triangle edge length of less than 500 m.

Nearly 760 million Nodes were included in the calculation process of the polygons for all OSM members. The smaller value compared to the total number of 1.47 billion existing Nodes from the full history file is caused by the filter that was applied to exclude boundaries and thereby limiting the number of Nodes per member to one million. The applied database of December 2011 provided 180 members who created or imported more than one million Nodes each.

During the triangulation process, approximately 370 million triangles were generated based on the 760 million Nodes. With the newly created triangles it was possible to determine the activity polygons for about 123,000 members. An example activity area of a member is shown in Figure 5.12. The distribution of activity area sizes for each of the three contributor groups (Figure 5.4) is displayed in Figure 5.13. For a better

overview, the area sizes have been divided into three individual size classes for each group.

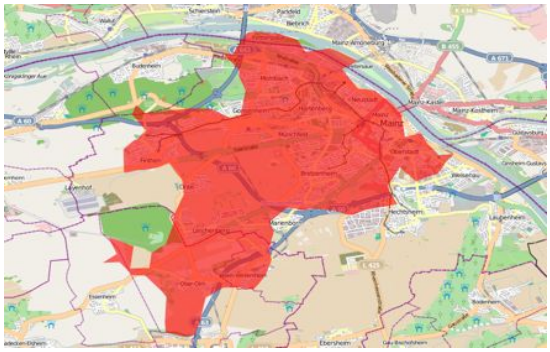


Figure 5.12.: Example activity area of a member of the OSM project.

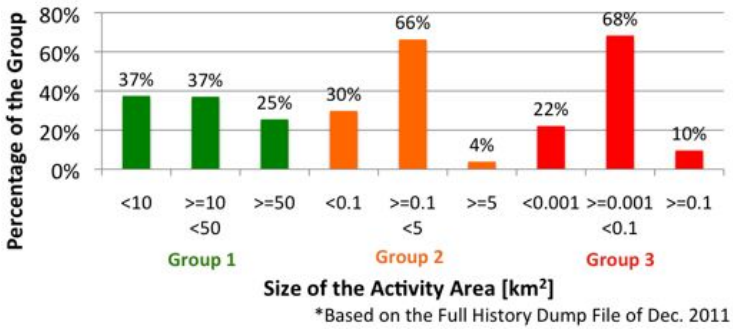


Figure 5.13.: Activity area sizes per OSM contributor group*.

For Group 1, which represents the most active OSM members, it could be determined that 37% of the 24,000 members have either an activity area of less than 10 km^2 or between 10 and 50 km^2; 25% of the group works in an area larger than 50 km^2. The second group shows a different pattern. In general, the activity areas tend to be much smaller compared to the first group. Almost exactly two-thirds of the 73,000 members of this group are active in an area between 0.1 and 5 km^2. The lower threshold of this activity area would be comparable to approximately 15 soccer fields (one soccer field is approximately 0.007 km^2) or more than one and a half times the size of the *central business district* (CBD) of London, England (2.9 km^2). For Group 3, with almost 96,000 members, it was not possible to generate an activity area polygon for all members because of the insufficient number of edits. However, more than two-thirds of the 26,000 members who provided enough information have an activity area between

one and 15 soccer fields.

In order to give additional information about the reliability of the generated activity areas, the number of created Nodes within the calculated areas was computed. The results provided in Figure 5.14 show that for Group 1, about 41% of the members have more than 66% of their created Nodes within the newly generated area, while for the remaining 59% of the members, this threshold could not be reached.

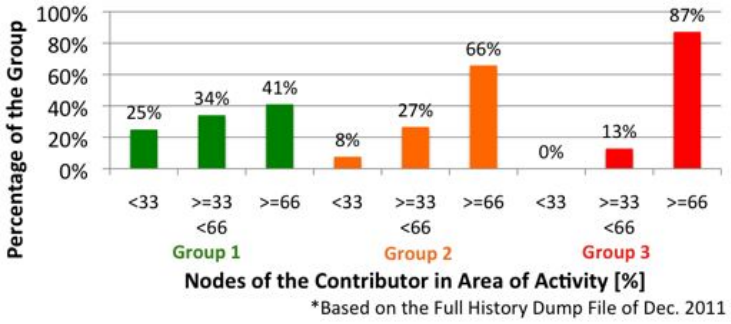


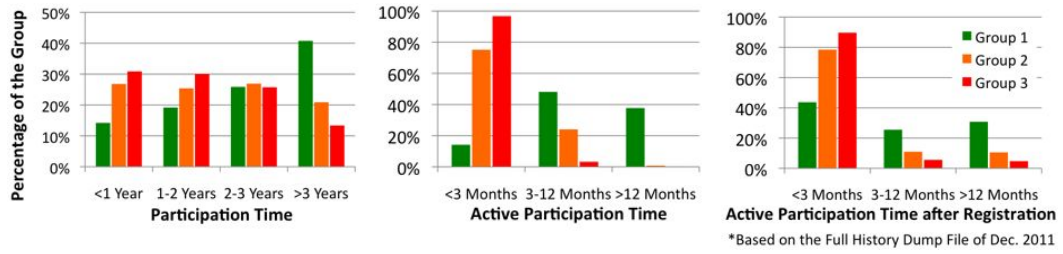
Figure 5.14.: Nodes of a contributor in area of activity*.

This supports the previously discussed results (Figure 5.11), which showed that more than half of the members in this group added new data in a variety of areas and did not focus only on one area, e.g., the home town. The second group of OSM members that was analyzed with respect to their activity areas showed that more than 66% of the created Nodes were within the generated activity area for almost two-thirds of the members. Group 3 showed a very high value for the group, with approximately 87% of the Nodes being within the activity area; however this value is less meaningful due to the small activity area.

5.6. Activity time frame of a member

Apart from the size of the area in which a member contributes data to the OSM project, it is obviously important in what time frame a member generates new data, i.e., how active a member is after registration for the project. Do members only collect data in the first few months, or can they be identified as long-term contributors? Figure 5.15(a) shows the percentage of each contributor group in relation to the years they have been registered for the project. The majority, 40%, of the most active contributor group (Group 1) has been registered for more than three years. However, it can also be determined that the increase of Group 1 members has not been consistent over

the years. On the other hand, an increase can be determined for Groups 2 and 3 in recent years. Figure 5.15(b) illustrates the actual time frame in which the members have been active since their registration. As expected, nearly all members in Group 3, who contributed the least amount of data to the project, were only active for less than three months. A similar pattern can be found in Group 2. Here about three-quarters of the members contributed for about three months, while the remaining members actively collected information for up to three to 12 months. In the most active Group 1, almost half of the members, 48%, contributed to the project actively between three to 12 months, while another 38% were involved for more than 12 months.



(a) (b) (c)

Figure 5.15.: (a) Participation, (b) active participation and (c) active participation after project registration.

The created changesets revealed the results shown in Figure 5.15(c) for our analysis of the most active times of the members. The bars in the diagram represent the average timeframe in which the members of the different groups performed their changesets. For Group 3, this means that almost 90% of the changes were created within the first three months after registering to the project, while only 6% were made after the first three to 12 months, and 5% within 12 months. The results of the second group show that 79% of all changes were made during the first three months, while changes between the third and twelfth months or later are fewer than 11%. Group 1 shows a slightly different pattern. Although an average of nearly 44% of all changes of a member are made within the first few months, the contributions of these members to the project can still last up to 12 months and more.

Additionally, an analysis was conducted that provides information about the percentage of members in each group who have made at least one change within the past six months, between six and twelve months, or within the past twelve months. Table 5.2 provides the results of the analysis in absolute and relative values. A slight decrease in activity can be determined for Groups 2 and 3. However, 60% of the members of

the Senior Mappers Group were active within the past six months or the past six to twelve months. In total, nearly three-quarters of all members of the most active group contributed information within the past 12 months.

Table 5.2.: Number of active members of the last six and 12 months (absolute and relative values).

Group	For the Last 6 Months	Between 6 and 12 Months	For the Last 12 Months
1	14,340 (59%)	14,350 (60%)	17,800 (74%)
2	20,800 (28%)	19,200 (26%)	34,650 (47%)
3	15,100 (22%)	13,840 (20%)	28,000 (40%)

5.7. Conclusions and future work

Various results of different analyses regarding the number of registered and truly active members of an online VGI community were presented in this article. To be able to retrieve the desired information, different datasets of the OSM project were investigated, all of which originate from the middle of December 2011. Several sources on the Internet have reported on the large number of contributors to the OSM project, which exceeded 500,000 registered members in December 2011. However, the results have shown that only 38% of the total number of members, around 192,000, carried out at least one change during their membership.

For a more detailed analysis, the members were divided into three groups according to their number of contributions to the project; Senior Mappers who created more than 1,000 Nodes, Junior Mappers who created 10 to 1,000 Nodes, and Nonrecurring Mappers who only created less than 10 Nodes. The Senior Mapper group represents the smallest group in the database with about 24,000 members. This means that only 5% of all members actively contribute to the project in a productive way. The other two contributor groups provide larger user numbers, with 73,100 for the Junior Mappers and 96,000 for the Nonrecurring Mappers group; their contributions, however, are very limited compared to the first group. Overall, 312,000 members never contributed to the project at all.

The evaluation of the changeset files of the project revealed that nearly the same number of members worked on the project every weekday with the exception of Sunday, which showed a slightly larger number. Further, almost 87,000 different members made at least one change to the OSM database in 2011. Breaking down the numbers per month, week, and day showed that 17,000, 6,500, and approximately 2,000 members

contributed to the project, respectively. These numbers indicate that roughly about 3% of all members made at least one change a month.

By applying four different methods, the countries of origin were determined for the 192,000 members who had completed at least one change in the database. The majority of the members are located in Europe (72%), while the remaining members (28%) are divided as follows: North America (12%), Asia (8%), South America (3%), Australia (2%), Africa (2%), and Oceania (1%). Further analysis showed that more than half of the Senior Mappers members collected information for the OSM project in at least two different countries.

The triangulation of all created Nodes in the database resulted in the determination of an activity area polygon for each of the OSM members. The results showed slightly different patterns for each contributor group. Two-thirds of the "Nonrecurring Mappers" (Group 3) has an activity area between one and 15 soccer fields in size. For the "Junior Mappers" (Group 2), the activity area increases for about two-thirds of the members to a size between 15 soccer fields and an area one and a half times the size of the CBD of London (2.9 km^2). The most active "Senior Mappers" (Group 1) can be divided into one-third of members that cover up to 10 km^2, one-third that cover between 10 and 50 km^2, and one-thirds that cover an area of about 50 km^2. Further research needs to be conducted to analyze whether and to what extent these numbers might change in the future.

An analysis of the timeframe in which the members contributed data to the project showed that the majority of the members contributed most of their information within the first three months of their membership. When comparing the conducted results with prior findings (Budhathoki 2010), there are a few similarities that can be addressed. Both analyses showed that only about a third of all members ever contributed to the project and that only a small number of contributors collected information over a longer period of time. Further research will provide more information about possible reasons for the reduced workload by the members. It may be based on adequately covered areas that do not need additional information or a general loss of interest for the project. However, these conclusions are only speculative and need to be researched in more detail. For the future of the project, these factors will play a major role, since VGI data does not solely rely on data collection, but also on maintaining the data to keep it as accurate and up-to-date as possible (Qian et al. 2009). Additionally, the development of the number of members per country in the coming years needs to be observed. An analysis that goes beyond the general activity of the members and focuses on changes within the activity areas of the members or whether members edit

or improve the objects of other members could be conducted as well. Some first investigations regarding these factors have already been published (Mooney and Corcoran 2012b).

Additionally to these suggested analyses with focus on user activities, the results gathered in this paper could provide a valuable foundation to answering questions regarding VGI data quality such as: Which type of contributors (e.g., Senior Mappers) created a particular dataset of interest? How many activity areas of members intersect with each other or within a predefined area? Similar to the approach used in the analysis of Wikipedia and "The Roles of Local and Global Inequality Contribution" (Arazy and Nov 2010), it could be tested if the quality of OSM data varies, depending on whether the member who edited the information is very familiar with the area or not.

References

Anthony, D., Smith, S. W., and Williamson, T. (2007). The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia. In: *Dartmouth Computer Science Technical Report TR2007-606*. Hanover, NH: Dartmouth College.

Arazy, O. and Nov, O. (2010). Determinants of Wikipedia Quality: The Roles of Global and Local Contribution Inequality. In: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW)*. (Feb. 6–10, 2010). Savannah, GA, USA.

Bennett, J. (2010). *OpenStreetMap: Be Your Own Cartographer*. 1st. Birmingham, UK: Packt Publishing.

Berg, M.D., Cheong, O., Kreveld, M.V., and Overmars, M. (2008). *Computational Geometry: Algorithms and Applications*. 3rd. Santa Clara, CA, USA: TELO.

Budhathoki, N. (2010). Participants' Motivations to Contribute to Geographic Information in an Online Community. Ph.D. Dissertation. University of Illinois, Urbana-Champaign, Urbana, IL, USA.

Coleman, D., Georgiadou, Y., and Labonte, Y. (2009). Volunteered Geographic Information: The Nature and Motivation of Producers. *International Journal of Spatial Data Infrastructures Research*, 4, 332–358.

Foursquare (2012). *Foursquare is Joining the OpenStreetMap Movement!* URL: `http://blog.foursquare.com/2012/02/29/foursquare-is-joining-the-openstreetmap-movement-say-hi-to-pretty-new-maps/` (visited on 03/19/2012).

Girres, J.F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS*, 14, 435–459.

Goetz, M. and Zipf, A. (2011). Extending OpenStreetMap to Indoor Environments: Bringing Volunteered Geographic Information to the Next Level. In: *Urban and Regional Data Management: Udms Annual 2011*. Delft, The Netherlands.

Goodchild, M.F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211–221.

Google (2011). *Introduction of Usage Limits to the Maps API - Google Maps*. URL: `http://googlegeodevelopers.blogspot.de/2011/10/introduction-of-usage-limits-to-maps.html` (visited on 06/27/2012).

Google (2012). *Map Maker*. URL: `http://groups.google.com/group/google-map-maker/browse_thread/thread/7ba81462f965c1dd` (visited on 03/07/2012).

Haklay, M. (2010). How good is OpenStreetMap Information: A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. *Environment and Planning B*, 37, 682–703.

Haklay, M., Basiouka, S., Antoniou, V., and Ather, A. (2010). How many Volunteers does it take to Map an Area well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47, 315–322.

Haklay, M. and Weber, P. (2008). OpenStreetMap: User-generated street map. *IEEE Pervasive Computing*, 7, 12–18.

Heipke, C. (2010). Crowdsourcing Geospatial Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65, 550–557.

Javanmardi, S., Ganjisaffar, Y., Lopes, C., and Baldi, P. (2009). User Contribution and Trust in Wikipedia. In: *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing*. (Nov. 11–14, 2009). Washington, DC, USA.

Knorr, E. (2003). *2004: The Year of Web Services*. URL: `http://www.cio.com/article/32050/2004_The_Year_of_Web_Services` (visited on 03/07/2012).

Lee, D.T. and Schachter, B.J. (1980). Two Algorithms for Constructing Delaunay Triangulation. *International Journal of Computer and Information Sciences*, 9, 219–242.

Lin, Y. (2011). A Qualitative Enquiry into OpenStreetMap Making. *New Review of Hypermedia and Multimedia*, 17, 53–71.

Mooney, P. and Corcoran, P. (2011). Accessing the History of Objects in OpenStreetMap. In: *Proceedings of AGILE 2011: The 14th AGILE International Conference on Geographic Information Science*. (Apr. 18–21, 2011). Utrecht, The Netherlands.

Mooney, P. and Corcoran, P. (2012a). Characteristics of heavily edited Objects in OpenStreetMap. *Future Internet*, 4, 285–305.

Mooney, P. and Corcoran, P. (2012b). The Annotation Process in OpenStreetMap. *Transaction in GIS*, 16, 561– 579.

Neis, P., Zielstra, D., and Zipf, A. (2012). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4, 1–21.

Neis, P. and Zipf, A. (2008). OpenRouteService.org is Three Times "Open": Combining OpenSource, OpenLS and OpenStreetMaps. In: *Proceedings of the GIS Research UK 16th Annual conference GISRUK 2008*. (Apr. 2–4, 2008). Manchester, UK.

Nestoria (2012). *Why (and How) We've Switched Away from Google Maps*. URL: `http://blog.nestoria.co.uk/why-and-how-weve-switched-away-from-google-ma` (visited on 03/19/2012).

Nielsen, J. (2006). *Participation Inequality: Encouraging More Users to Contribute*. URL: `http://www.nngroup.com/articles/participation-inequality/` (visited on 09/23/2013).

Nokia (2012). *Map Creator*. URL: `http://blog.maps.nokia.com/travelling-out-and-about/introducing-nokia-map-creator` (visited on 03/07/2012).

OpenStreetMap (2011). *World Boundaries - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Mapnik#World_boundaries` (visited on 12/01/2011).

OpenStreetMap (2012a). *Bing - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Bing` (visited on 07/28/2012).

OpenStreetMap (2012b). *Statistics*. URL: `http://www.osm.org/stats/data_stats.html` (visited on 07/28/2012).

OpenStreetMap (2012c). *Stats - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Stats` (visited on 03/12/2012).

OpenStreetMap (2012d). *Yahoo! Aerial Imagery - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Yahoo!_Aerial_Imagery` (visited on 07/28/2012).

O'Reilly, T. (2005). *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. URL: `http://oreilly.com/web2/archive/what-is-web-20.html` (visited on 09/21/2013).

OSMStats (2012). *Statistics of the Free Wiki World Map*. URL: `http://osmstats.altogetherlost.com` (visited on 02/01/2012).

Over, M., Schilling, A., Neubauer, S., and Zipf, A. (2010). Generating Web-based 3D City Models from OpenStreetMap: The Current Situation in Germany. *Computers, Environment and Urban Systems*, 34, 496–507.

Population, World (2011). *Data Sheet*. URL: http://www.prb.org/Publications/Datasheets/2011/world-population-data-sheet/data-sheet.aspx (visited on 12/12/2011).

Qian, X., Di, L., Li, D., Li, P., Shi, L., and Cai, L. (2009). Data Cleaning Approaches in Web 2.0 VGI Application. In: *Proceedings of the 17th International Conference on Geoinformatics*. (Aug. 12–14, 2009). Fairfax, VA, USA.

Ramm, F. (2009). *Krautsourcing 2.0 Beta - The State of Germany. State of the Map, Amsterdam*. URL: http://www.geofabrik.de/media/2009-07-11-sotm-state-of-germany.pdf (visited on 11/28/2011).

Ramm, F. and Stark, H.J. (2008). Crowdsourcing Geodata. *Géomatique Suisse*, 6, 315–318.

Ramm, F., Topf, J., and Chilton, S. (2010). Cambridge, UK: UIT.

TomTom (2012). *Map Share*. URL: http://www.tomtom.com/en_gb/maps/map-share/ (visited on 03/07/2012).

Wikipedia (2012). *Statistics about Wikipedia*. URL: http://en.wikipedia.org/wiki/Special:Statistics (visited on 02/20/2012).

Wunsch-Vincent, S. and Vickery, G. (2007). Participative Web: User-Created Content: Web 2.0, Wikis and Social Networking. In: *Organisation for Economic Co-operation and Development*. Paris, France.

Zielstra, D. and Hochmair, H.H. (2011). Digital Street Data: Free versus Proprietary. *GIM International*, 25, 29–33.

Zielstra, D. and Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

6. The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011

Authors

Pascal Neis, Dennis Zielstra, and Alexander Zipf

Journal

Future Internet

Status

Published: 29 December 2011 / Accepted: 19 December 2011 / Revised: 16 December 2011 / Submitted: 1 December 2011

Reference

Neis, P., Zielstra, D., and Zipf, A. (2012). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4(1):1-21.

Contribution statement

Pascal Neis conducted all analyses for this study and wrote the majority of the manuscript. Both co-authors supported this publication through continuing discussions about the methods and the results of the study. Furthermore, extensive proof-reading by both co-authors led to substantial improvements to the manuscript.

.....	.....
Prof. Dr. Alexander Zipf	Dennis Zielstra

Abstract

The *OpenStreetMap* (OSM) project is a prime example in the field of *Volunteered Geographic Information* (VGI). Worldwide, several hundred thousand people are currently contributing information to the "free" geodatabase. However, the data contributions show a geographically heterogeneous pattern around the globe. Germany counts as one of the most active countries in OSM; thus, the German street network has undergone an extensive development in recent years. The question that remains is this: How does the street network perform in a relative comparison with a commercial dataset? By means of a variety of studies, we show that the difference between the OSM street network for car navigation in Germany and a comparable proprietary dataset was only 9% in June 2011. The results of our analysis regarding the entire street network showed that OSM even exceeds the information provided by the proprietary dataset by 27%. Further analyses show on what scale errors can be reckoned with in the topology of the street network, and the completeness of turn restrictions and street name information. In addition to the analyses conducted over the past few years, projections have additionally been made about the point in time by which the OSM dataset for Germany can be considered "complete" in relative comparison to a commercial dataset.

Keywords: Volunteered Geographic Information (VGI); OpenStreetMap; geodata; quality assessment; Germany; street network.

6.1. Introduction

The *OpenStreetMap* (OSM) project has a history of nearly seven years now (2011). Similar to Wikipedia, the information gathered can be described as *User-Generated Content* (UGC) (Anderson 2007, Diaz et al. 2011). However, unlike Wikipedia, it is not encyclopedia information that is being gathered; instead, users are contributing their geodata to OSM. This data of geographical relevance is compiled by volunteers, saved in a database, and made 'freely' available to everyone via the *World Wide Web* (WWW) (Coast 2007, Nelson et al. 2006). OSM is a well-known project in the field of *Volunteered Geographic Information* (VGI) (Elwood 2008, Goodchild 2007a, Goodchild 2007b), which others also describe as crowd sourced (geodata) (Chilton 2012, Heipke 2010, Hudson-Smith et al. 2009a, Ramm and Stark 2008). Furthermore, especially in connection with the term of Web 2.0 (O'Reilly 2005), it is also referred to as Neogeography (Goodchild 2008, Hudson-Smith et al. 2009b, Rana and Joliveau 2009, Turner 2006, Walsh 2008). Others again describe it as Collaborative Mapping (Fischer 2008)

or the Wikification of GIS (Sui 2008). The successful development of user generated content in recent years had an increasing impact on a variety of research fields and particularly OSM has been the focus of many new developments such as routing applications, 3D city models and *Location-based Services* (LBS) (Fritz et al. 2009, Mooney and Corcoran 2010, Neis and Zipf 2008, Over et al. 2010).

Within the past few years, the OSM membership numbers have rapidly developed from a few hundred in mid-2004 to more than half a million registered members in November 2011. But what drives such a large number of people to participate in a voluntary project, and where lies their motivation? A few suggestions have been made to describe the motivation of the volunteers, such as a certain need for self-representation by the members or the project's fun factor, and a degree of interest in technical terminology, equipment, and the WWW (Goodchild 2007a). Either way, according to our research, approximately 150 new active members have joined the project each day since the beginning of 2011. Despite these numbers, OSM tends to experience a problem that is similar to other online portals that are based on UGC: participation inequality. This term describes the phenomenon of a 90-9-1 rule that most of the projects inherit (Nielsen 2006). Ninety percent of all members merely consume and are described as lurkers, 9% contribute to the project at irregular intervals, and just 1% of the members is actively involved and counts for the largest number of contributions to the project. A similar scenario exists within the OSM project. At the beginning of 2008 there were 30,000 total members in OSM, and about 10% of them actively contributed to the project (Ramm and Stark 2008). In 2009, there were approximately 200,000 registered members, yet still, only roughly 10% of these were active members (Ramm 2009a) and contributed to the project. In the same year 98% of all project data was provided by approximately 5% about 10,000 members. In 2010, only 5% of the 330,000 members were active contributors, and 98% of the data was provided by about 3.5%, which represents 12,000 of the total registered members.

While some countries such as the USA and France imported large datasets from other providers of freely available data (e.g., from governmental agencies such as the US Census Bureau), within the scope of the project, in Germany, OSM relies on its large number of participants. In 2009 nearly 50% of the entire changes in the OSM database were made within Germany (Ramm 2009b). However, in 2010 this value lowered to approximately 30%.

Despite this decrease, the aforementioned numbers give a first impression of how the OSM project has become a potential competitor to public and commercial geodata providers, not just in Germany but also worldwide. We see a revolutionary paradigm

shift on how map data is being collected. MapQuest and Bing Maps are some of the first international companies that are reacting to this trend and have started offering maps based on OSM data (e.g., open.mapquest.com) or additional LBS's such as route planning, and address and area search functions on their websites.

Although the OSM project shows a high membership number and contributors are active, it needs to be noted that most of the VGI projects, and thus also OSM, rely on volunteers that do not necessarily have professional qualifications and background in geodata collection or surveying (Goodchild 2007a). Furthermore, contributing to the project depends largely on technical aspects such as specific equipment, e.g., a PC/laptop, an internet connection, or potentially also a GPS receiver or Smartphone. The population density of the specific areas naturally plays a role too. Thus, the probability that more densely populated areas are mapped or more complete than sparsely populated areas is beyond question. However, the local knowledge of most participants should in fact make them local experts (Goodchild 2009). This raises numerous questions: How does the data perform in comparison (relative completeness and attribute accuracy) with other proprietary geodata providers? Can a difference be detected between urban and rural areas? How has the OSM project data developed in recent years? Can a projection be made for the data's future development?

The remainder of the paper is structured as follows: The following section gives an overview of prior OSM quality research conducted in recent years. Also, the study area and data preparation steps applied during our research are being discussed. The next section provides the results conducted during our research with regards to the OSM street network evolution for the years 2007 to 2011. This is followed by a discussion of the results and aspects of future work.

6.2. OpenStreetMap quality assessment history

Numerous scientists have investigated the quality of VGI and particularly OSM in recent years, and further research is currently still being conducted. In 2008 a discussion about the need for research with regard to the accuracy and correctness of the compiled information within the world of Web 2.0 was sparked (Nelson et al. 2006, Goodchild 2007b, Flanagin and Metzger 2008, Maué and Schade 2008). Preliminary direct comparison analyses with regard to OSM were conducted for Great Britain in 2008 comparing *Ordnance Survey* (OS) geodata with OSM (Ather 2009, Haklay 2010); in Germany in 2009 (Zielstra and Zipf 2010) OSM data has been compared with the commercial Multinet dataset from TomTom (still known as TeleAtlas at the time).

A few months later a similar comparison for Germany was conducted, but with the street dataset from Navteq, a different proprietary geodata provider (Ludwig et al. 2010). Both analyses came to similar conclusions despite using slightly different methods: OpenStreetMap data shows a high degree of detail in urban areas; however, this detail richness declines significantly in rural areas. The main difference between the two studies was that one analysis included methodology to show the geographical discrepancies within Germany (Zielstra and Zipf 2010), while the other merely advised how complete OSM is in relative comparison with another dataset.

In France, a similar approach was used to analyze OSM data and further studies were conducted at the same time (Girres and Touya 2010). The results showed the advantage and flexibility, but also the problem of the heterogeneity of the data specifically for this country. The latter is the result of the different data sources that have been used in OSM and also the differences in the work by the project participants in France.

In 2011, the first studies that analyzed the quality of OSM outside of Europe were conducted (Zielstra and Hochmair 2011b). In this particular case the OSM project data has been compared with proprietary data from TomTom (TeleAtlas) and Navteq for the entire state of Florida (USA) and four specific cities within the USA. In comparison to the results for Germany or England, the discrepancies between the rural and urban areas in the USA showed an opposite tendency. In Florida, the rural data was, in parts, even more complete than that of the proprietary datasets in the relative comparison conducted. This can probably be attributed to the TIGER (U.S. Census Bureau) street data import to the OSM database for the entire United States. Other analyses were conducted comparing the impact of OSM on shortest path generation for pedestrians in Germany and the US (Zielstra and Hochmair 2011a, Zielstra and Hochmair 2012). Apart from England, no studies have been conducted to date over a period of several years and for an entire country (Haklay and Ellul 2011).

6.2.1. Study area and data preparation

A long history in geodata quality research has provided a variety of publications dealing with the definition of characteristics of geodata that can be used as quality parameters (Brassel et al. 1995, Van Oort 2006). In 2002 the International Organization of Standardization (ISO) set a standard that defines the quality attributes of geodata in ISO 19113:2002 (principles for describing the quality of geographic data) and ISO 19114:2003 (framework for procedures for determining and evaluating quality). The defined parameters for the quality of geodata according to ISO 19113:2002 are completeness, logical consistency, positional accuracy, temporal accuracy, and thematic

accuracy.

Within the scope of this article, we shall consider all these parameters with the exceptions of positional and thematic accuracy. The completeness of the street network is determined via a relative comparison between OSM and a commercial dataset provider. Furthermore, we will show the development in the urban and rural areas over a certain time period. The logical consistency will subsequently be evaluated with the help of an internal test, whereby topological and thematic consistency will be determined. The temporal accuracy will be verified in a simple form by means of an object‘s time stamp in the OSM dataset.

The study area for this article relates to all of Germany; however, each of the studies conducted takes place on different scales with the smallest scale being the municipal level. A variety of OSM datasets with three-month intervals were used starting from January 2007 to June 2011. Overall, 19 datasets were prepared for Germany, representing the different dates. In addition to clipping the Germany dataset from the entire OSM database dump-file[1] for each point in time, it proved to be a challenge to work with the different API (application programming interface) versions to extract the data. The first dataset included in the analysis (2007) was taken from OSM API Version 0.4; however; Version 0.6 is the latest version (2011) of the API. The data could be converted, but unfortunately it did not always feature all of the latest attributes. With the older version of the API ways would still consist of segments in the OSM database, while with API 0.5 ways were mapped by referencing to a node. Furthermore, since API 0.6 appeared in April 2009, anonymous edits in OSM are no longer permitted, and a user ID and user name have since been included with every object. This means that at certain points during our analysis it was not possible to retrieve information for the entire time frame up to 2007, but instead only to 2009. For the comparison analyses, the TomTom Multinet 2011 commercial dataset has been used. The respective data was imported into a PostgreSQL/PostGIS database with OSM‘s OSMOSIS program. In contrast to other available OSM programs that import the data into a database, the application used has the advantage of not filtering, preparing, or optimizing the data during the import procedure to the database. All analysis procedures in this paper are either based on PostgreSQL/PostGIS functions or specific tools developed in Java implementing the GeoTools open source toolkit.

[1] http://planet.openstreetmap.org/

6.3. OSM street network evolution

6.3.1. User activity and data development

The number of OSM participants in Germany increases from year to year. To date (June 2011), a total of more than 40,000 different members have actively contributed to the project. Slightly different numbers of contributors have generated the three OSM object types: Nodes, Ways and Relations (Figure 6.1). Another interesting fact, also presented in Figure 6.1, is shown by the lines in orange, light blue, and purple, representing the users that generated a total of 98% of the data volume of each respective object type. Here 98% of about 74 million Nodes can be attributed to approximately 8,500 members, 98% of about 11 million Ways to approximately 7500, and 98% of about 171,000 Relations to approximately 2,600. These numbers are based on the information in the database on who has been saved as the last owner of each Node, Way, or Relation.

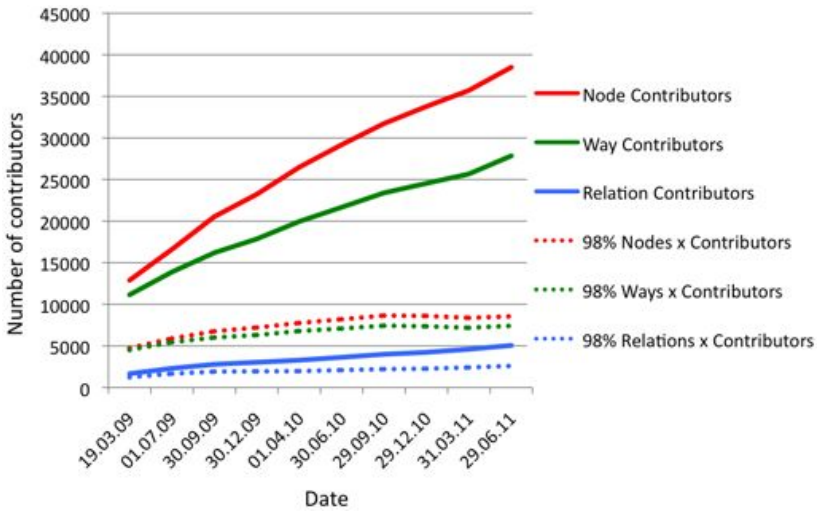


Figure 6.1.: Number of OSM contributors in Germany from 2009 to 2011.

To be able to give more information about the results conducted during our analysis, we need to introduce the three object types that are used in the OSM project/database in greater detail. A Node is the basic object in the database and constitutes a coordinate. Ways represent lines or surface objects and constitute references to Nodes. Objects can be linked via Relations, which relate to each other. The results of our analysis showed a general pattern of an increase in the number of OSM Node and Way objects over the past few years (2007–2011) (Figure 6.2 & 6.3), which was expected

due to the general trend that OSM showed in Germany in recent years.

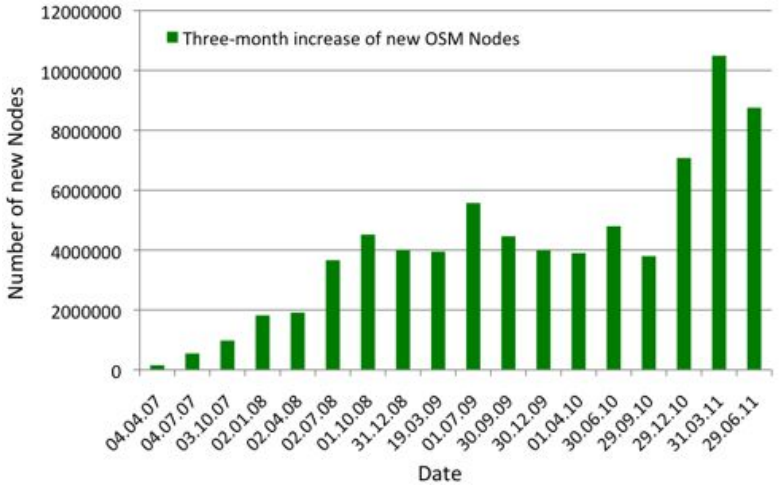


Figure 6.2.: Development of OSM nodes in Germany

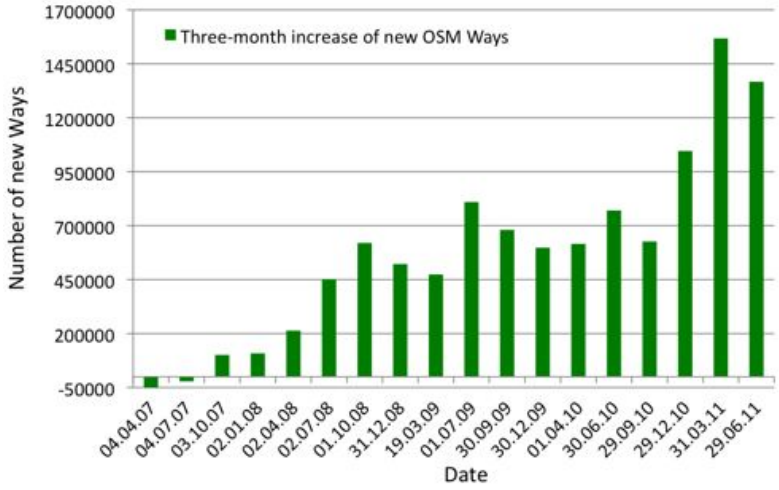


Figure 6.3.: Development of OSM ways in Germany.

However, due to the three-month interval used in this analysis, other factors can be interpreted and distinguished. The data clearly shows that during the summer months the members are more active than during the winter months. Also, an above-average increase in data can be noticed at the turn of the year 2010/2011 and in spring 2011. The high proportion of new objects during these points in time can be attributed to the release of the Bing aerial images for digitalization purposes. The negative trend

for Ways in OSM at the beginning of 2007 is due to the API changes during that year. With these changes the data schema and representation has been adjusted, and the total number of Ways was affected in this way; however, no data was lost by this change.

In prior studies the development of the German street network has only been compared to a commercial dataset (TomTom) for a period of eight months (Zielstra and Zipf 2010). Neither definite statements about when different street types can be considered relatively complete nor a projection for the future were given. In our analysis the strongest increase in transportation-related routes in OSM for Germany to date could be distinguished in the third quarter of 2008 (+180,000 km) (cf. Figure 6.4 & 6.5). The year 2008 was also when the most transport routes were added to the OSM database in general with a total length of almost 530,000 km. Since 2008 the annual expansion has decreased over time, and a slight change is discernible for the first half of 2011, where the trend slightly increases again. However, if this tendency continues for the rest of 2011, a small increase in the total street network will be detected for this specific year.

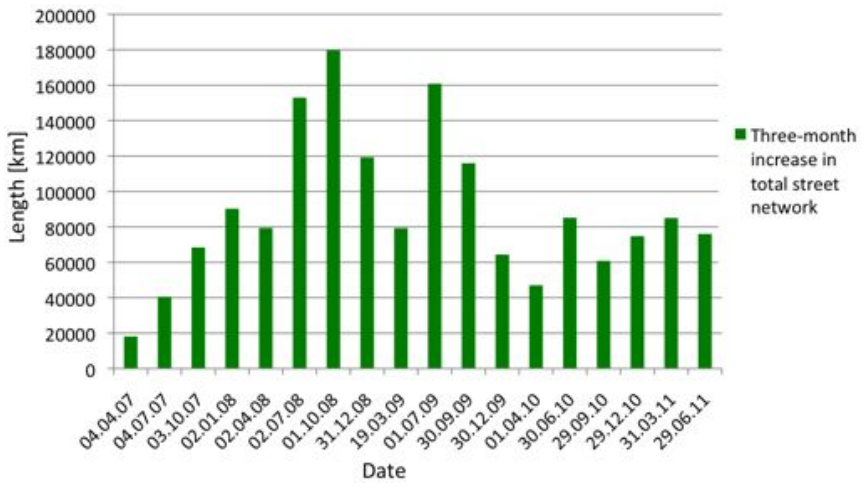


Figure 6.4.: Increase in German OSM street network (three-month interval).

After gaining this first general impression about the development of the German OSM dataset, we conducted further analyses to give more detailed information about the street network. Due to the fact that every country‘s street network consists of several different street categories, it seemed mandatory to consider these in our analysis. Thus, the various different street categories, which can also be found on the OSM Map

Features web page[2], have been divided into four groups for the sake of clarity and for enhanced research and comparison methods; namely, motorways/dual carriageway, district/municipal roads, roads to/in residential areas, and other roads such as service roads and dirt/forest trails (Figure 6.6).

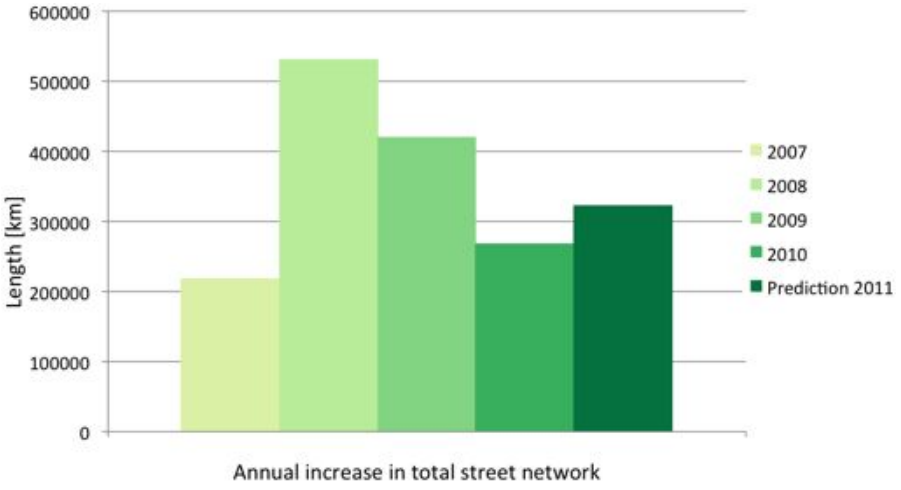


Figure 6.5.: Annual increase in German OSM street network (2007–2011).

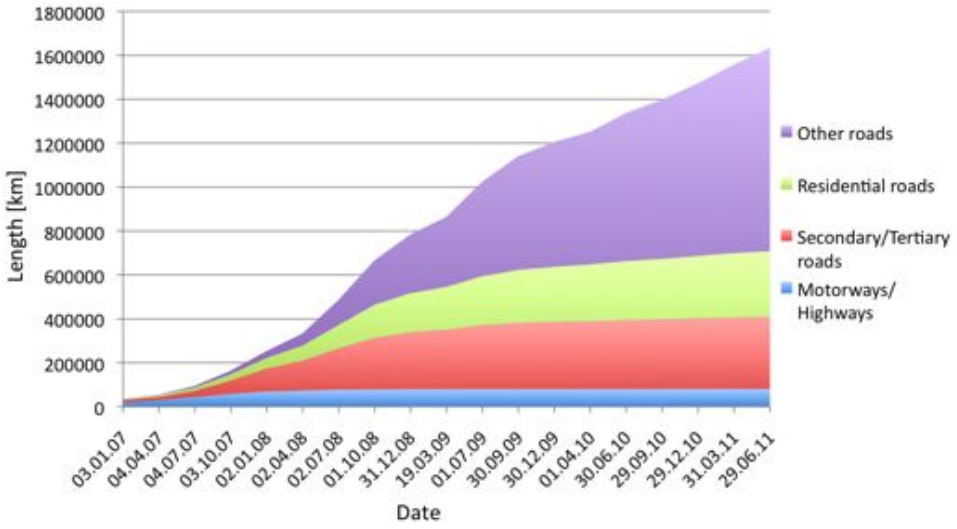


Figure 6.6.: Development of OSM street network in Germany by street category (2007–2011).

[2]http://wiki.openstreetmap.org/wiki/Map_Features

Tracing the growth of the different categories, it can be noted that from a specific point in time, most categories do not expand any further. This indicates which category should be close to "completion" or in which category there are still new streets being added. It needs to be noted though that for comparison the TomTom dataset is suitable only for street network data for car-specific navigation, three out of the four categories. The "other routes" category can be compared only to TomTom to a certain degree. In this fourth category, OSM has a far higher street network than the commercial provider. Based on the presumptions stated above and the comparison with the corresponding TomTom category street lengths, we reached the following conclusions. First, motorways and expressways were completely recorded for Germany by the middle of 2008. Second, all municipal roads for all of Germany were recorded by the middle of 2009. Third, streets that are close to or within residential areas are not fully recorded yet. Fourth, at the end of 2009 there were more segments in the "other routes" class in OSM than in the total TomTom commercial dataset. Fifth, in the middle of 2010 OSM surpassed TomTom in the total number of streets recorded. However, a high number of field and forest trails caused this advantage for OSM. Finally, most data contributions in 2011 are isolated street networks close to or within residential areas, but "other route" data, such as forest and field trails, are also increasing.

The development of the individual categories in comparison to prior research results (Zielstra and Zipf 2010) and the commercial TomTom Multinet dataset from 2011 is depicted in Figure 6.7. The assumptions mentioned above with regard to the development and completeness of the total street network can here be confirmed too.

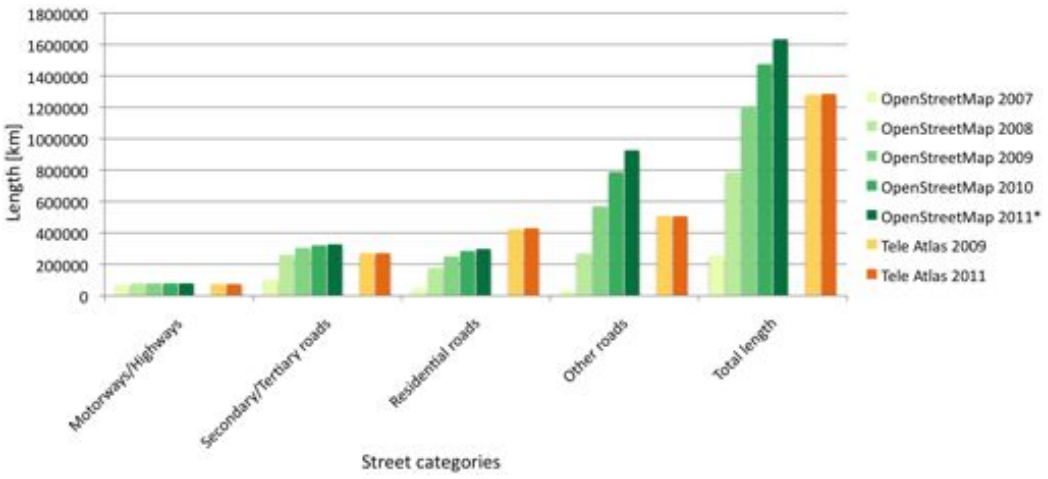


Figure 6.7.: Development of OSM street network in comparison to TomTom.

In June 2011, our studies for Germany showed that OSM had provided a street network for car navigation that is approximately 9% smaller than that of TomTom (Table 6.1). However, OSM's total street network is approximately 27% larger in comparison with TomTom's. In terms of pedestrian-related data and information, the OSM Germany dataset is even approximately 31% larger.

Table 6.1.: Total street length of TomTom Multinet 2011 and OSM in June 2011.

Street Network	TomTom Multinet 2011	OSM June 2011	%
Total street network	Approximately 1,283,000 km	Approximately 1,630,000 km	OSM 27% longer street network
Street network for car navigation	Approximately 777,000 km	Approximately 705,000 km	TomTom 9% longer street network
Street network for pedestrian navigation	Approximately 1,185,000 km	Approximately 1,552,000 km	OSM 31% longer street network

In addition to the relative geometric completeness in comparison with another dataset, the internal completeness within the street network with regard to the street names is also important. This factor, sometimes also referred to as attribute accuracy (Ather 2009), plays a significant role in applications such as routing applications that are being built on the specific dataset. Our results showed that a total of approximately 16% of streets in OSM have neither a name nor a route number (e.g., A 61) that could be used for car navigation. However, these results vary by street type in significant ways (Figure 6.8).

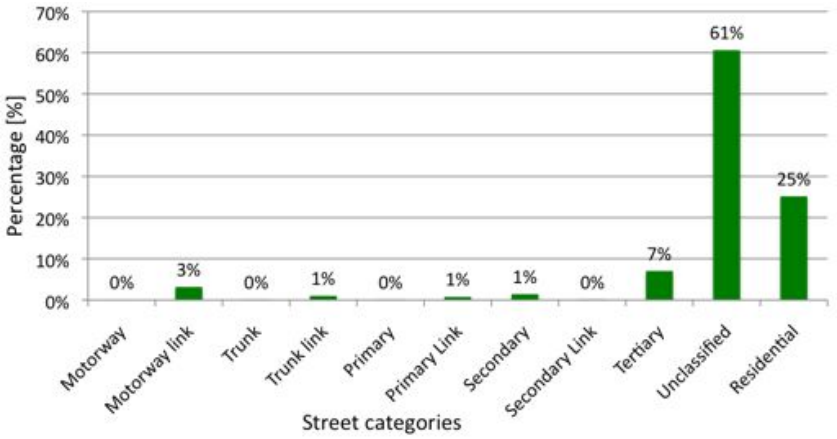


Figure 6.8.: Distribution of streets without name or route number attribute information by street category (June 2011).

The results clearly show that the majority of the unnamed streets are streets that are either within or close to residential areas. The "unclassified" street category could lead to confusion in this case, since streets that have a linking function between villages are included within this category. Another reason for this high value could be the fact that many of these particular routes (e.g., country lanes) have been digitized from satellite images, thus the local knowledge to add the specific name of each route is missing.

6.3.2. Data completeness and population density

For further, more detailed studies of the route length of the total street network, the dataset was divided into the smallest possible German administrative units: municipalities and town boundaries. Detailed presumptions about the data development and the relative completeness with regard to population and area can be provided as a consequence of calculating the length of the route network for the different modes of transportation within the specified boundaries.

The administrative areas used in our analysis (12,387 in total) feature the number of inhabitants for the years 2008 and 2009 and were obtained from the TomTom Multinet dataset. The entire administrative area dataset is subdivided into six groups considering different population numbers. The first group ($\geq$1,000,000) represents metropolitan areas; the second group ($\geq$500,000 and $<$1,000,000) large towns; the third ($\geq$100,000 and $<$500,000) towns; the fourth ($\geq$50,000 and $<$100,000) medium-sized towns; the fifth ($\geq$10,000 and $<$50,000) small towns; and the last ($<$10,000) rural towns. With regard to the entire administrative area of Germany, this means that approximately 73% of the entire population lives in population groups one to five, covering one third of the entire area of Germany. Conversely, around 27% of the population lives in population group six (rural towns) and is distributed over two-thirds of the total area of Germany.

For our analysis we considered the development of three different street networks: total street network, and car and pedestrian networks (Figure 6.9). The rows in Figure 6.9 visualize the expansion for each individual network and percentage of new data per year. It is evident that over the past four years, the route network of the individual groups has developed in correlation to their population density. While the general route network has been less active in the more densely populated areas, an increase in new data can still be seen in the more sparsely populated areas. It is also clearly discernible that the largest overall increase in new streets occurred in 2008.

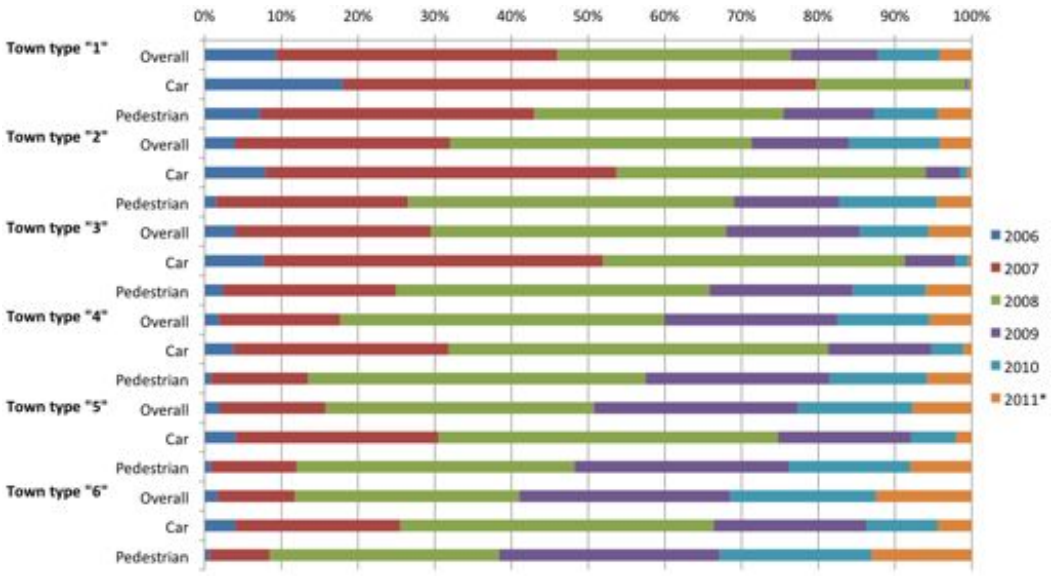


Figure 6.9.: Development of OSM street network by town type (June 2011).

Another aspect that has been included in our analysis was the difference in total length of the route networks by town or municipality (Figure 6.10). The results showed that the aforementioned approximately 9% of missing data is mainly distributed over the sparsely populated areas. It is also clearly discernible that OSM provides more overall data in comparison to the proprietary dataset with regard to the total and pedestrian route network length.

When expressed in route network lengths, this means that in mid-2011, OSM was still lacking approximately 3% (21,000 km) in the small-town population group and approximately 6% (48,000 km) in the rural-town population group. Using these highly detailed studies for the increase in street data for the different town types and the analyses of the differences in route network lengths in comparison with a commercial dataset, projections could be made of the time frame within which the dataset could be completed (at least in a relative comparison to another dataset, since neither of the datasets represent ground truth). As Figure 6.10 indicates, there is currently still a lack of data for less densely populated areas in OSM. Figure 6.9 shows the development of data by population group. In line with the expansion rate of this graph, 6% (14,000 km) of new streets were added to group 5 for car navigation in 2010 and nearly 10% (33,000 km) to group 6. By mid-2011, 2% (5,000 km) of new streets were added to group 5 and slightly less than 4% (16,000 km) to group 6. This means that if there is an increase in new street data that remains at least at the same level and does not

decline as shown in Figure 6.4 & 6.5, the German street network for population groups 5 and 6 will be almost completely covered by the middle to end of 2012.

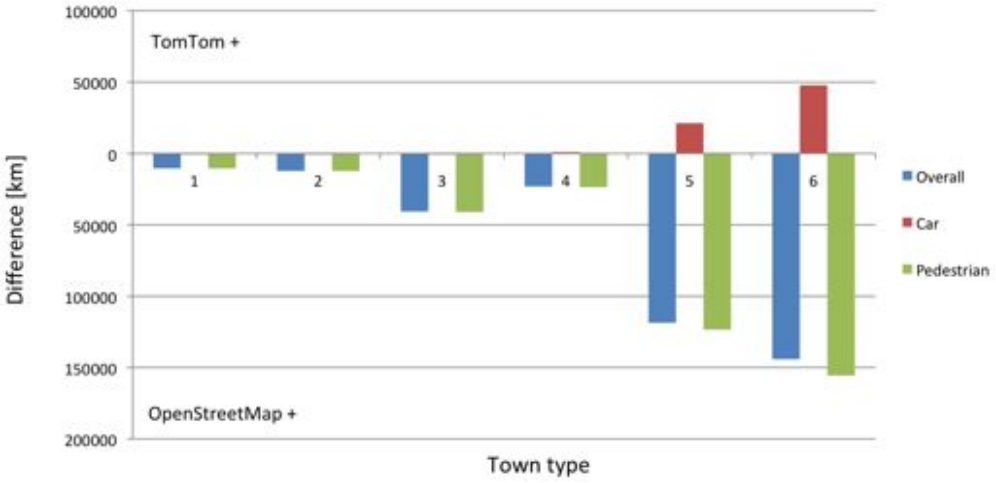


Figure 6.10.: Relative difference by town type and street network (June 2011).

With regard to the correlation between TomTom's commercial dataset and OSM, and the relative route network comparisons by town or municipality area, the following statement can be made (cf. Figure 6.11). Overall there exists an 85% correlation in total length between the OSM and the TomTom dataset for a total of 87% of the area of Germany.

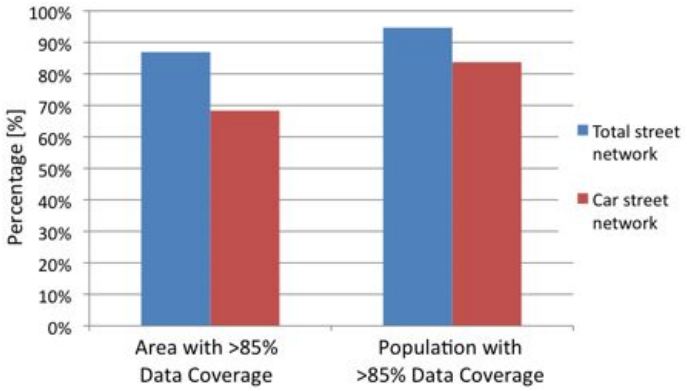


Figure 6.11.: Correlation between OSM data coverage and area, and OSM data coverage and population (June 2011).

For data related to car navigation, this value decreases to approximately 69%. Considering the population density, this means that nearly 95% of the inhabitants of Ger-

many are covered by 85% data coverage. In the case of car navigation data, this value decreases again to nearly 84% of the population.

Although OSM's total route length already exceeds that of TomTom, there are still areas in Germany within which TomTom has more data present than does OSM. According to the previous results gathered with regard to population density, these are typically areas in which the population tends to be low. The following two maps show where the differences in the total route network (Figure 6.12(left)) and the route network for car navigation (Figure 6.12(right)) can be found, based on the administrative areas for municipalities and towns.

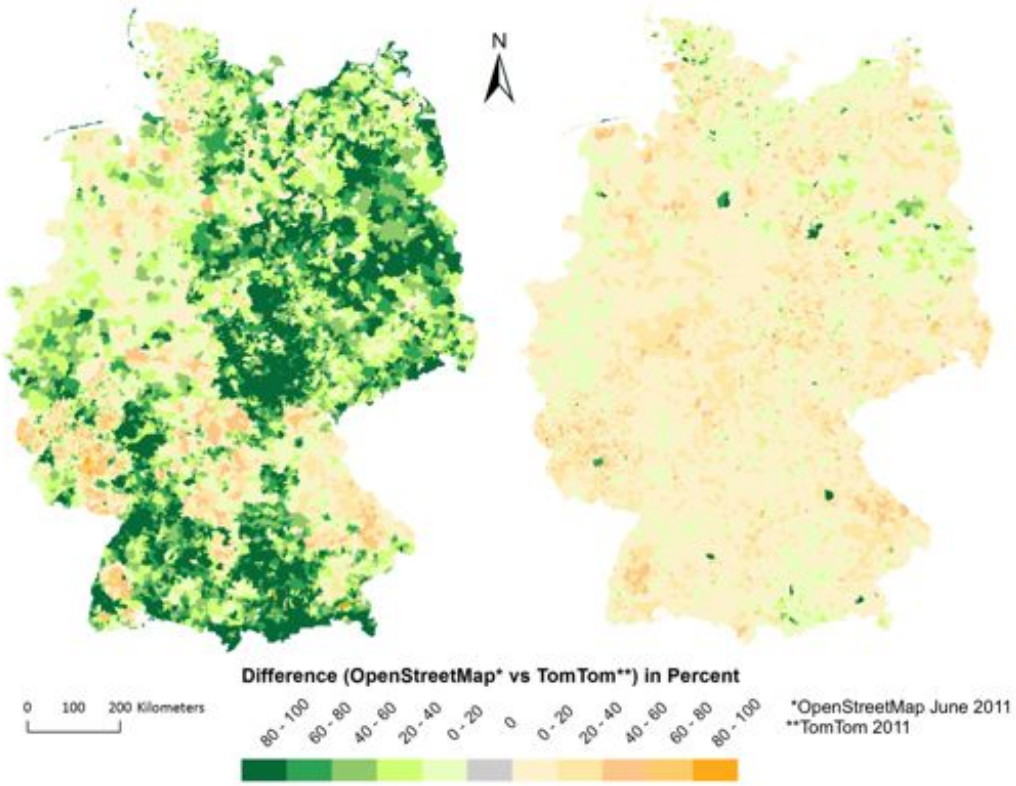


Figure 6.12.: Relative difference between TomTom and OSM for total route network (left) and for car navigation network (right) (June 2011).

The results gathered from several analyses over time showed that data collections in municipalities in the southeast of Germany show a good total route network; however, the same areas still lack data specific to car navigation. Upon closer examination, routes within these areas showed that although they were geometrically present in the dataset, attributes associated with these routes would not give a definitive street

category. This error occurs often when streets are digitized by a contributor from aerial images, but due to the lack of local knowledge about the area, no statement can be made on the category of the street. The second information that could be derived from the maps was that TomTom has less data available in the total route network for the eastern part of Germany, while OSM generally shows a higher total route network length in this area. Overall, with the exception of a few areas, this statement can be made for large parts of all of Germany. However, with regard to the route network for car navigation, this situation is, as mentioned before, somewhat worse.

A cloud diagram allows us to visualize the towns and municipalities according to their population and relative differences between the TomTom and OSM datasets, in particular, the network for car navigation (Figure 6.13). The graph clearly shows a decrease in discrepancies between the datasets with growing population density. These discrepancies can be positive and negative for each dataset. Additionally we can see that data differences in the class of rural towns (10,000–50,000 inhabitants) can vary between 10% and 20%.

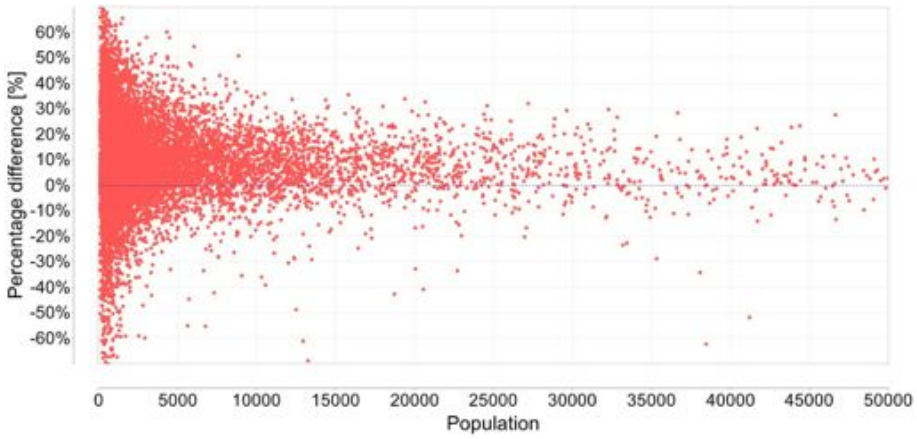


Figure 6.13.: Correlation between dataset differences and population density (June 2011).

Different numbers of members have been gathering data for OSM in each administrative area that we analyzed. A simplified number of participants per square kilometer can be calculated by dividing the total number of participants per administrative area by the size of the area. Our results showed that with an increasing number of participants, the relative difference between the datasets decreased (Figure 6.14). However, what is more important, a statement can be made on how many participants are required to gather all data to receive a sophisticated dataset.

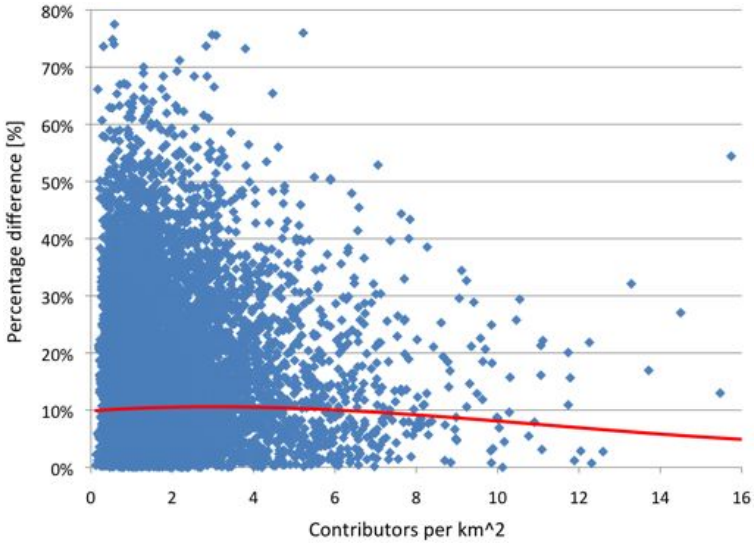


Figure 6.14.: Correlation between data completeness and number of contributors (June 2011).

Bearing in mind with the current data collection trend in Germany, completeness for car navigation data of more than 90% could already be achieved in relative comparison to the commercial dataset with an average of two project participants per square kilometer. According to the trend line, more than six participants are required to achieve a dataset that is close to "complete".

6.3.3. Topology errors and turn restrictions

A graph is generally required for a routing application that represents a street network and also comprises nodes and edges. Due to this fact, it is essential that the graph is topologically correct and that it does not contain any errors. OSM data is not routable in its standard form (Chen and Walter 2009, Schmitz et al. 2008); however, within the OSM project, attempts are being made to record the street data correctly topologically, but this topology cannot be used directly for routing without additional data preparation. During this preparation, procedure junctions must be localized by searching for nodes that are used by several streets, and streets must be attributed to these nodes accordingly. However, errors do occur in the OSM dataset. We have examined the entire route network for Germany to find possible topology errors. In doing so, we identified errors in the topology similar to those visualized in Figure 6.15. The first possible error that can occur is that the junction cannot be determined as

such, as the ways do not share a common node (1). Second, duplicate nodes or ways can cause an error (2), and third, the streets do not cross or lack information and they simply overlap (3).

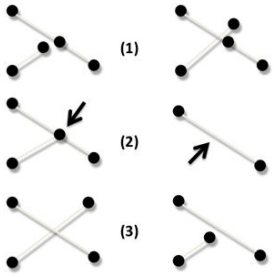


Figure 6.15.: OSM topology error types.

We converted the annual datasets (2007–2011) of OSM into routable street networks and searched for possible topology errors. The topology errors for non-linked streets were determined by measuring the distance between the two applicable streets, which should not be greater than 1 m. It can be clearly seen that the number of such errors has decreased over the years and remains high only for routes of cyclists or pedestrians (Figure 6.16).

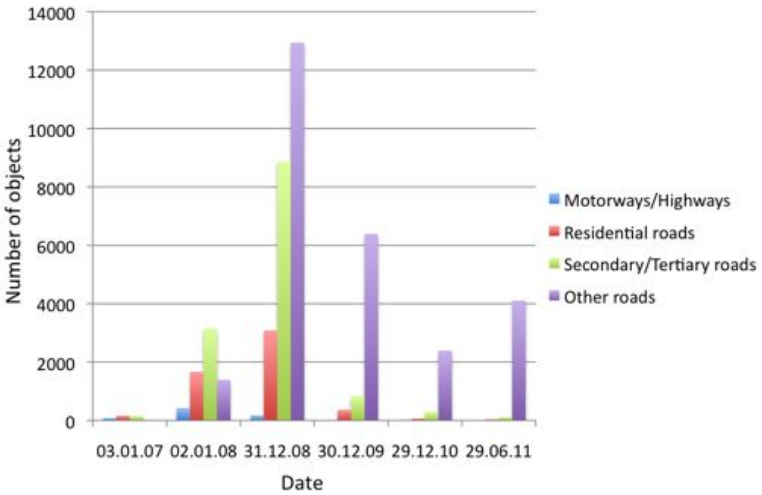


Figure 6.16.: OSM topology errors.

The results of the second analysis for possible double streets also showed that the quality has continually improved here, at least in the street network for car navigation (Figure 6.17). The number of errors for the third analysis, which shows the results

of the error for intersecting streets without any shared nodes (Figure 6.18), remains relatively constant, with the exception of the "other routes" data group. During random sampling, it happened that some of the errors that were identified were based on attribute errors in the dataset. For example, the information that the street is in fact a bridge was missing.

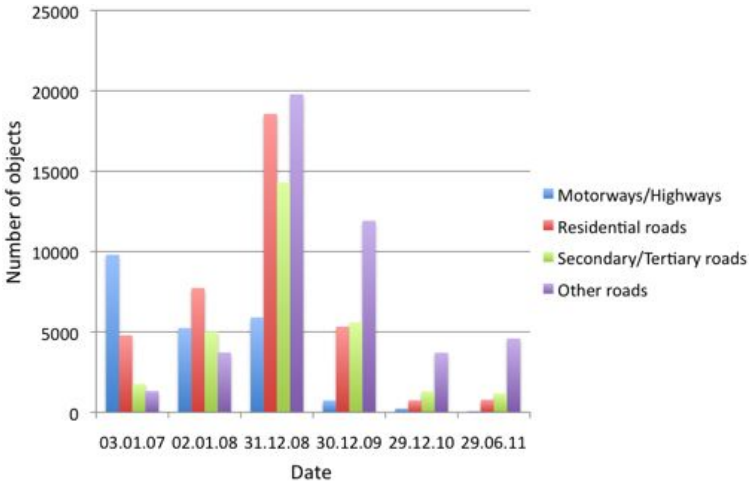


Figure 6.17.: OSM duplicate nodes or ways errors.

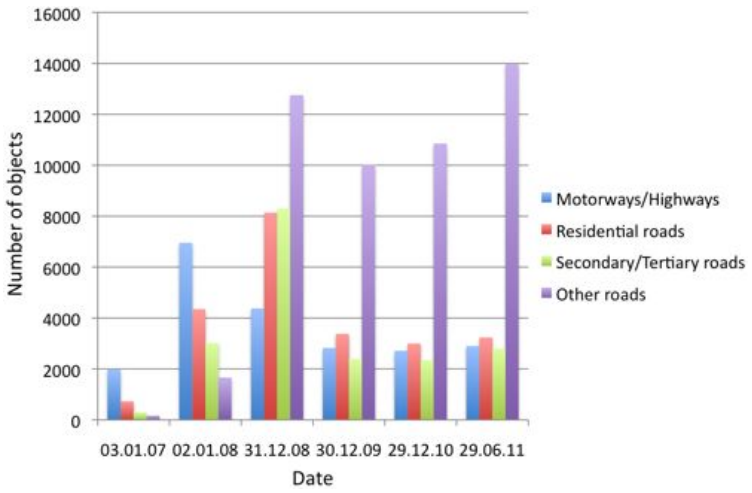


Figure 6.18.: Lack of information errors.

Turn restrictions constitute an essential component of routing applications. In a worst-case scenario, serious street accidents can occur should they be absent or incorrect. There are several different types of turn restrictions. In general, two types

can be differentiated: requirements and prohibitions. Requirements prescribe the only possible way(s) to turn or travel at a junction. Prohibitions, on the other hand, indicate where it is not permitted to travel. In the following preliminary comparison (Table 6.2), the total number of turn restrictions of TomTom and OSM for Germany are compared.

Table 6.2.: Total number of TomTom and OSM turn restrictions in Germany.

Data Provider	Date	Total	Standardized
TomTom	2011	Approximately 176,000	Approximately 174,000
OpenStreetMap	June 2011	Approximately 21,000	Approximately 28,000

The difference between TomTom and OSM totals almost 146,000. As such, TomTom currently has five times more turn restrictions available for Germany than does OSM. Although the number of turn restrictions available in the OSM dataset is continually increasing, it will probably take several more years before OSM achieves the same level as TomTom, based on the current status and development. The biggest issue during this analysis was to adjust TomTom‘s dataset, read the turn restrictions, and convert them in such a way that the OSM data would be applicable for a comparison. In addition to the distribution of information for turn restrictions over several attribute tables and datasets in TomTom, the existing restrictions also had to be filtered. For example "automatically calculated" turn restrictions or those prohibiting turning into a "residents only" street were among the restrictions that have been filtered out of the TomTom dataset. In addition to the total number of differences described above, a comparison by street category was also conducted (Figure 6.19).

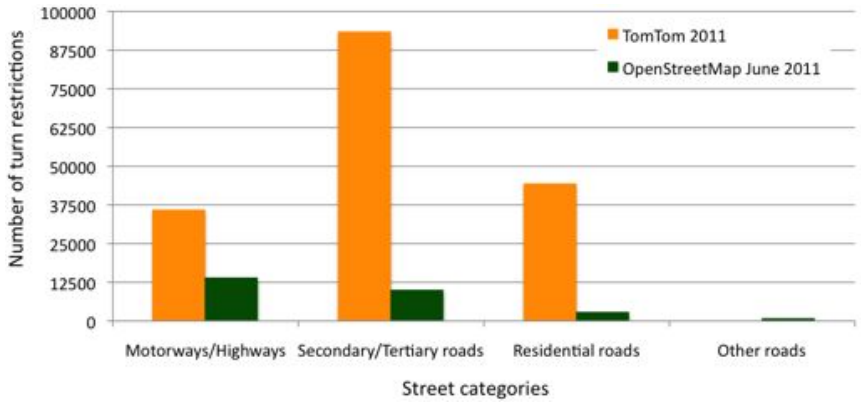


Figure 6.19.: Number of turn restrictions by street category in Germany for TomTom and OSM (June 2011).

For a further analysis, we organized the standardized turn restrictions according to their appearance in the different population groups (Figure 6.20). The results showed that a large number of missing objects fall into the rural groups. However, the graph also shows that objects are missing in urban areas as well.

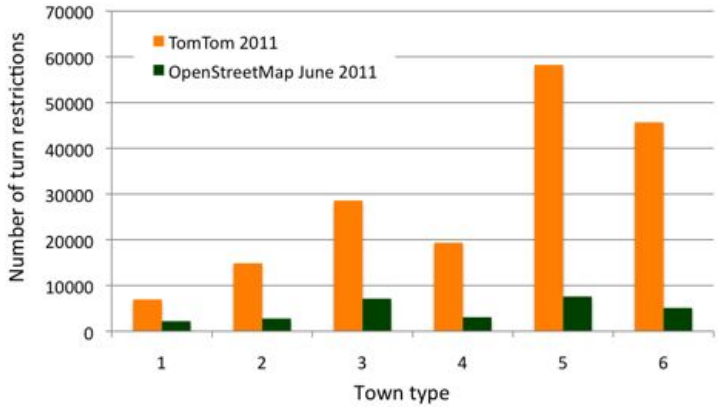


Figure 6.20.: Number of turn restrictions by town type in Germany for TomTom and OSM (June 2011).

A further important quality parameter, and the final aspect of our analysis, is the temporal accuracy of the geodata. The OSM dataset allowed us to analyze this accuracy factor by identifying the street time stamp of each object in the dataset. According to the information retrieved from the dataset, which included the time stamp of each route network object, approximately one third of the data originated during 2011 and 2010, and another third during 2009 and 2008 (cf. Figure 6.21).

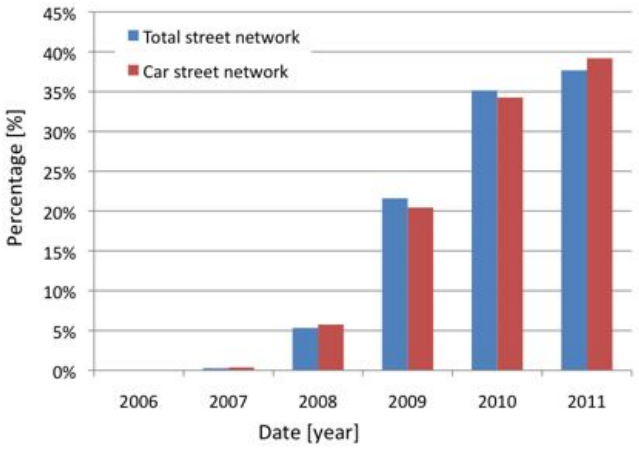


Figure 6.21.: Actuality of the OSM route network.

6.4. Conclusions and future work

In this article, we outlined the development of Volunteered Geographic Information in Germany from 2007 to 2011, using the OpenStreetMap project as an example. Specifically, we considered the expansion of the total street network and the route network for car navigation. With a relative completeness comparison between the OSM database and TomTom‘s commercial dataset, we proved that OSM provides 27% more data within Germany with regard to the total street network and route information for pedestrians. On the contrary, OSM is still missing about 9% of data related to car navigation. According to our projection for the future, this discrepancy should disappear by the middle or end of 2012, and the OSM dataset for Germany should then feature a comparative route network for cars as provided by TomTom.

In addition to the route network comparisons, we conducted further analyses regarding topology errors and the completeness of street name information. The results showed that the OSM dataset is not flawless; however, the trend shows that the relative and absolute number of errors is decreasing. Thus, it can also be discerned that not only is new data being added to the project database but also quality assurance is becoming a major factor within the OSM community. Our findings with regard to turn restrictions within the OSM database, which are of critical importance to navigation, showed that based on the current development rate and activity, it will take more than five years for OSM to catch up with the information found in the proprietary dataset used in our analysis. This slower development in comparison to the regular street data collection can have several reasons. It can be based on the fact that turn restrictions cannot be seen in the regular OSM map and thus are less appealing for contributors to be added. Some members might also not be familiar with the importance of turn restrictions for the dataset or do not understand how to implement them correctly.

Overall, a certain trend can be distinguished from our studies, as well as in all other studies conducted to date for the countries that were examined. Preliminary statements and conclusions in the past were that OSM data is sufficient for use with map applications. Today we can say that, at least in countries in which the OSM project is well developed, the data is becoming comparable in quality to other geodata from commercial providers regarding the different factors analyzed in this paper such as temporal accuracy and geometric accuracy.

However, several questions remain and further research is still needed. One important factor that has not been addressed yet is the importance of whether users who contribute data to OSM should also maintain it. Also, it is unclear whether missing

attribute information, such as street types or names, if added at a later date, could be analyzed and provided useful insights. So far it seems as if processing within the OSM project is closely related to visual factors, meaning that most data is collected in areas where there are white spots on the map, and thus no information is available. We will investigate specific questions regarding this user behavior in detail in the near future. It will be important to obtain further information on the project‘s participants and data contributors. These are some of the questions that need to be addressed: Are OSM mainly long-term contributors or are most of them so-called "submarine users"; that is, do they appear for a short period, add information, and then disappear again? Do members only add new data, or do they also edit existing information? Can an activity radius or area be determined for the participants of the project? Is the administrative area of an entire country completely covered by volunteers of the project or are data contributions by agencies playing a major role in certain areas?

It will continue to be important to carry out studies about the quality assurance of VGI. Preliminary suggestions have been made on how consistency of compiled VGI data could be achieved by improving quality during production and providing quality metadata for the users (Brando and Bucher 2010).

References

Anderson, P. (2007). What is Web 2.0? Ideas, Technologies and Implications for Education. In: *JISC*.

Ather, A. (2009). A Quality Analysis of OpenStreetMap Data. M.E. Thesis. University College London, London, UK.

Brando, C. and Bucher, B. (2010). Quality in User-Generated Spatial Content: A Matter of Specifications. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

Brassel, K., Bucher, F., Stephan, E., and Vckovski, A. (1995). Completeness. In: *Elements of Spatial Data Quality*. Oxford, UK: Elsevier81–108.

Chen, H. and Walter, V. (2009). Quality Inspection and Quality Improvement of Large Spatial Datasets. In: *Proceedings of the GSDI 11 World Conference: Spatial Data Infrastructure Convergence: Building SDI Bridges to Address Global Challenges*. (June 15–19, 2009). Rotterdam, The Netherlands.

Chilton, S. (2012). Crowdsourcing Is Radically Changing the Geodata Landscape: Case Study of OpenStreetMap. In: *Proceedings of the UK 24th International Cartography Conference*. (Nov. 15–21, 2009). Santiago, Chile.

Coast, S. (2007). *OpenStreetMap. Workshop on Volunteered Geographic Information.* URL: `http://www.ncgia.ucsb.edu/projects/vgi/` (visited on 11/11/2011).

Diaz, L., Granell, C., Gould, M., and Huerta, J. (2011). Managing User-generated Information in Geospatial Cyberinfrastructures. *Future Generation Computer Systems*, 27, 304–314.

Elwood, S. (2008). Volunteered Geographic Information: Future Research Directions Motivated by Critical, Participatory, and Feminist GIS. *GeoJournal*, 72, 173–183.

Fischer, F. (2008). Collaborative Mapping - How Wikinomics is Manifest in the Geoinformation Economy. *GEOInformatics*, 2, 28–31.

Flanagin, A. J. and Metzger, M. J. (2008). The Credibility of Volunteered Geographic Information. *GeoJournal*, 72, 137–148.

Fritz, S., McCallum, I., Schill, C., Perger, C., Grillmayer, R., Achard, F., Kraxner, F., and Obersteiner, M. (2009). Geo-Wiki.Org: The Use of Crowdsourcing to Improve Global Land Cover. *Remote Sensing*, 1, 345–354.

Girres, J.F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS*, 14, 435–459.

Goodchild, M.F. (2007a). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211–221.

Goodchild, M.F. (2007b). Citizens as Voluntary Sensors: Spatial Data Infrastructure in the World of Web 2.0. *International Journal of Spatial Data Infrastructures Research*, 2, 24–32.

Goodchild, M.F. (2008). Spatial Accuracy 2.0. In: *Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*. (June 25–27, 2008). Shanghai, China.

Goodchild, M.F. (2009). NeoGeography and the Nature of Geographic Expertise. *Journal of Location Based Services*, 3, 82–96.

Haklay, M. (2010). How good is OpenStreetMap Information: A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. *Environment and Planning B*, 37, 682–703.

Haklay, M. and Ellul, C. (2011). *Completeness in Volunteered Geographical Information - The Evolution of OpenStreetMap Coverage in England (2008–2009).* URL: `http://povesham.wordpress.com/2010/08/13/completeness-in-volunteered-geographical-information-%E2%80%93-the-evolution-of-openstreetmap-coverage-2008-2009/`.

Heipke, C. (2010). Crowdsourcing Geospatial Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65, 550–557.

Hudson-Smith, A., Batty, M., Crooks, A., and Milton, R. (2009a). Mapping for the Masses: Accessing Web 2.0 through Crowdsourcing. *Social Science Computer Review*, 27(4), 524–538.

Hudson-Smith, A., Batty, M., Milton, R., and Batty, M. (2009b). NeoGeography and Web 2.0: Concepts, Tools and Applications. *Journal of Location Based Services*, 3, 118–145.

Ludwig, I., Voss, A., and Krause-Traudes, M. (2010). Wie gut ist Open Street Map? Zur Methodik eines automatisierten objektbasierten Vergleiches der Strassennetze von OSM und NAVTEQ in Deutschland. *GIS Science*, 4, 148–158.

Maué, P. and Schade, S. (2008). Quality of Geographic Information Patchworks. In: *Proceedings of AGILE*. (May 5–8, 2008). Girona, Spain.

Mooney, P. and Corcoran, P. (2010). Using OSM for LBS - An Analysis of Changes to Attributes of Spatial Objects. In: *Advances in Location-Based Services*. Vol. 34. Berlin/Heidelberg, Germany: Springer496–507.

Neis, P. and Zipf, A. (2008). OpenRouteService.org is Three Times "Open": Combining OpenSource, OpenLS and OpenStreetMaps. In: *Proceedings of the GIS Research UK 16th Annual conference GISRUK 2008*. (Apr. 2–4, 2008). Manchester, UK.

Nelson, A., Sherbinin, A. de, and Pozzi, F. (2006). Towards Development of a High Quality Public Domain Global Roads Database. *Data Science Journal*, 5, 223–265.

Nielsen, J. (2006). *Participation Inequality: Encouraging More Users to Contribute*. URL: `http://www.nngroup.com/articles/participation-inequality/` (visited on 09/23/2013).

O'Reilly, T. (2005). *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. URL: `http://oreilly.com/web2/archive/what-is-web-20.html` (visited on 09/21/2013).

Over, M., Schilling, A., Neubauer, S., and Zipf, A. (2010). Generating Web-based 3D City Models from OpenStreetMap: The Current Situation in Germany. *Computers, Environment and Urban Systems*, 34, 496–507.

Ramm, F. (2009a). *Crowdsourcing Geodata. Society of Cartographers Summer School, Southampton*. URL: `http://www.geofabrik.de/media/2009-09-08-crowdsourcing-geodata.pdf` (visited on 12/28/2011).

Ramm, F. (2009b). *Krautsourcing 2.0 Beta - The State of Germany. State of the Map, Amsterdam*. URL: `http://www.geofabrik.de/media/2009-07-11-sotm-state-of-germany.pdf` (visited on 11/28/2011).

Ramm, F. and Stark, H.J. (2008). Crowdsourcing Geodata. *Géomatique Suisse*, 6, 315–318.

Rana, S. and Joliveau, T. (2009). NeoGeography: An Extension of Mainstream Geography for Everyone Made by Everyone? *Journal of Location Based Services*, 3, 75–81.

Schmitz, S., Neis, P., and Zipf, A. (2008). New Applications Based on Collaborative Geodata - the Case of Routing. In: *Proceedings of XXVIII INCA International Congress on Collaborative Mapping and Space Technology*. (Nov. 4–6, 2008). Gandhinagar, Gujarat, India.

Sui, D. (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.

Turner, J.A. (2006). Introduction to Neogeography. In: Cambridge, MA, USA: O'Reilly Media.

Van Oort, P.A.J. (2006). Spatial Data Quality: From Description to Application. Ph.D. Thesis. Wageningen University.

Walsh, J. (2008). The Beginning and End of Neogeography. *GEOconnexion International Magazine*, 7, 28–30.

Zielstra, D. and Hochmair, H.H. (2011a). A Comparative Study of Pedestrian Accessibility to Transit Stations using Free and Proprietary Network Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2217, 145–152.

Zielstra, D. and Hochmair, H.H. (2011b). Digital Street Data: Free versus Proprietary. *GIM International*, 25, 29–33.

Zielstra, D. and Hochmair, H.H. (2012). Comparison of Shortest Path Lengths for Pedestrian Routing in Street Networks Using Free and Proprietary Data. In: *Proceedings of the Transportation Research Board - 91st Annual Meeting*. (Jan. 22–26, 2012). Washington, DC, USA.

Zielstra, D. and Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

7. Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions

Authors

Pascal Neis, Dennis Zielstra, and Alexander Zipf

Journal

Future Internet

Status

Published: 3 June 2013 / Accepted: 16 May 2013 / Revised: 14 May 2013 / Submitted: 8 April 2013

Reference

Neis, P., Zielstra, D., and Zipf, A. (2013). Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Future Internet*, 5(2):282-300.

Contribution statement

Pascal Neis conducted all analyses for this study and wrote the majority of the manuscript. Both co-authors supported this publication through continuing discussions about the methods and the results of the study. Furthermore, extensive proof-reading

by both co-authors led to substantial improvements to the manuscript.

.....	.....
Prof. Dr. Alexander Zipf	Dennis Zielstra

Abstract

Volunteered Geographic Information (VGI) projects and their crowdsourced data have been the focus of a number of scientific analyses and investigations in recent years. Oftentimes the results show that the collaboratively collected geodata of one of the most popular VGI projects, *OpenStreetMap* (OSM), provides good coverage in urban areas when considering particular completeness factors. However, results can potentially vary significantly for different world regions. In this article, we conduct an analysis to determine similarities and differences in data contributions and community development in OSM between 12 selected urban areas of the world. Our findings showed significantly different results in data collection efforts and local OSM community sizes. European cities provide quantitatively larger amounts of geodata and number of contributors in OSM, resulting in a better representation of the real world in the dataset. Although the number of volunteers does not necessarily correlate with the general population density of the urban areas, similarities could be detected while comparing the percentage of different contributor groups and the number of changes they made to the OSM project. Further analyses show that socio-economic factors, such as income, can have an impact on the number of active contributors and the data provided in the analyzed areas. Furthermore, the results showed significant data contributions by members whose main territory of interest lies more than one thousand kilometers from the tested areas.

Keywords: Volunteered Geographic Information; OpenStreetMap; urban areas; collaborative mapping; comparison.

7.1. Introduction

In the past few years, collaborative mapping projects, with the main goal of collecting and distributing freely available geodata, have attracted significant attention from academia, leading towards their integration into research projects in a semantic and meaningful way. There are a variety of projects available on the Internet on which mostly volunteers share their expertise and information.

One of the most well-known examples of a *User-Generated Content* (UGC; Anderson 2007) online portal is the free online-encyclopedia Wikipedia. Other projects focus on the collection of a diverse type of data and information specifically containing geographic objects and their corresponding information. Due to the voluntary approach to data collection efforts for this particular type of information, it was initially termed

Volunteered Geographic Information (VGI; Goodchild 2007). In 2007, early questions were raised about "the phenomenon of VGI, and the use of VGI in doing science" (Kuhn 2007), especially in the area of *Geographic Information Science* (GIScience). Thus, UGC or VGI have developed into popular interdisciplinary research topics in recent years. Most VGI analyses focus on the *OpenStreetMap* (OSM) project, which provides a plethora of research questions due to its data diversity. An increasing number of studies in the past investigated data quality indicators, the motivation and activity spectrum of the community that share the information or applications that were developed based on the collected information found in OSM. Although different factors have been analyzed in these prior articles which will be discussed in more detail in the second section of this paper, the analyses usually show a similar pattern: The larger the population in the predefined, analyzed area, the stronger the data quality of the collaboratively collected information in OSM gets. To the best of the authors' knowledge, only minor studies on other countries not located in Europe have been carried out and reported. More importantly, no comparative investigation on OSM data for different selected world regions has been conducted.

The second major factor for collaborative projects such as OSM, after a well designed project infrastructure, is the worldwide community. The volunteers build the foundation of the project and guarantee the detailed data contributions and temporal accuracy of the data, thus it is important to investigate the different worldwide collaboration efforts. The main objective of this paper is to determine similarities and differences in the pattern of VGI data contributions and user activity spectrums for different worldwide urban areas. We hypothesize that different factors such as contributor concentration, population density and socio-economic parameters such as income can influence the data contributions to OSM. The analysis is conducted for 12 world regions representing at least one urban area for each continent.

The remainder of the paper is organized as follows: The following section gives a brief overview of the OSM project and prior OSM research. The next section introduces the study areas and applied data preparation steps. The third section presents the results conducted for the selected urban areas between 2005 and 2012. The last two chapters summarize and discuss the achieved outcomes and provide an outlook on potential future research.

7.2. Volunteered geographic information: The OpenStreetMap project

The OSM project is one of the most popular and well-known VGI platforms on the Internet. The main goal of the project, since its initiation in 2004, is to create a freely available database of geographic features (OpenStreetMap 2013). Contributions and edits to OSM can be made by any internet user that is registered to the project. This open approach to data contribution allowed the project to gradually attract new members and grow rapidly in the past few years to more than one million registered members at the time of writing (OpenStreetMap 2012d). Aside from the aforementioned registration to the project to be able to make edits to the map, the volunteers are commonly equipped with GPS enabled devices that allow the collection of new information in the field. Others prefer to trace new data from aerial imagery from their home computer. The imagery information is provided by a variety of sources and allows the active users to trace data even for areas that are not close to the editor's location. In addition to digitized features, attribute information about the created objects can be simultaneously added to the database. For a certain number of countries, contributions to the project were achieved through large data imports from commercial or governmental data sources whose licenses are compatible to the OSM license. Some examples can be found in the United States, the Netherlands and Austria, where partial or complete data representations of the road networks were enabled through this approach. For France, building information and the CORINE (Coordination of Information on the Environment), land cover information were imported to OSM in 2009.

There are no strict limitations, rules or standards to the type of information which should be contributed to the OSM project. Merely a de facto standard, represented by the "Map-Features", helps to guide the contributors with their work (OpenStreetMap 2012c). This guide describes the most common elements and objects and their corresponding attributes that can be found in the OSM project. The map features are mostly attributed with a key and value combination also referred to as "tags". The collected real world information in the database is represented by three data types. Nodes, which represent any point feature, Ways which represent lines such as roads and areas such as buildings, and Relations which contain information about how Nodes and Ways are related to each other.

In recent years, the OSM project, data, and contributors have been the center of attention for many research disciplines. In 2009, early outcomes showed a density

of OSM data in Germany with potential applicability for 3D location based services (Schilling et al. 2009). The results also showed some first indications of a correlation between improved data quality in areas with a higher population density. Similar results were conducted for London and entire England in 2008–2009, highlighting the decrease in data quality when moving away from bigger cities and that: “more affluent areas and urban locations are better covered than deprived or rural locations” (Haklay 2010, Haklay and Ellul 2011). In 2010, findings for Germany resembled the pattern of urban locations found in England: “If coverage is needed only in the densely populated urban areas of Germany, OpenStreetMap may already be an interesting and very cost-efficient-alternative” (Zielstra and Zipf 2010).

A similar statement was made for England: "Most accurate tiles are located in major urban areas such as London, Liverpool, Manchester or Birmingham" (Haklay et al. 2010). A different analysis for Germany stated that: "At the national level, the quality of OSM is highest regarding relative object completeness" and "quality differs locally, and even in a single town the different aspects of quality may vary" (Ludwig et al. 2011). A study conducted for France showed a heterogeneous OSM data pattern, which "is particularly explained by the coexistence of different data sources, processes of capture, and contributors' profiles, highlighting the importance of following accepted and well-defined specifications" (Girres and Touya 2010). Additionally, the analysis revealed that the more volunteers contributed within an area, the more recent the objects were, i.e. providing a better temporal quality of OSM itself. Similar to all prior findings, an analysis for Ireland showed that the data completeness in OSM loosely correlates with the population density (Ciepluch et al. 2010).

However, contradicting results to the pattern found in Europe could be determined for the US. In this particular case urban areas only showed similar data completeness between OSM and commercial providers in Florida, while rural areas were more complete in OSM (Zielstra and Hochmair 2011). This difference was primarily based on the TIGER/Line data import in OSM for the US and not due to active data contributions (Zielstra and Hochmair 2011).

The heterogeneous pattern of the OSM project is not limited to data completeness and accuracy factors. The community and its active contributors show a similar distribution. At the beginning of 2012, about 75% of all members who made at least one change to the database were located in Europe, while the rest was distributed over the world (Neis and Zipf 2012). Especially some countries with higher population values such as USA, China and India show relatively small OSM project communities. Although the project was initiated in the UK, the most active community of the project

in recent years can be found in Germany (Neis and Zipf 2012). The latest results showed that about 25% of all active OSM members are located in Germany. Thus, it is not surprising that the road network completeness shows good results, sometimes exceeding commercial providers for this particular area (Neis et al. 2012b). Solely attribute information such as road names, speed limits and turn restrictions are missing for parts of the German dataset (Ludwig et al. 2011, Neis et al. 2012b).

Aside from road network parameters, which are the main focus of most conducted analyses, a few other publications also confirm OSM data to be suitable for 3D city models in urban areas (Song and Sun 2010, Goetz and Zipf 2012). In summary, almost all prior studies show that urban areas provide better data completeness in OSM than rural areas (Hagenauer and Helbich 2012, Koukoletsos et al. 2012), which is sometimes also referred to as "urban bias in VGI" (Mooney et al. 2013). However, each individual case study needs to be analyzed for its particular purpose (Mooney et al. 2013, Mondzech and Sester 2011). Chances are that: "When one moves away from large urban centers the major issue for quality becomes one of coverage - in many rural areas there is little or no OSM coverage at all" (Mooney and Corcoran 2012).

While most studies analyze the quantity and quality of the collaboratively collected information in OSM, others focus on the motivational factors of the volunteers that contribute to VGI projects (Budhathoki et al. 2008, Coleman et al. 2009, Lin 2011). Possible motivational factors might be the unique ethos, or that geospatial information should be freely available to everyone. For others, learning new technologies, self-expression, relaxation and recreation or just pure fun can play a major role (Budhathoki 2010). Three independent surveys in 2009 (Budhathoki 2010), 2010 (Stark 2010) and 2011 (Lechner 2011) gave more insight on demographic aspects of the OSM project. The majority of the contributors to the project, about 97%, were males. Two out of the three surveys (Budhathoki 2010, Lechner 2011) showed that on average 65% of the respondents were between 20 and 40, and about 23% between 40 and 50 years old. Furthermore, about 56% had a high-school or higher education degree. About 50% of the respondents in one survey considered their current profession as computer science related (Lechner 2011) and another survey showed that about 50% had some sort of GIS background (Budhathoki 2010), highlighting that "the OSM community does not constitute with GIS amateurs as is speculated in VGI" (Budhathoki 2010).

Community-based projects, websites and portals such as OSM are oftentimes affected by so-called "participation inequality". A 90-9-1 rule can usually be applied to most of these projects (Nielsen 2006). This rule highlights that about 90% of the members of community-based projects are usually only consuming the collaboratively collected

information, while 9% occasionally contributes to the project and only 1% demonstrate a very active pattern. This rule can be applied to projects such as Wikipedia (Anthony et al. 2007, Javanmardi et al. 2009) and has also been tested for the OSM project. In 2007, about 28% of the 120,000 members of the project actively contributed any data (Budhathoki 2010).

In 2011, about 38% of the 500,000 members made at least one change to the dataset (Neis and Zipf 2012). Additionally, only 3% of all members actively contributed to the project each month. However, considering values from prior years the recent number of active contributors is not increasing in the same pattern as the total number of registered members. At the end of 2012, of the almost 1 million registered OSM members on average only 18,000 members, less than 2% actively contributed to the project each month (OpenStreetMap 2012b).

The increasing popularity of OSM also comes with caveats such as cases of vandalism, similar to developments seen in Wikipedia. An analysis carried out in 2012 revealed that for a timeframe of one week at least one case of vandalism could be detected in the OSM database each day (Neis et al. 2012a). It needs to be noted though that these cases of vandalism can also be accidently created by new or inexperienced members and are not always intentional.

7.3. Selected urban areas and data sources

Several definitions from different sources help to distinguish urban or agglomerated areas from rural areas, which is a crucial point for the analysis presented in this article. Demographia defines urban areas as: "A continuously built up land mass of urban development that is within a labor market (i.e., metropolitan area or metropolitan region), without regard for administrative boundaries (i.e., municipality, city or commune)" (Demographia 2012). The identification of these areas is usually based on maps and satellite images that estimate the continuous urbanized area (Demographia 2012). It is also important to distinguish between urban areas and metropolitan areas in which: "A metropolitan area is an urban area plus the satellite cities around the urban area and the agricultural land in between." Since these factors could potentially influence the results of the analysis conducted, it was decided to use urban areas instead of metropolitan areas to avoid forests, agricultural and other uninhabited areas in the selected regions.

A variety of online sources allow Internet users to retrieve freely available urban area information. However, oftentimes sources show inconsistencies in their provided

information due to different geographical definitions of urban areas (Forstall et al. 2009). Since none of the available sources, such as Natural Earth Data or CORINE, provided the information needed for a comprehensive comparison of worldwide urban areas, it was decided to trace the urban area boundaries based on Bing satellite imagery. The center of each polygon was primarily based on the location of the city name feature in the standard OSM map. The urban area sizes that were implemented during the polygon generation and their corresponding population information were retrieved from Demographia (2012). Figure 7.1 shows a world map highlighting the selected regions for which urban area polygons were generated.

Figure 7.1.: Overview of the selected urban areas.

During the urban area selection for the analysis it was decided to choose at least one large, well-known urban area (city) for each continent to provide world wide information. In total, 12 urban areas and their related area extent, absolute and population density information were chosen as shown in Table 7.1.

After the areas of interest were defined, generated and included all desired information for analysis, an OSM history dump file was retrieved from the OSM project (OpenStreetMap 2012a). This particular file includes the entire history (versions) of all geodata that is included in the worldwide OSM database until October 19, 2012. Doing so enabled us to analyze the potential development of the datasets for each urban area for the past few years by clipping the information from the worldwide dataset and applying Java based tools that were specifically developed for this research project.

Table 7.1.: Selected urban areas. Source: Demographia (2012).

Country	City	Population in 2011	Area (km^2)	Density (/km^2)
Germany	Berlin	3,453,000	984	3,509
Argentina	Buenos Aires	13,639,000	2,642	5,162
Egypt	Cairo	14,718,000	1,658	8,877
Turkey	Istanbul	13,576,000	1,399	9,704
South Africa	Johannesburg	7,618,000	2,525	3,017
United Kingdom	London	8,586,000	1,623	5,290
United States	Los Angeles	14,900,000	6,299	2,365
Russia	Moscow	15,512,000	4,403	3,523
Japan	Osaka-Kobe-Kyoto	17,011,000	3,212	5,296
France	Paris	10,755,000	2,845	3,780
South Korea	Seoul-Incheon	22,547,000	2,163	10,424
Australia	Sydney	3,785,000	1,788	2,117

7.4. Results

A number of different analyses were conducted to provide detailed information on the development of OSM data, number of contributors and member activities in the selected urban areas in relation to population and other socio-economic factors such as income. The main goal was to identify significant differences or similarities between the selected urban areas to approve or reject findings from prior research, which solely focused on European cities and selected areas in the US.

7.4.1. Contributor numbers and activity spectrums

One of the most important factors of projects such as OSM that rely on volunteered data contributions is the activity of the community in the project. The active members do not only contribute new data but also keep existing data up to date or improve it over time. It was shown that an increasing number of contributors within an area also improve the positional accuracy of the geodata, one of many geodata quality assessment criteria (Haklay et al. 2010). Figure 7.2 shows the development of the OSM community for each urban area from January 2007 to September 2012. The absolute number of OSM members has been normalized by the population density in each urban area to reduce the impact of the size of the city area on the results. The urban area names appear next to Figure 7.2 in descending order based on the values retrieved from the datasets. The diagram shows that Berlin, Paris, Moscow and London have higher values in comparison to other urban areas that were tested when considering

the relation between the number of OSM members and the total population density. Generally, there are three groups that can be distinguished. The four aforementioned cities fall into a group with the highest values, while Los Angeles and Sydney create the second group with average values.

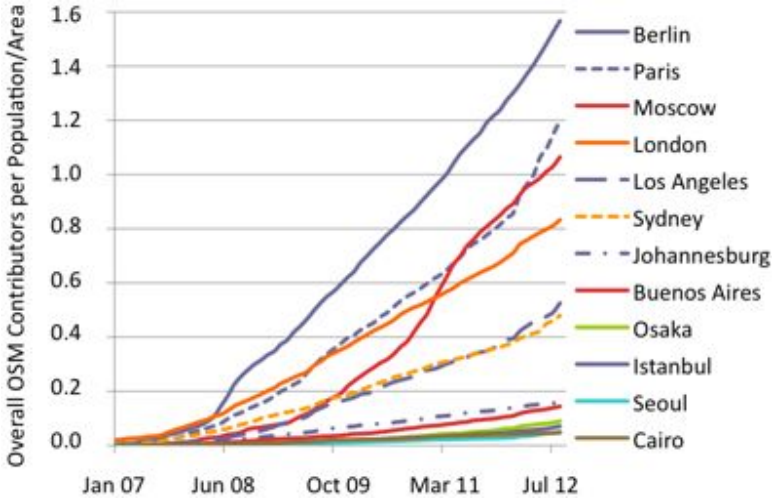


Figure 7.2.: Number of OpenStreetMap (OSM) contributors per Population/Area-ratio (Jan. 2007–Sept. 2012).

The third and last group inherits all other urban areas i.e., Johannesburg, Buenos Aires, Osaka, Istanbul, Seoul und Cairo with values smaller than 0.20. The statistical analysis showed, however, that there is no correlation between the number of contributors and the population density in the tested areas (Spearman's rho R_S value of -0.140). Figure 7.2 also shows the significantly different increase in member numbers for the different areas within the past four years.

The urban areas showing the smallest values for current active contributors to the project also reveal the least impressive increase over time. Although the total number of OSM members in an area might give some prior impression about the potential data contributions that could occur in an area, it does not take the actual activity spectrum of the members into consideration. A smaller group of very active data contributors could achieve similar results in data collaboration efforts as a large group of mappers with very limited contributions. Thus, the following analysis divided the registered members for each urban area into different mapper groups as introduced in a prior publication(Neis and Zipf 2012). Different thresholds, based on the number of Nodes an OSM member created, helped to distinguish the different mapper groups.

If a mapper created less than ten nodes she/he falls into the "Nonrecurring Mapper" group, less than 1000 Nodes identify "Junior Mappers" and more than 1000 Nodes identify mappers that are part of the "Senior Mapper" group. Figure 7.3(left) shows how many members have actively contributed to the project by making at least one edit (creation, modification or deletion) to an object, while Figure 7.3(right) shows the distribution of the members based on the aforementioned classification schema.

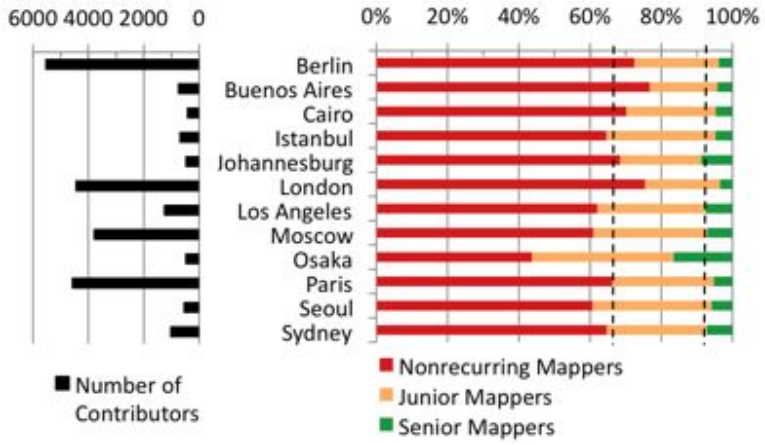


Figure 7.3.: (left) Number of contributors; and (right) Distribution of mapper groups per urban area (Sept. 2012).

The results presented in Figure 7.3(left), representing the absolute numbers of OSM contributors, show a similar pattern as the relative values shown in Figure 7.2, with larger values for all European cities and Moscow. Seoul, Istanbul, and Cairo, the three urban areas with the highest population density values of all tested areas, do not show similarly high values for the number of contributors, indicating that population density is not a major factor for data contributor numbers in OSM. The vertical dashed lines in Figure 7.3(right) represent the average values for each corresponding mapper group for all areas. This additional information helps to prove that almost all urban areas, with the exception of Osaka, show similar distributions for the individual mapper groups. Only 6.6% (with a standard deviation of 3.5%) of all contributors in an urban area added a large amount of information with more than 1000 Nodes ("Senior Mapper"), while 65.5% of the mappers fall into the "Nonrecurring Mapper" category (standard deviation of 8.7%).The rest belongs to the "Junior Mapper" group with an average of 27.9% and a standard deviation of 5.9%. Due to the small number of members in Osaka and possibly data imports, the results gathered for this particular case do not match the general pattern of all other tested urban areas and can be considered as an

outlier.

Next to the general contributions that a member makes to the project, it is also important to investigate the active time frame of a member. Prior research has shown that the contributors that fall into the different mapper categories also collect information for different time frames (Neis and Zipf 2012). While for the Wikipedia project prior research revealed that active contributors usually edit at least one article per month, in this analysis the time frame was expanded to three months to retrieve more meaningful results about the number of active contributors in each tested area. Figure 7.4(left) gives an overview of the number of members that have been active between August and October 2012 by creating at least one Node in the selected urban areas.

Figure 7.4(right) shows the percentage of the total contributors in each area for the designated timeframe, again divided into the different mapper groups. On average about 16% of the total number of members that created at least one Node in the tested urban areas, have been active between August and October 2012. Figure 7.4(right) also shows that the amount of contributors that are part of the "Senior Mapper Group" is very low with an average value of less than 3%.

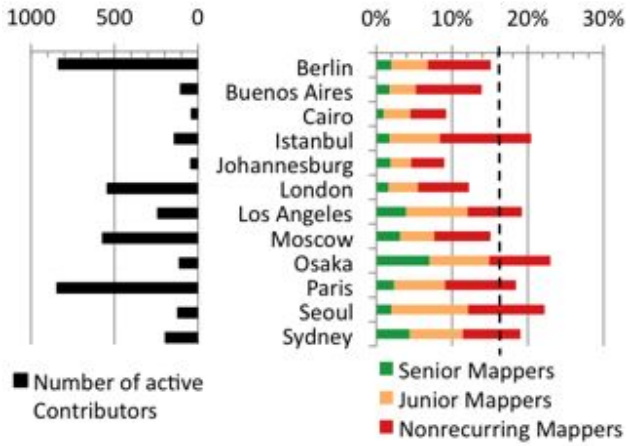


Figure 7.4.: (left) Number of active contributors; and (right) Percentage of mapper group contributions per urban area (Aug.–Oct. 2012).

7.4.2. Dataset quantity

A first impression about the quality of the OSM dataset can be gathered by investigating the quantity of the collected information in the designated areas. Figure 7.5 shows the total number of Nodes, Ways and Relations collected in each urban area per km^2. The results show that Paris has the highest contributed object density, which

is partially based on a large data import of cadastral building information and not necessarily only based on active contributions. However, for Osaka different data imports were applied to the OSM dataset as well but did not show the same results as in Paris when considering the object density. Especially the concentration of Relation information, which is added via a more complex process and is usually conducted by experienced contributors, separates the more advanced cities such as Berlin, London, Moscow and Paris from less complete cities such as Cairo, Istanbul and Seoul.

Figure 7.5 also supports prior findings about the strong concentration of OSM data in European urban areas in comparison to other continents, in which Istanbul is the only exception for a European urban area with lower data density and Moscow an exception for a non-European urban area with strong data collection efforts. Overall the number of collected OSM objects in the tested areas correlates with the number of contributors per area. For Nodes and Ways the results showed a R_S value of 0.6783 and for Relations one of 0.720.

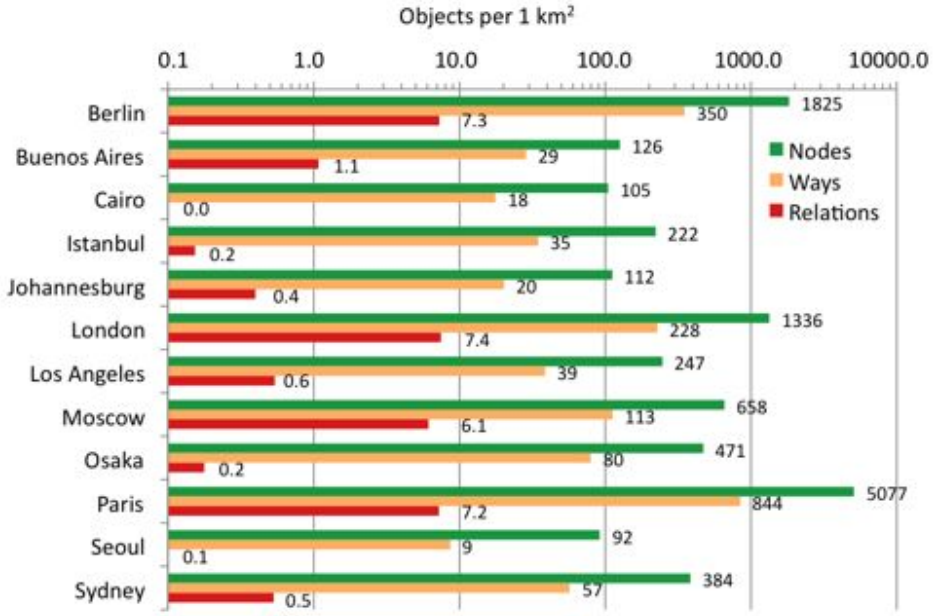


Figure 7.5.: Density of nodes, ways & relations per km^2 (Oct. 2012).

7.4.3. Temporal dataset quality

The timeliness of the collected information plays a major role in the quality assessment of a geodataset such as OSM. Each object in the OSM database is related to a unique timestamp, which represents the time at which the object was edited the last time.

Additionally each edited object has a version-number indicating how many times the object has been changed since its first creation. By utilizing this information and combining it with the OSM history dump file, it is possible to determine when an object has been created and when and how often it has been edited. Figure 7.6 shows the collected temporal accuracy information for all 12 analyzed areas based on the OSM history dump file dated 19 October 2012.

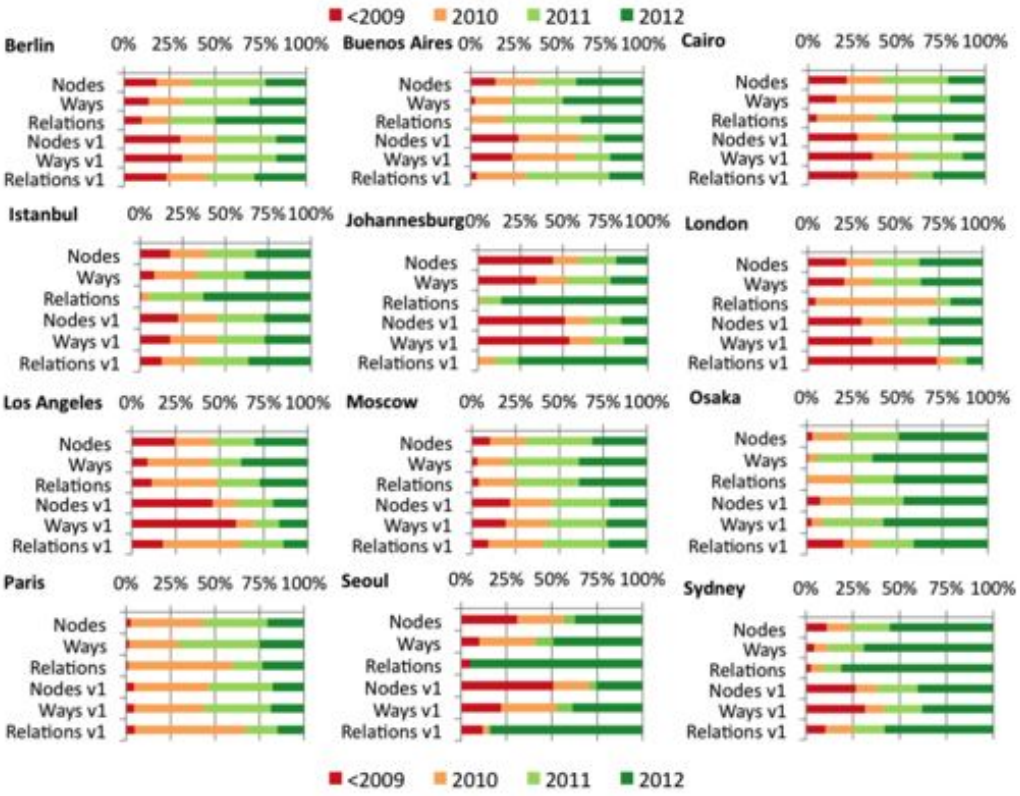


Figure 7.6.: Distribution of currency and data versions per urban area (Oct. 2012).

The first three Nodes, Ways, and Relations bars of each diagram represent the distribution of the currency of the objects based on the timestamp information of the entire dataset. The larger the dark green area of the bar, the more up-to-date is the particular dataset. The first three bars are followed by a second set of Nodes v1, Ways v1 and Relations v1 bars for each tested area. The information provided in these bars shows when an object, i.e., the first version of the object, was created. Therefore, it is possible to determine for how long and how much data has been contributed to each area over the past few years. The comparison of the three upper and three lower bars

provide detailed information if a dataset has been updated by the OSM community. If the bars show a similar pattern to each other e.g. Nodes and Nodes v1, it would indicate that the data has not been updated since its creation.

The comparison of the diagrams for Berlin shows that the currency information is similar to the objects labeled as first version. Moscow, Paris and Buenos Aires show similar patterns. The aforementioned data imports for France are also represented in the diagrams for Paris, with a strong concentration of data contributions in 2010. For London, the results revealed that the majority of Relations were created before or in 2009, yet the timestamp of most Relations is dated within 2010. The strong increase in Relations in the dataset in 2009 can again be attributed to a data import. However, the diagram also shows that after the import the community has been updating the information in 2010, thus most Relations show a 2010 timestamp. Sydney proves to be one of the most up-to-date datasets in OSM. One possible reason for this pattern could be the OSM license change in 2012. Due to the license change, all data that was contributed by members that explicitly did not agree with the new license was deleted from the database, including all data that was imported from sources whose license was no longer compatible with the new OSM license. After the deletion of the data it was partially recollected by the local members that agreed to the new license. Cairo, Johannesburg, Seoul and Los Angeles show less up-to-date datasets in comparison to the other analyzed areas. The temporal OSM dataset quality is better in areas with a stronger community activity. The correlation between the number of contributors in an area and the number of newly created objects in 2012 proof to be high for Nodes (R_S 0.755), Ways (R_S 0.720) and Relations (R_S 0.832). With regards to the contributions by the individual mapper groups the numbers showed that on average 94.6% of all Nodes, 92.8% of all Ways and 84.8% of all Relations were collected by Senior Mappers.

7.4.4. Local and external mappers

Although OSM is based on collaborative data collection efforts by volunteers, companies such as Yahoo Imagery (until 2011) and Microsoft Bing[1] have been supporting the project by providing their satellite imagery to the project's members to trace information directly from the images. This also allowed members to collect information for areas that they maybe never physically visited or where no local knowledge is available. The OSM project does not provide any direct information about the home location of the members. However, prior research has introduced different methods on how to

[1] http://opengeodata.org/microsoft- imagery-details

determine an activity spectrum or area of a member (Neis and Zipf 2012). One of the introduced processes has been applied to determine the location of a member for the selected urban areas. The process is collecting information from all changesets provided by OSM for each member. Changesets are rectangle shaped polygons that surround the area in which a designated member has been making changes to the dataset. By utilizing the center points of all changesets that were created for a single member it is possible to create a final polygon which represents the main activity area of a mapper (Neis and Zipf 2012). Based on the newly created polygons a distance between the polygon and the designated urban area can be measured. The measured distance gives clues whether the data contributor is a local or external mapper. For classification purposes the differences were divided into three groups representing a distance of less than 100 km, more than 100 km and less than 1000 km and more than 1000 km. A distance of less than 100 km to the corresponding urban area would indicate a local mapper and a distance larger than 1000 km would represent an external mapper. Figure 7.7 shows the distribution of local and external mappers for all tested areas. The distribution of the contributors represented in the figure is based on the Senior Mapper group. Similar results were retrieved when conducting an analysis with the Junior and Nonrecurring Mapper groups.

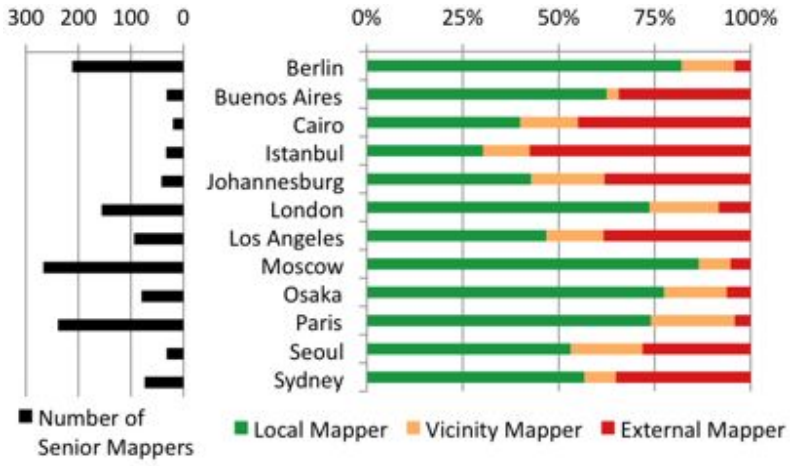


Figure 7.7.: (left) Number of senior mappers per urban area; and (right) Distribution of senior local or external mappers per urban area (Oct. 2012).

Figure 7.7 clearly shows that not all urban areas merely rely on local mappers. Areas with higher member numbers generally show smaller contributions by external mappers. The statistical analysis, however, does not show any correlation between the

two variables for the selected areas (R_S -0.16). Cairo, Istanbul, Johannesburg and Los Angeles revealed some surprising results. Although Los Angeles has a higher number of members in OSM than Johannesburg, both urban areas show similar patterns when considering external mapper activities. The largest external mapper contributions were found for Cairo (almost 50%) and Istanbul (more than 50%), indicating that the main activity area of these mappers is more than 1000 km away from these particular urban areas. Reasons for these patterns could be the increased Internet accessibility in other countries or the popularity of these particular areas for tourism, which attracts more external mappers. However, this statement is only based on speculation.

7.4.5. Average contributions by active OSM members

The analyses presented thus far focused on the estimation of the absolute number of OSM members and the determination of active members in each urban area. Another important factor that needs to be considered is the quantity in which an active member contributes to the project. In the following analysis only active members that are part of the aforementioned Senior Mapper group and that did not participate in any data imports for the urban areas have been investigated. The main goal was to retrieve detailed information about the average number of active days a Senior Mapper spends on data contributions to the project and how many objects on average were created in this timeframe in each urban area. The analysis was conducted for the three months prior to the creation date of the history dump file, October 19, 2012. Figure 7.8 shows the average number of active days of all Senior Mappers and the number of Nodes, Ways and Relations created in this time frame for all tested areas.

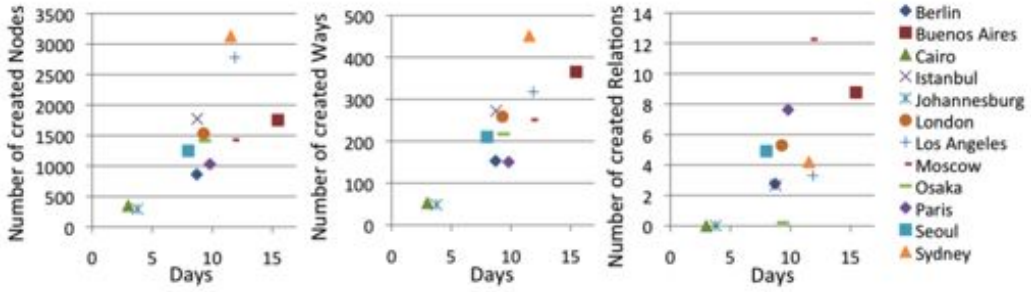


Figure 7.8.: Average senior mapper activity timeframe and contributions per urban area (Aug.–Oct. 2012).

Although the absolute number of members for each tested urban area varies, the

distribution of the different mapper groups proved to be alike. A similar result was revealed for the average contribution values and active days of the Senior Mappers. The total average values of all urban areas combined showed that a Senior Mapper is active for about 9 to 10 days and creates around 1466 Nodes, 229 Ways and 4 Relations (Table 2).

Table 7.2.: Activity timeframe and contributions of a senior mapper (Aug.–Oct. 2012).

Parameter	Min	Max	Mean value	Standard deviation
Active days	3	15.5	9.3	3.4
Nodes	292.8	3133.0	1466.2	848.3
Ways	48.7	451.5	229.1	118.8
Relations	0.0	12.3	4.3	3.8

Additionally, Figure 7.8 shows that particularly areas with very small communities can generate positive and negative outliers in this analysis such as Cairo and Buenos Aires. Similar to the result gathered during the temporal dataset quality analysis, Sydney appears to take an outlier role due to the same cause i.e., OSM license change. The stronger Nodes and Ways contributions can be accredited to remapping efforts by the OSM community.

7.4.6. Impact of socio-economic factors

GPS-enabled devices, Smartphones and computers with internet access have become omnipresent in many countries worldwide. The existence of these devices in each country does not automatically imply however that all citizens have access to them or have the financial resources to purchase them. Figure 7.3 has shown that the OSM community of an urban area does not necessarily relate to the population density. Thus, one question that remains is if other socio-economic factors, such as income, have an impact on the development of an internet community for portals such as OSM. The *Gross National Income* (GNI) is defined by the World Bank (2012) as: "The value of all final goods and services produced in a country in one year (gross domestic product) plus income that residents have received from abroad, minus income claimed by nonresidents." The corresponding GNI per capita is defined as: "A country's *gross national product* (GNP) divided by its population." For the following analysis it is important to apply the GNP value, which represents each individual urban area and not the entire country. The values for 2012 utilized in this analysis are provided by The Brookings Institution (2013). Figure 7.9 show the results gathered for the tested

urban areas when comparing the OSM contributor density with the GNP per capita.

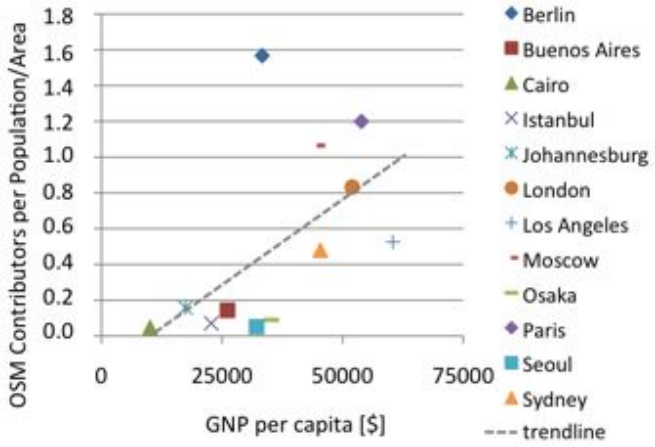


Figure 7.9.: Contributor density (Oct. 2012) and GNP per capita (2012).

The statistical analysis of the results shown in Figure 7.9 revealed a distinct correlation between the number of members in OSM and the GNP with an R_S value of 0.664. Furthermore, a number of outliers such as Berlin with a slightly lower GNP but larger OSM community could be determined. The opposite situation can be seen for Los Angeles and Sydney; both countries have a higher GNP in comparison to other tested areas but only show small OSM communities. Usually, it would be expected that these cities would show a higher concentration of OSM members based on their GNP values, if income is considered as an influential factor. However, the outliers identified in this analysis also showed that other factors next to population density or income must have an influence on the development of an OSM community in the selected areas.

7.5. Conclusions and future work

The analyses presented in this article provided detailed information about the concentration of OSM geodata and its contributors for 12 selected worldwide urban areas. The main objective of the article was to determine similarities or significant differences between the selected areas regarding their data growth and collection efforts by the OSM community. The results showed that the urban areas provide significantly different data concentrations in OSM, which can be caused by data imports for selected areas or large differences between community contributions. The results also highlighted the differences between European and other world regions in OSM. Especially the number of OSM members can differ largely in this case. With the exception of Istanbul, all

tested European areas show higher OSM member concentrations than other areas with high population density values such as Cairo or Seoul. Moscow proved to be a positive example outside of Europe with a large OSM community.

When splitting the OSM contributors into different groups, based on their number of edits made to the map data, all tested areas show similar patterns. About 7% of the data contributors are very active "Senior Mappers" while 28% fall into the Junior Mapper category with fewer contributions. The largest group of data collectors is represented by the "Nonrecurring Mappers" with 66%. The determination of the active time frames of the members showed that about 16% of all OSM contributors in each area have been active within three months by making at least one edit to the map.

However, only 3% of the members can be considered as very active "Senior Mappers". The data also revealed that the absolute number of active OSM members has no impact on the activity spectrum of the volunteers. The most active "Senior Mappers" created on average about 90% of the data in the urban areas and worked on about 9 of the tested 90 days and created almost 1500 Nodes, 230 Ways and 4 Relations total. The temporal data quality proved to be highly influenced by the size of the community in each urban area, which confirms similar findings for France (Girres and Touya 2010). Smaller communities do not guarantee continuing data collection or correction efforts and thus make the datasets outdated.

Further results were gathered by analyzing and comparing local to external data contributors. Especially urban areas with lower OSM community member numbers show large (sometimes more than 50%) external member data contributions. Especially Cairo, Istanbul, Johannesburg and Los Angeles rely on these non-local members. In general, this pattern contradicts in certain aspects the main idea behind VGI projects as defined by Goodchild (2009) in which "local volunteers" should be the main source of information. However, Neis and Zipf (2012) already proved that more than 50% of the worldwide "Senior Mappers" of the OSM project contribute data to two or more countries and do not limit their efforts to local areas. Due to the fact that the population density did not provide enough evidence of impacting OSM member numbers, other socio-economic factors were taken into consideration. It was hypothesized that income might be a major influential factor. The analysis showed that urban areas with higher income values such as Sydney, Los Angeles, Seoul and Osaka could potentially inherit larger OSM communities than currently available but still show a correlation between income and OSM contributor ratio. Berlin has a slightly lower average income in comparison to other tested areas and a relatively high member density, but can be considered as an exceptional case. Overall the conducted analyses do not completely

confirm prior results gathered for England where "more affluent areas and urban locations are better covered than deprived or rural locations" (Haklay and Ellul 2011). However, a more comprehensive investigation with additional urban areas, which increases the sample size, could improve the findings of our analysis and statistical results presented.

Questions remain about potential other reasons that would explain why urban areas such as Los Angeles or Seoul only show small OSM communities and not similar success as in Europe. Possibly differences in Internet access, culture, mentality, personal interests or acquaintance to the project due to language barriers could play a role. Others would argue that countries with freely available datasets, e.g., provided by the government such as the TIGER/Line datasets in the US, are slowing down data contribution efforts in OSM. Other influential indicators could most likely only be determined by conducting an extensive survey.

The assessment of the quality of the data collected by external OSM members in comparison to local members was not part of this study. However, it was clearly shown that large data contributions have been made in selected areas by members that maybe never collected data locally in person and lack the "local expertise" (Goodchild 2009) that are making VGI projects unique. Based on these findings, investigations planned for the future will reveal some answers to questions such as: Do external or remote members provide a better, equal or worse data quality when contributing to the project? A similar approach to the one chosen during the analysis of Wikipedia and "The Roles of Local and Global Contribution Inequality" (Arazy and Nov 2010) could provide some meaningful insights. Geometric differences such as inconsistencies in positional accuracy will most likely be limited due to the high resolution images that the mappers can utilize when tracing data for OSM, as long as they are not outdated. However, a metadata analysis including street names, street types or turn restrictions could introduce some of the caveats of remote data contributions in OSM.

Acknowledgments

The authors would like to thank Jamal Jokar Arsanjani for his valuable comments towards the improvement of this paper.

References

Anderson, P. (2007). What is Web 2.0? Ideas, Technologies and Implications for Education. In: *JISC*.

Anthony, D., Smith, S. W., and Williamson, T. (2007). The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia. In: *Dartmouth Computer Science Technical Report TR2007-606*. Hanover, NH: Dartmouth College.

Arazy, O. and Nov, O. (2010). Determinants of Wikipedia Quality: The Roles of Global and Local Contribution Inequality. In: *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work (CSCW)*. (Feb. 6–10, 2010). Savannah, GA, USA.

Budhathoki, N. (2010). Participants' Motivations to Contribute to Geographic Information in an Online Community. Ph.D. Dissertation. University of Illinois, Urbana-Champaign, Urbana, IL, USA.

Budhathoki, N., Bruce, B., and Nedovic-Budic, Z. (2008). Reconceptualizing the Role of the User of Spatial Data Infrastructure. *GeoJournal*, 72, 149–160.

Ciepluch, B., Jacob, R., Mooney, P., and Winstanley, A. (2010). Comparison of the Accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In: *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences*. (July 20–23, 2010). Leicester, UK.

Coleman, D., Georgiadou, Y., and Labonte, Y. (2009). Volunteered Geographic Information: The Nature and Motivation of Producers. *International Journal of Spatial Data Infrastructures Research*, 4, 332–358.

Demographia (2012). *World Urban Areas. 8th Annual Edition*. URL: `http://www.demographia.com/db-worldua.pdf` (visited on 12/28/2012).

Forstall, R.L., Greene, R.P., and Pick, J.B. (2009). Which are the Largest? Why Published Lists of Major Urban Areas vary so Greatly. *Tijdschrift voor economische en sociale geografie*, 100, 277–297.

Girres, J.F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS*, 14, 435–459.

Goetz, M. and Zipf, A. (2012). Towards Defining a Framework for the Automatic Derivation of 3D CityGML Models from Volunteered Geographic Information. *International Journal of 3-D Information Modeling*, 1, 1–16.

Goodchild, M.F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211–221.

Goodchild, M.F. (2009). NeoGeography and the Nature of Geographic Expertise. *Journal of Location Based Services*, 3, 82–96.

Hagenauer, J. and Helbich, M. (2012). Mining Urban Land Use Patterns from Volunteered Geographic Information by Means of Genetic Algorithms and Artificial Neural Networks. *International Journal of Geographical Information Science*, 26, 963–982.

Haklay, M. (2010). How good is OpenStreetMap Information: A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. *Environment and Planning B*, 37, 682–703.

Haklay, M., Basiouka, S., Antoniou, V., and Ather, A. (2010). How many Volunteers does it take to Map an Area well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47, 315–322.

Haklay, M. and Ellul, C. (2011). *Completeness in Volunteered Geographical Information - The Evolution of OpenStreetMap Coverage in England (2008–2009).* URL: `http://povesham.wordpress.com/2010/08/13/completeness-in-volunteered-geographical-information-%E2%80%93-the-evolution-of-openstreetmap-coverage-2008-2009/`.

Javanmardi, S., Ganjisaffar, Y., Lopes, C., and Baldi, P. (2009). User Contribution and Trust in Wikipedia. In: *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing*. (Nov. 11–14, 2009). Washington, DC, USA.

Koukoletsos, T., Haklay, M., and Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transaction in GIS*, 16, 477–498.

Kuhn, W. (2007). National Center for Geographic Information and Analysis and Vespucci Specialist Meeting on Volunteered Geographic Information. In: *Volunteered Geographic Information and Giscience*. Santa Barbara, CA, USA.

Lechner, M. (2011). Nutzungspotentiale crowdsource-erhobener Geodaten auf verschiedenen Skalen. Ph.D. Dissertation. University Freiburg, Freiburg, Germany.

Lin, Y. (2011). A Qualitative Enquiry into OpenStreetMap Making. *New Review of Hypermedia and Multimedia*, 17, 53–71.

Ludwig, I., Voss, A., and Krause-Traudes, M. (2011). A Comparison of the Street Networks of Navteq and OSM in Germany. *Advancing Geoinformation Science for a Changing World*, 1, 65–84.

Mondzech, J. and Sester, M. (2011). Quality Analysis of OpenStreetMap Data based on Application need. *Cartographica*, 46, 115–125.

Mooney, P. and Corcoran, P. (2012). The Annotation Process in OpenStreetMap. *Transaction in GIS*, 16, 561– 579.

Mooney, P., Corcoran, P., and Ciepluch, B. (2013). The Potential for using Volunteered Geographic Information in Pervasive Health Computing Applications. *Journal of Ambient Intelligence and Humanized Computing*, 4(6), 731–745.

Neis, P., Goetz, M., and Zipf, A. (2012a). Towards Automatic Vandalism Detection in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 315–332.

Neis, P., Zielstra, D., and Zipf, A. (2012b). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4, 1–21.

Neis, P. and Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project - The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 146–165.

Nielsen, J. (2006). *Participation Inequality: Encouraging More Users to Contribute.* URL: `http://www.nngroup.com/articles/participation-inequality/` (visited on 09/23/2013).

OpenStreetMap (2012a). *Planet : Complete OSM Data History.* URL: `http://planet.osm.org/planet/full-history/` (visited on 12/28/2012).

OpenStreetMap (2012b). *Statistics of the Free Wiki World Map.* URL: `http://osmstats.altogetherlost.com` (visited on 12/28/2012).

OpenStreetMap (2012c). *Wiki : Map Features-Summary of Commonly Used Tags for Main Elements Used to Describe Features within OSM.* URL: `http://wiki.osm.org/wiki/Map_Features` (visited on 12/28/2012).

OpenStreetMap (2012d). *Wiki : Stats.* URL: `http://wiki.osm.org/wiki/Stats` (visited on 12/28/2012).

OpenStreetMap (2013). *The Free Wiki World Map.* URL: `http://www.osm.org` (visited on 05/25/2013).

Schilling, A., Over, M., Neubauer, S., Neis, P., Walenciak, G., and Zipf, A. (2009). Interoperable Location Based Services for 3D Cities on the Web Using User Generated Content from OpenStreetMap. In: *Proceedings of the 27th Urban Data Management Symposium.* (June 24–26, 2009). Ljubljana, Slovenia.

Song, W. and Sun, G. (2010). The Role of Mobile Volunteered Geographic Information in Urban Management. In: *Proceedings of 18th International Conference on Geoinformatics.* (June 18–20, 2010). Beijing, China.

Stark, H.J. (2010). Umfrage zur Motivation von Freiwilligen im Engagement in Open Geo-Data Projekten. In: *Proceedings of FOSSGIS Anwenderkonferenz für Freie und Open Source Software für Geoinformationssysteme.* (Mar. 2–5, 2010). Osnabrück, Germany.

The Brookings Institution (2013). *Global MetroMonitor.* URL: `http://www.brookings.edu/research/interactives/global-metro-monitor-3` (visited on 01/10/2013).

World Bank (2012). *Beyond Economic Growth, Glossary.* URL: `http://www.worldbank.org/depweb/english/beyond/global/glossary.html#30` (visited on 12/27/2012).

Zielstra, D. and Hochmair, H.H. (2011). A Comparative Study of Pedestrian Accessibility to Transit Stations using Free and Proprietary Network Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2217, 145–152.

Zielstra, D. and Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

8. Towards Automatic Vandalism Detection in OpenStreetMap

Authors

Pascal Neis, Marcus Goetz, and Alexander Zipf

Journal

ISPRS International Journal of Geo-Information

Status

Published: 22 November 2012 / Accepted: 16 November 2012 / Revised: 5 November 2012 / Submitted: 8 October 2012

Reference

Neis, P., Goetz, M., and Zipf, A. (2012): Towards Automatic Vandalism Detection in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1(3):315-332.

Contribution statement

Pascal Neis conducted all analyses for this study and wrote the majority of the manuscript. Both co-authors supported this publication through continuing discussions about the methods and the results of the study. Furthermore, extensive proof-reading by both co-authors led to substantial improvements to the manuscript.

.....	.....
Prof. Dr. Alexander Zipf	Marcus Goetz

Abstract

The *OpenStreetMap* (OSM) project, a well-known source of freely available worldwide geodata collected by volunteers, has experienced a consistent increase in popularity in recent years. One of the main caveats that is closely related to this popularity increase is different types of vandalism that occur in the projects database. Since the applicability and reliability of crowd-sourced geodata, as well as the success of the whole community, are heavily affected by such cases of vandalism, it is essential to counteract those occurrences. The question, however, is: How can the OSM project protect itself against data vandalism? To be able to give a sophisticated answer to this question, different cases of vandalism in the OSM project have been analyzed in detail. Furthermore, the current OSM database and its contributions have been investigated by applying a variety of tests based on other Web 2.0 vandalism detection tools. The results gathered from these prior steps were used to develop a rule-based system for the automated detection of vandalism in OSM. The developed prototype provides useful information about the vandalism types and their impact on the OSM project data.

Keywords: investigation; vandalism; detection; OpenStreetMap; Volunteered Geographic Information (VGI).

8.1. Introduction

User generated content (UGC; Diaz et al. 2011) is a well-known phenomenon of the Web 2.0 movement. With the ubiquitous availability of GPS-enabled devices, such as smartphones, cameras or tablets, users not only collected UGC, but also started to geo-reference their collected information. This new trend became popular under the term "*Volunteered Geographic Information*" (VGI; Goodchild 2007), or crowd-sourced geodata (Heipke 2010). Amateurs and professionals collaboratively collect, share and enhance geodata for specific VGI platforms. Essentially, everybody is able and allowed to use available VGI for their own applications and services at no charge.

One of the most popular and most manifold projects for VGI is the *OpenStreetMap* (OSM) project (Neis et al. 2012, Goetz and Zipf 2012, Mooney et al. 2011). Initially aiming at the creation of a global web map, the project soon turned far beyond that. Nowadays, OSM can be considered as a global geodata database that everybody can access, edit and use. A rapidly growing community and the emerging demand for open geodata made OSM one of the most used non-proprietary online maps and a source of geodata and information for many third-party applications, such as route planning

& geocoding (Neis and Zipf 2008), 3D (Over et al. 2010) or indoor (Goetz and Zipf 2012) applications. Although users of OSM have to register prior to contributing, the OSM model has been designed under an open-access approach. However, although this open approach can be considered as the key to OSM's success, it can also be a source of a variety of problems. While most contributions are legitimate, some attacks by lobbyists or spammers result in vandalism. One of the most popular examples for vandalism in OSM is the case of two employees of a popular search engine deleting OSM features in the London area (OpenGeoData 2012). However, there has not yet been a general investigation on the impact and quantity of vandalism in OSM, and no distinct task force against vandalism has been implemented so far in OSM (cf. Wikipedia's "Subtle Vandalism Taskforce" - Wikipedia 2012b). The OSM "Data Working Group" can be contacted when "serious Disputes and Vandalism" (OpenStreetMap 2012b) are encountered. However, "Minor incidents of vandalism should be dealt with by the local community" (OpenStreetMap 2012b).

Nevertheless, following the idea of collective knowledge, it is assumed that vandalism is detected and corrected by other OSM contributors within an (unknown) period of time. In the case of OSM, vandalism can occur intentional and unintentional, contradicting the traditional definition of the term "vandalism". However, this strongly depends on the change itself (vandalizing a geometry is probably more obvious than vandalizing semantic information), as well as on the area in which the edit has been performed, e.g., vandalism in a metropolitan area will probably be detected and corrected pretty fast, whereas vandalism in a very rural area will potentially remain for a very long period of time.

In general, data validation and vandalism detection needs to be distinguished from each other. While data validation incorporates different methodologies for quality assurance, vandalism focuses on active data corruptions. In this paper, we will focus on the latter one. Although automatic methods for the detection of vandalism in OSM ought to be of interest for both the contributors as well as the consumers of the project, only a few tools or other significant developments have been accomplished in this field in recent years. At the time of writing, only two tools are available that can observe a predefined area of interest. The tools are "OpenWatchList" (OpenStreetMap 2012k), which was developed under open source standards, and "OSM Mapper" developed by ITO (2012). Both tools provide the possibility to register to an RSS feed, which provides information about the latest changes in a distinct area. The functionality of the tools can be compared to so-called watchlists in Wikipedia, which track edits that were made to selected articles (Van den Berg et al. 2011). There are many other

similarities between OSM and Wikis, which nowadays are widespread on the Internet, such as the reliance on "radical trust" (Caminha and Furtado 2012) and citizens that become the "sources of data" (Van den Berg et al. 2011). What all these tools have in common is that they inform registered users about every single edit in a predefined area, rather than evaluating the corresponding edit and highlighting potential cases of vandalism. That is, the user still has to (manually) investigate every single edit, which is a very time-consuming and tedious effort. Nevertheless, vandalism detection in OSM plays an important role and will gain more importance because of the increasing popularity of the project, which depends highly on its data quality (Neis et al. 2012, Brando and Bucher 2010).

Regarding Wikipedia, there are a couple of approaches available that focus on vandalism (cf. Chin et al. 2010, Potthast 2010, Adler et al. 2011, Mola-Velasco 2011). As mentioned by Van den Berg et al. (2011), it could be useful to apply similar approaches, methodologies and technologies, which have already been utilized in other open source projects and Web 2.0 encyclopedias, to detect vandalism in OSM and/or revert unconstructive changes.

Therefore, the main contribution of this paper is an investigation of vandalism in OSM, as well as the development of a rule-based system for the automated detection of vandalism in OSM. A comprehensive set of rules has been defined by investigating past vandalism incidents, the current OSM database and its contributions, as well as related Wikipedia vandalism detection tools. Our system incorporates the user's individual reputation, as well as the performed action. Both parameters are evaluated independent from each other, which allows an individual definition of weights for both parameters after the corresponding detection. Essentially, this enables dedicated OSM members (patrols) to individually define appropriate weights, as individual patrols probably judge the importance of a user's reputation differently. Obviously, the change itself is a very important component, which needs to be investigated. Therefore, the change is tested against well-known community criteria, such as provided attributes, or community accepted map features and more. Incorporating the user's reputation is another important factor, as investigations on vandalism for other Web 2.0 projects (i.e., Wikipedia) revealed that in most cases, vandalism is performed by new (probably inexperienced) community members instead of more experienced community members (West et al. 2010). The developed system architecture has been prototypically implemented and added to a minutely updated OSM database. In this way, it was possible to apply some initial early results to refine the defined base rules and further improve the automatic vandalism detection.

The remainder of this article is structured as follows: The following section gives a general introduction to the OSM project, while the third section of the paper describes the different types of vandalism that can occur in OSM. Sections four and five describe our developed rules-based system and the results that were gathered after applying our methodology. The last section discusses our findings and presents some suggestions for future work and research.

8.2. OpenStreetMap

Founded in 2004 at the University College London, the OSM project's goal is to create a free database with geographic information for the entire world. A large variety of different types of spatial data and features, such as roads, buildings, land use areas or *Points-of-Interest* (POI), are collected in the database. Following the Web 2.0 approach of a collaborative creation of massive data, any user can start contributing to the project after a short online registration. Essentially, every registered user is able to add new elements, as well as alter or delete existing ones. This simple approach - similar to Wikipedia - led to more than 700,000 registered members by August 2012 (OpenStreetMap 2012e). In total, almost 200,000 members made at least one edit, and roughly 3% of all members made at least one change per month to the database by the end of 2011 (Neis and Zipf 2012), i.e., those that can be considered as regular contributors.

Several different research studies about the quality and completeness of OSM data in comparison to other data sources (e.g., governmental or commercial) have been published in recent years (e.g., Neis et al. 2012, Zielstra and Hochmair 2011, Ludwig et al. 2011, Girres and Touya 2010, Haklay 2010). In Europe, OSM shows an adequate level of data coverage for urban areas, which allowed the development and distribution of map services or other applications, such as *Location Based Serviced* (LBS). However, less populated areas do not show the same completeness level in OSM, which makes the dataset unreliable in those areas. Thus, OSM data can be very use-case dependent, and the requirements must be carefully considered (Mooney et al. 2013).

As stated above, every community member can alter the current OSM database; however anonymous changes are not supported (Neis et al. 2012). This is a crucial difference between Wikipedia and OSM and brings at least a small advantage: "OSM users are identified by their usernames, which can be blocked. In Wikipedia, users are identified by username or IP-address and more than one user might use the same IP-address" (Van den Berg et al. 2011). However, after uploading the data to the

community, the change is instantly live, i.e., applied to the productive system. Essentially, there is no quality or vandalism control prior to the publication, in contrast to, for example, Google MapMaker (Google 2012), where experienced users review new submitted content. General thoughts on how an OSM user should react in the case he or she detects vandalized data in the project can be found on the vandalism webpage in the OSM Wiki (OpenStreetMap 2012i).

Registered users can contribute data to OSM in different ways. The classic approach is to collect data with a GPS receiver, which afterwards can be edited with one of the various freely available editors, such as Potlatch or JOSM. Since November 2010, users are explicitly allowed to trace data from Bing aerial imagery and add the data to OSM (Bing 2010). This allows a user to collect information without physically being at a distinct location (thus a user from Germany can also provide data about a city in France). Regardless of how the data is collected, users can provide additional information, such as street names or building types, about the different OSM features. In the eight years since its initiation, 1.5 billion geo-referenced points (nodes in OSM terminology), 144 million ways (both linestrings and simple polygons) and 1.5 million relations (for describing relationships, such as turn restrictions or complex polygons with holes) (OSMstats 2012) have been collected as of today (August 2012).

Each of the objects contains a version number, a unique ID, the name of the last editor and the date of the last modification (i.e., the date of creation for new objects). Furthermore, so-called tag-value pairs containing additional (user-provided) information are attached to each feature. Any modification made by a user to a feature in OSM is stored in a so-called changeset, containing information about the change itself, as well as the editor and the time of the edit.

If a user wishes to implement the data of the OSM project for an application (for example a map for public transportation), a planet file, which contains all information of the latest database of the project, can be downloaded (OpenStreetMap 2012h). However, as people are contributing data every minute, a dataset will become outdated after a short period of time (probably even after one minute). To avoid the deployment of a full OSM database every time an update is required (which is hardly feasible for a minutely or hourly updated database), the OSM project provides so-called *OSM Change-Files* (OSC), also referred to as "Diff", which can be downloaded. These files only contain the latest changes to the database and are available for different time frames, such as every minute, hour or day. The format of the OSC-files will be explained in more detail in a later section of this paper. Figure 8.1 shows a simplified version of the OSM project infrastructure.

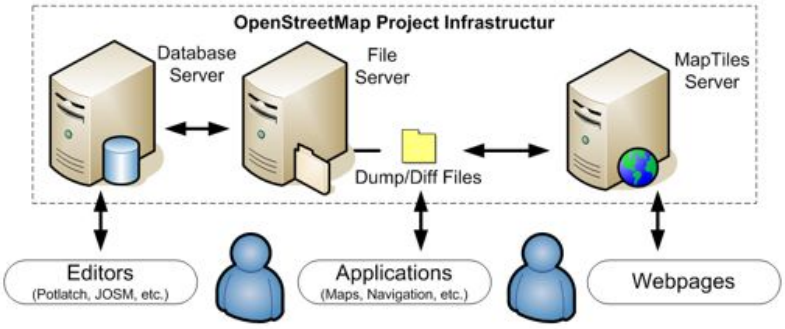


Figure 8.1.: The OpenStreetMap infrastructure/geostack (simplified).

By using one of the freely available OSM editors, the contributors can edit any object of the project's database. External applications, such as routing or mapping applications, can use the project's data by retrieving the dump- and diff-files from the database.

On average, nearly 700 new members have registered to the project each day between January and March 2012. According to Neis and Zipf (2012), nearly 30% of those newly registered contributors will become active contributors (and not only a registered user). That is, each day in 2012, 230 new OSM members started contributing to OSM. Table 8.1 contains the average number of edits per OSM object (node, way and relation) per day between January and June 2012.

Table 8.1.: Number of daily edited OSM objects (January–June 2012).

Number of ...	Node	Way	Relation
Daily created objects	1,200,000	130,000	1,500
Daily modified objects	170,000	70,000	3,500
Daily deleted objects	195,000	16,000	280
Users who daily edited	2,000	1,940	560

Considering these numbers for estimating future processing workloads for our suggested tool, it can be determined that (on average) every minute, 830 node creations, 190 node modifications and 135 node deletions will have to be performed. Furthermore, 90 new, 48 modified and 11 deleted ways and one new, two modified and 0.2 deleted relations have to be processed. Those numbers can obviously vary during the day, but they give a first indication on how much data will be edited.

8.3. Types of vandalism

The open approach of data collection in the OSM project can cause a variety of types of vandalism. It is possible that a contributor purposely or accidently makes changes to the dataset that are harming the project's main goal. Common vandalism types that appear in the actual OSM geodata database are (based upon OpenStreetMap (2012i)):

- a new object with no commonly used attributes
- a non-regular geometrical modification of an object ("Graffiti")
- a non-common modification of the attributes of an object
- randomly deleting existing objects
- an overall abnormal behavior by a contributor
- generating fictional and non-existing objects
- inappropriate use of automated edits (bots) in the database
- application of mechanic edits (e.g., selecting 100 buildings and adding the key building: roof: shape = flat)

Other vandalism types can be found in the OSM project, but that are not solely limited to the actual geodata that is stored in the database are:

- Copyright infringements, e.g., tracing data from Google Maps
- Disputes on the project-wiki
- Disruptive behavior or spamming of a member (e.g., in his edits, in the forum, on the mailing lists or on her/his user page)

The following Figure 8.2 shows an example of "Graffiti" vandalism in 2011 in Zwijndrecht (The Netherlands). The users who caused this disorder of the features used the Potlatch OSM editor, which directly applies the changes to the live OSM database.

Potthast et al. (2008) and West et al. (2010) manually analyzed vandalism in Wikipedia to learn about the specific characteristics. For our analysis, we used a similar approach and manually analyzed 204 user blocks of the project to gather more detailed information about vandalism in OSM. Members of the OSM Data Working Group or moderators are allowed to block other OSM members for a short period of time (between 0 and 96 h) (OpenStreetMap 2012g). The blocked members need to login to the main OSM website and read their notification to be able to unblock their accounts.

Figure 8.2.: Example of "Graffiti" vandalism in OSM in Zwijndrecht (The Netherlands) (OpenStreetMap 2012j).

There are several reasons for such a block, such as importing data that infringes the Import-Guidelines (OpenStreetMap 2012c), creating mass-edits which do not follow the Code-of-Conduct (OpenStreetMap 2012a) or vandalizing the data in some sort of way. Out of these 204 users that were blocked between 7 October 2009 and 31 July 2012, we were able to determine 51 cases of data vandalism events, not counting members that were blocked multiple times. The geographical pattern showed that cases of vandalism can be found worldwide in the OSM database with a slight focus on larger cities. Table 8.2 summarizes our results of manually collected vandalism cases and their characteristics.

Table 8.2.: Characteristics of vandalism in OSM (October 2009–July 2012).

Feature	Value	Description
Fictional Data	33.3%	The user created some fictional data
Editing Data	33.3%	The user modified some existing data, e.g., did some non-regular geometrical modification
Deleting Data	43.1%	The user deleted some existing data
New User	76.4%	The data was vandalized by a new project member
Potlatch Editor	82.4%	OSM editor which was used during the vandalism

When a vandalism event on one of the articles in Wikipedia is detected, it is usually reverted within a matter of minutes (Kittur and Kraut 2008). In our OSM analysis, 63% of the vandalism events were reverted within 24 h and 76.5% within 48 h. Some outliers could be determined, which needed more than 5 d up to a maximum of 29 d.

As mentioned before, there are three types of objects used in the database: nodes, ways and relations. All three object types can be created, modified or deleted by each member of the project. Essentially, each member can also alter objects that have been created or altered previously by a different community member. Changes to the OSM database can be analyzed by investigating the "Diff"-files (also termed OSM-Change files). Every "Diff"-file, whether it contains information about each minute, hour or day, provides information about all changes that have been made to the database within this distinct time frame. The most recent edits, that is, those changes that happened after the creation of the previous diff file, are grouped according to their action ("create", "modify" or delete") and object type ("node", "way" or "relation"). However, the diff-files only contain the geo-reference of the nodes (i.e., longitude and latitude), but not the actual geometry of way- or relation-features. Additionally, if a feature has been modified, the file does not provide any specific information about the actual change. If, for example, a node has been moved, the change file will only contain the new location of the node and not both locations (or the difference vector). The action types "create" and "delete" are only valid for the creation or deletion of a geometric object (e.g., the creation of a node or the deletion of a complete relation). In contrast, whenever additional (semantic) information is added (i.e., create), altered (i.e., modify) or removed (i.e., delete), this change is represented as a modify change, rather than a create or delete action. For a more accurate analysis of the object (especially for a modify action), it is therefore often necessary to gather the history of the object. If the change constitutes the creation of an object, it is necessary to have a closer look at what type of object has been created and, for example, to check which attributes have been used for the initial creation. In the case of a modification of an existing object, additional aspects need to be considered, such as: Has the geometry or the attributes of the object changed? Being able to provide information about these particular changes, the former and the updated object need to be compared. Additional useful information can be gathered by determining the former object owner, the version number and the date of the former object. If the change that has been made to the database is a deletion of an OSM object, similar characteristics of the modification analysis process, such as determining the type of object and what the prior object metadata looked like, should be considered.

8.4. Rule-based vandalism detection system

The detection of the introduced types of vandalism in the database allowed for the prototypical implementation of a rule-based decision system, named *OSMPatrol*. One of the main goals of the prototype was to detect the vandalism as fast as possible. For this purpose we used the OSM "Diff"-files and the contained information about changes to the database that are made each minute. Additionally, an OSM database was created to be able to compare the former and the new OSM object. As described, users in OSM can basically provide any kind of additional (semantic) information by tagging the corresponding OSM feature with key-value pairs. Both, the key and the value can typically contain any arbitrary content. Nevertheless, there are community-accepted and well-known OSM Map Features (OpenStreetMap 2012d). To be able to compare and judge the vandalism-likelihood of the additional (semantic) information, our prototypical application matches the added (or altered) information against the well-known OSM Map Features. This is achieved by parsing the OSM wiki page for the map features, extracting the different features (i.e., key-value pairs) and storing them in a database. It was decided to use both the English and the German version of the website as a reference, because these typically contain the most details. As of 13 August 2012, there was a total of 1,139 map features in our database. When evaluating the individual changes, the applied key-value pair is tested against the database with all the map features. Additionally, to retrieve more information about the OSM user, we built a similar database as implemented by Neis (2012). This contributor table contains detailed information such as:

- How many nodes, ways and relations an OSM contributor creates, modifies or deletes?
- What is her/his date of registration? When did she/he start to contribute?
- How often did she/he use one of the most common Tags on an OSM object such as: address, amenity, boundary, building, highway, landuse, leisure, name, natural, railway, sport or waterway?

Both tables are stored in the *OSMPatrol* PostgreSQL database. Figure 8.3 shows the complete architecture of the developed prototype in relation to the OSM project architecture. To retrieve the OSM "Diff"-files, which contain the changes that were made to the database per minute, the OSMOSIS tool is applied. OSMOSIS (OpenStreetMap 2012f) is an open-source command line JAVA tool, which processes OSM data in several different ways. In our particular case, it was also important to update

the newly created OSM PostgreSQL database with the "Diff"-file information to be able to compare the former and newly created or updated OSM objects.

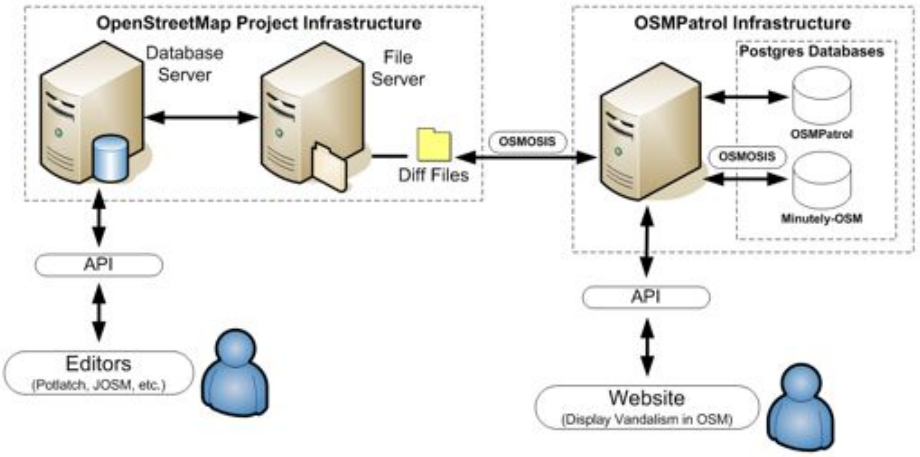


Figure 8.3.: OSM & *OSMPatrol* architecture.

For a better explanation of the different processing steps that are executed every minute to detect potential vandalism in the database, Figure 8.4 shows a UML sequence diagram.

The process starts with the download of the latest OSM "Diff"-file via OSMOSIS. As soon as the download has finished, the main tool, *OSMPatrol*, analyzes this file for signs of vandalism. Prior to testing each edit of the retrieved file, the tool requests two lists that we defined and contain the following information: (a) a list of users who have a history of vandalism incidents (black-list); and, (b) a list of users who should not be considered during the test for vandalism (white-list). During the reviewing process of each OSM edit of the "Diff"-file, information, such as the contributor reputation, the quality of the attributes that have been used and, if necessary, the former object versions, are requested. If an edit is detected as vandalism, it will be stored in an extra table with some additional information. After testing all edits, the OSMOSIS tool will update the OSM database with the regular "Diff"-file, which is an important last step.

Besides the regular architecture and the sequence of the vandalism tool, a major aspect is the assignment of each edit to the corresponding vandalism type. As described in the diagram (Figure 8.4), every edit in OSM will be tested. One assumption that can be made is that new contributors that just joined the project are more prone to errors, mistakes or vandalism in comparison to a more experienced member. Thus, the overall value, which describes if an edit is a potential act of vandalism, is separated into

two parts: the first value summarizes the contributor reputation and the second value rates the edit itself. Additionally, it is possible to filter the results by the corresponding vandalism type and/or by the edits that were created by new contributors.

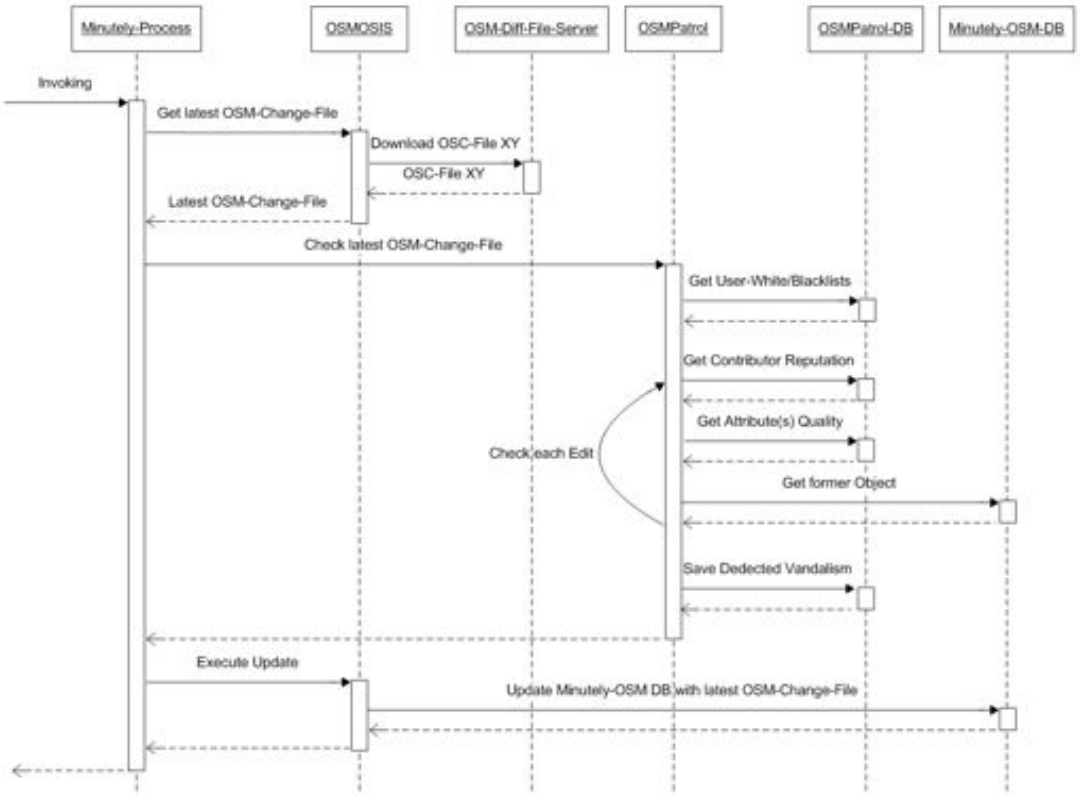


Figure 8.4.: UML sequence diagram of the vandalism detection tool (*OSMPatrol*).

However, to be able to determine a value that represents the reputation of a member, several values need to be integrated. The created objects and corresponding tags that were used during the creation should be equally considered in the reputation value of a user. The aforementioned contributors of the project create more nodes/ways, thus the relation object gets a smaller weight. The used Tags are divided into the "Top12" used Tags of the project. Those contain all established key categories of the OSM Map Features (OpenStreetMap 2012d) list. The reputation value for a contributor can be between 0 and 100%, where 0 represents the value of a new member and 100% an expert member. The following list shows the weight of each aspect to calculate the final user reputation:

- A maximum of 20% for the number of created nodes by the contributor

- A maximum of 20% for the number of created ways by the contributor
- A maximum of 12% for the number of created relations by the contributor
- A maximum of 4% for the number of each of the "Top12" Tags (address, amenity, boundary, building, highway, landuse, leisure, name, natural, railway, sport or waterway)

The registration date of a member was not taken into account for determining the reputation of a contributor, but it was used during the vandalism detection process. Figure 8.5 shows a UML activity diagram for the process of potential vandalism detection of an OSM edit.

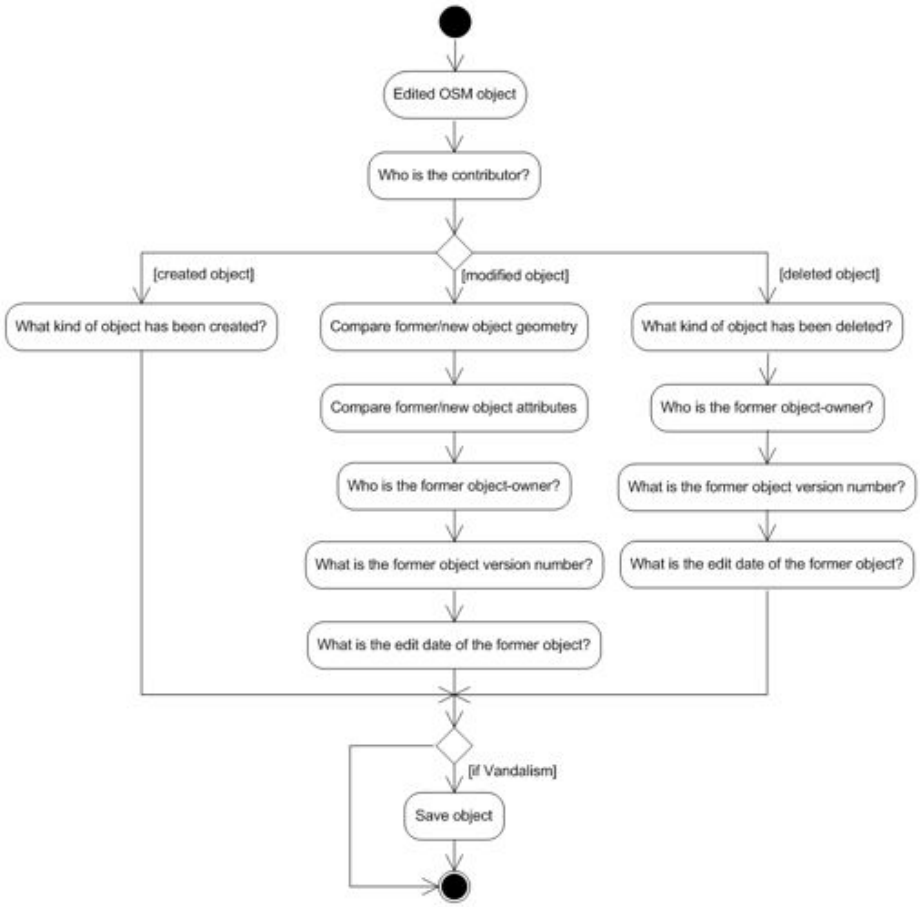


Figure 8.5.: UML activity diagram to detect the types of vandalism of an OSM edit sequence diagram of the vandalism detection tool (*OSMPatrol*).

In general, it can be distinguished between the creation, modification and deletion of an object. In a first step, the contributor who edited the object will be determined. Depending on what type of edit has been made, a value above 0% determines if an edit can be considered as vandalism or not.

If the edit is a newly created object, only the attributes of the object, if available, will be used for designation. At the same time, all attributes will be matched against the map features list (cf. above). If a combination is not available in the map feature list, the vandalism-value of this edit is increased. Thus, the overall vandalism value of an edit can increase according to its number of tags.

If the edit is a modification or deletion of an object, some additional parameters will be checked. The last version of the object is compared with the edit (i.e., the new version), to get answers to questions such as: Who is the former object-owner and does he show a high user reputation value? What is the former object version number, and what is the edit date of the former object? All three items will be used again to increase the overall vandalism value of an edit. Additionally, if the edit is a modification, the geometries will be compared with each other to detect if an object has been moved more than, e.g., 11 m. Also the tags of the latest and the former object will be compared to analyze which tag (key/value) has been changed, added or removed.

8.5. Experimental results

After testing the developed prototype for small areas with known heavy and light vandalism cases, we conducted our final analysis by running the prototype on a dedicated server for one week (14 August 2012–21 August 2012). The server's hardware consisted of a 16 core CPU with 2.93 GHz, 35 GB of RAM and overall 3 TB hard disk space with speeds between 5,400 and 7,200 RPM. During the testing phase, *OSMPatrol* detected about seven Mio "vandalism" edits of 9,200 different users for the entire week. During the same time frame, around 16 Mio edits were made to the OSM database.

This means that the prototype marked 44% of all edits as possible vandalism. The following Figure 8.6 shows the distribution of the affected amounts of nodes, ways and relations. Additionally, the figure provides information about how many of the affected objects were detected during a creation, modification or deletion.

As described by Neis and Zipf (2012), this week basically represents an average week (regarding contribution behavior), meaning that the OSM members contribute to the project in a similar way every other week. A similar statement can be made

about the vandalism edits. We were not able to determine a particular day of the week on which a higher number of vandalism edits took place. Interestingly, almost 1/3 of the 9,200 users who were detected as possible vandalism committers were new contributors to the project. The following Figure 8.7 shows the distribution of edits that were detected as vandalism based on the user reputation. About 50% of the users, for which *OSMPatrol* detected a possible case of vandalism, have a user reputation larger than 66%, indicating that also experienced contributors' actions could be recognized as vandalism. Based on the collected results, users with a reputation level larger than 66% committed 48% of the detected possible vandalism cases. According to these values, about 43% of all detected vandalism edits were committed by new users of the project with a low reputation. Overall, almost 1/3 (36%) of all vandalism users were new users.

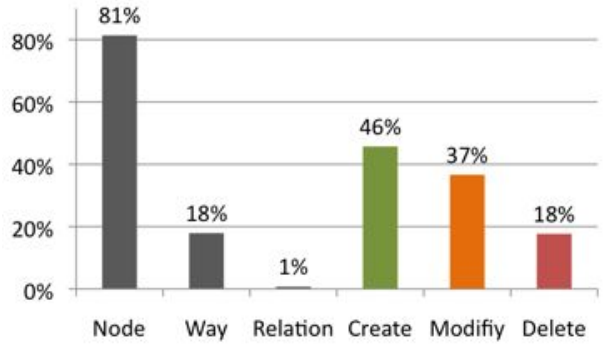


Figure 8.6.: Distribution of objects and edit-types in the detected vandalism (14–21 August 2012).

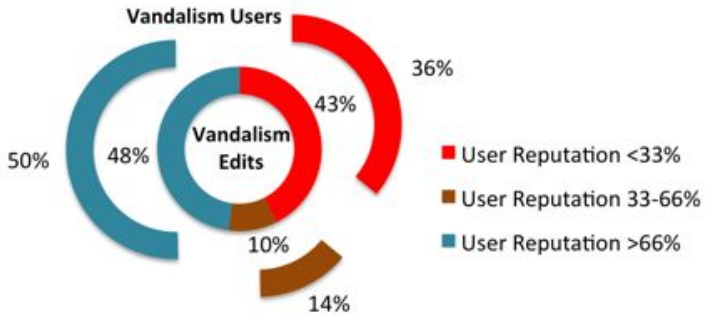


Figure 8.7.: Distribution of vandalism users and vandalism edits based on the user reputation (14–21 August 2012).

Different methods were applied to detect potential cases of vandalism. For each saved vandalism edit there are additional attributes available, e.g., timestamp, user and her/his reputation, vandalism value and a text comment with some information why the edit is marked as vandalism. Based on these values we created three basic filters:

1. Show all edits of new users and/or users with a very low reputation (<5%).
2. Show all users who modified or deleted more than 500 objects within one hour.
3. Show all users who modified node objects and moved the object for more than 500 m.

Based on the first filter, the edits of almost 500 users had to be checked every day. Filter two shows almost 75 users and filter three around 100 users per day in our test phase in August 2012. After investigating at least 30% of the edit provided by these users, it was possible to find at least one real vandalism case per day without taking a deeper look into the type of user edits. Every second case of these vandalism cases was reverted by other OSM contributors within one or two days. In some other cases, the users were blocked by the OSM DWG or the users were contacted via email to raise awareness about their editing errors to the live database. Generally, the results showed that it is feasible to detect real vandalism cases from our detected dataset. However, the analysis also shows that more tools are needed that support the user in analyzing the potential erroneous edit in OSM in an easier and more convenient way. One solution would be to provide a webpage or application that provides detailed information about the tag and/or geometrical changes of the history of an object, as introduced by Huggle (Wikipedia 2012a) and his analysis of Wikipedia articles. With such an application or service, it would be easier to validate the vandalism edits that were detected by *OSMPatrol*.

Within the tested week, about 85% of the detected vandalism edits were committed by only 1,000 users. This shows that the importance of using and maintaining our introduced users black and white list cannot be overestimated. A few users were detected with cases of vandalism based on a large number of objects that were deleted by these users, which were created by the same user in the past. These special cases may allow for future research on how to separate edits made on the user's own data or data contributed by others. *OSMPatrol* was also able to detect a lot of deletions in France, where many active OSM users are currently cleaning up the prior import of buildings to the database.

Overall, *OSMPatrol* was able to detect vandalism types committed by new users,

"illegal" imports and mass edits. However, it can be difficult to distinguish false positive vandalism types from actual cases committed by users with a high reputation level or by users who only delete one or two objects.

8.6. Discussion

When designing the rule base and actually implementing the prototype, a couple of issues and ideas for a more sophisticated prototype became apparent. Those have not been implemented yet, because incorporating those would come at the cost of not being able to perform minute vandalism detection (due to high computation costs). Nevertheless, they will be discussed in this section and maybe incorporated into the system later.

Regarding the individual user's reputation, it was questionable if (and how) to incorporate the project-membership time span (i.e., the time since a user has registered). Is this parameter an indicator for vandalism or not? Is a change by a user who has been registered for four years probably less vandalism than a change by a user who has registered one week ago? The pure incorporation of the project-belonging does not provide any new knowledge, because quite often users register without actually editing the database. However, after a couple of months (or years) they might come back and perform their very first edit—does this now mean that it is vandalism or not? Therefore, a combined consideration of the project-belonging and the activity in the community (a so-called activity-ratio) could lead to an additional indicator for the detection of vandalism; however a real implementation of this factor has not been realized in the presented prototype here.

As described above, massive changes on the geometry of a feature, such as moving a POI ten or more meters, is likely to be vandalism. However, being aware of this as a vicious contributor, it is possible to split the one geometric change into several small ones, which will probably not be detected. For example, instead of moving a POI 15 m at once to a different location, a user could also move the POI 15 times one meter to a different location. The former incident is detected as vandalism, whereas the latter one would not be detected, thus it is defined as being a safe change. Therefore, an extended and more sophisticated vandalism detector also needs to consider multiple changes for the same object over time. These might be detectable through time stamps that are close to each other for the same object by the same user.

Another questionable indicator for vandalism is the evaluation of the version number of an OSM feature. When creating a new feature, the version number is set to one.

The version is incremented with each change of this distinct object (regardless of the actual change). But, is a change on an object with a high version number more likely a type of vandalism than a change on an object with a low version number? What about the opposite situation? A couple of investigations revealed that there are so-called "heavily edited objects" (Mooney and Corcoran 2012) in OSM, but it is not known if changes on those objects are more likely vandalism or not. One could argue that the higher the version number is, the less likely the feature is prone to vandalism, because a larger version value also means more potential feature reviewers. But what about one single user who changes an object a couple of times? Does this also indicate the correctness of the object? These factors indicate that investigations that solely rely on the version numbers are not a good indicator for the vandalism likelihood of an object. Nevertheless, combining the version number and the amount of distinct editors of an object probably represents an appropriate indicator.

When defining the rule base and designing the prototype, it also became apparent that vandalism detection is closely related to data validation (to some extent). One example for a potential indicator for the vandalism likelihood of a change in OSM is the consideration of the neighborhood or surrounding of a newly created or edited object. For example, changing a road in a residential area from residential to primary is very likely vandalism (or a necessary correction of a previous mistake by an inexperienced user). When only considering the change itself, these types of edits will not be detected, whereas the incorporation of the objects in the neighborhood probably provides additional justification for the detector. The fact that the neighborhood needs to be regarded is furthermore underpinned by Tobler's first law of geography (Sui 2004), stating that, "Everything is related to everything else, but near things are more related than distant things".

Changing the name of an existing object or adding a name to a new object might also contain vandalism; however it is pretty hard (or potentially impossible) to properly distinguish between vandalism and validation of names. For example, extending the abbreviation of a name to its full name (e.g., changing str to street) is not vandalism. As a possible solution, comparing names to an existing dictionary of common terms might provide clarity.

In contrast to the aforementioned issues, some vandalism related aspects, such as the IP address, cannot (yet) be implemented in the client due to missing data (it is not possible to gather the IP address of a OSM contributor). However, having such information might be a good (additional) indicator for the vandalism likelihood. It can be investigated if the IP (and the access point) suits the area in which the change has

been performed. For example, if a user changes a street in a country that is hundreds of kilometers away, this might be more likely vandalism than changes of a street name next to the access point of a user, similar to what is described by West et al. (2010) for Wikipedia vandalism detection. Additionally it could be useful to save the IP address of a user that commits the vandalism to block the user from the project and prevent any future vandalism.

8.7. Conclusions and future work

In this paper, we investigated past vandalism incidents, the current OSM database and its contributions, as well as related Wikipedia vandalism detection tools. Based on the results gathered, we developed a comprehensive rule-base prototypical tool that allows automatic vandalism detection in OSM. It can be concluded that it is (to some extent) possible to detect vandalism by applying a rule-based methodology. During the testing phase in August 2012, the prototype marked around seven million edits as potential vandalism. By creating several filters, we were able to determine at least one real vandalism case per day. Overall, the detected vandalism was committed by all types of OSM users and not only by new users or users with a low reputation.

However, as discussed in the previous section, there are some limitations and restrictions. Although those can be solved, they will likely come at the cost of slow performance, thus the initial aim of a minute detection can probably not be realized (at least in the presented prototype and the current server configuration). Furthermore, although vandalism can be detected, it needs to be stated that a manual review of the correctness is still preferable.

As described beforehand, the aim of the conducted research is not the validation of OSM data, but the detection of vandalism. However, the separation between these two domains is not always clear. The focus lies on vandalism detection, because OSM (and especially the editors) already incorporate different methodologies for quality assurance and validation. For example, JOSM informs a user prior to the upload if there are any intersecting geometries or duplicated elements. However, the editors only inform the user, but do not refuse to actually upload the changes.

The principle for the vandalism detection of our prototypical implementation is similar to the basic approach of a firewall: prefer the detection of too many rather than detection of too little cases of vandalism. Thus, the prototype tends to detect more vandalism cases than there actually are in reality. That way, it is assured that there are less misses, but also more false positives. However, as the system only informs

about vandalism (instead of actually blocking vandalism), this is rather uncritical.

For future work, it will be important to enhance the API of the developed prototype. By providing the gathered results (i.e., the detected vandalism) via a well-defined interface, other application developers can use these results for their purposes. One possible (and desirable) application is a tool that enables users to register as a patrol for a distinct area. This way, a user can define a distinct region and/or distinct attributes and as soon as *OSMPatrol* detects a vandalism type that suites to a patrol's preference pattern, he or she is informed via. e-mail. Another potential application is a platform that enables well-known users to highlight edits as vandalism or non-vandalism and to maintain an appropriate white-list for the *OSMPatrol*.

In general, the topic of vandalism detection and prevention is also being discussed in the OSM community, e.g., limitation of the OSM API. As mentioned before, a possible solution could be to allow experienced users to review submitted content of new and, maybe, inexperienced users. This approach has already been implemented by the Google MapMaker platform (Google 2012). However, in this case, the question remains: Are there enough volunteers available that are willing to work on some manual data validation in the future?

References

Adler, B., Alfaro, L. de, Mola-Velasco, S., Rosso, P., and West, A. (2011). Wikipedia Vandalism Detection: Combining Natural Language, Metadata, and Reputation Features. In: *Computational Linguistics and Intelligent Text Processing*. Berlin, Germany: Springer.

Bing (2010). *Bing Engages Open Maps Community - Bing Maps Blog*. URL: `http://www.bing.com/community/site_blogs/b/maps/archive/2010/11/23/bing-engages-open-maps-community.aspx` (visited on 08/11/2012).

Brando, C. and Bucher, B. (2010). Quality in User-Generated Spatial Content: A Matter of Specifications. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

Caminha, C. and Furtado, V. (2012). Modeling User Reports in Crowdmaps as a Complex Network. In: *Proceedings of 21st International World Wide Web Conference*. (Apr. 16–20, 2012). Lyon, France.

Chin, S.C., Street, W.N., Srinivasan, P., and Eichmann, D. (2010). Detecting Wikipedia Vandalism with Active Learning and Statistical Language Models. In: *Proceedings*

of the 4th Workshop on Information Credibility (WICOW '10). (Apr. 27, 2010). Raleigh, NC, USA.

Diaz, L., Granell, C., Gould, M., and Huerta, J. (2011). Managing User-generated Information in Geospatial Cyberinfrastructures. *Future Generation Computer Systems*, 27, 304–314.

Girres, J.F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS*, 14, 435–459.

Goetz, M. and Zipf, A. (2012). Using Crowdsourced Indoor Geodata for Agent-Based Indoor Evacuation Simulations. *ISPRS International Journal of Geo-Information*, 1, 186–208.

Goodchild, M.F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211–221.

Google (2012). *Map Makers. Frequently Asked Questions*. URL: `http://www.google.com/mapmaker/mapfiles/s/faq.html` (visited on 09/12/2012).

Haklay, M. (2010). How good is OpenStreetMap Information: A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. *Environment and Planning B*, 37, 682–703.

Heipke, C. (2010). Crowdsourcing Geospatial Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65, 550–557.

ITO (2012). *OSM Mapper*. URL: `http://www.itoworld.com/static/openstreetmap_tools/osm_mapper.html` (visited on 08/05/2012).

Kittur, A. and Kraut, R.E. (2008). Harnessing the Wisdom of Crowds in Wikipedia: Quality through Coordination. In: *Proceedings of the 2008 ACM Conference on Computer Supported Cooperative Work*. (Nov. 8–12, 2008). San Diego, CA, USA.

Ludwig, I., Voss, A., and Krause-Traudes, M. (2011). A Comparison of the Street Networks of Navteq and OSM in Germany. *Advancing Geoinformation Science for a Changing World*, 1, 65–84.

Mola-Velasco, S.M. (2011). Wikipedia Vandalism Detection. In: *Proceedings of the 20th International Conference Companion on World Wide Web (WWW '11)*. (Mar. 28–Apr. 1, 2011). Hyderabad, India.

Mooney, P. and Corcoran, P. (2012). Characteristics of heavily edited Objects in OpenStreetMap. *Future Internet*, 4, 285–305.

Mooney, P., Corcoran, P., and Ciepluch, B. (2013). The Potential for using Volunteered Geographic Information in Pervasive Health Computing Applications. *Journal of Ambient Intelligence and Humanized Computing*, 4(6), 731–745.

Mooney, P., Sun, H., Corcoran, P., and Yan, L. (2011). Citizen-Generated Spatial Data and Information: Risks and Opportunities. In: *Proceedings of IEEE International Conference on Intelligence and Security Informatics*. (July 10–12, 2011). Beijing, China.

Neis, P. (2012). *How Did You Contribute to OpenStreetMap?* URL: `http://hdyc.neis-one.org` (visited on 08/01/2012).

Neis, P., Zielstra, D., and Zipf, A. (2012). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4, 1–21.

Neis, P. and Zipf, A. (2008). OpenRouteService.org is Three Times "Open": Combining OpenSource, OpenLS and OpenStreetMaps. In: *Proceedings of the GIS Research UK 16th Annual conference GISRUK 2008*. (Apr. 2–4, 2008). Manchester, UK.

Neis, P. and Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project - The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 146–165.

OpenGeoData (2012). *Google IP Vandalizing OSM*. URL: `http://opengeodata.org/google-ip-vandalizing-openstreetmap` (visited on 08/05/2012).

OpenStreetMap (2012a). *Automated Edits Code of Conduct - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Automated_Edits/Code_of_Conduct` (visited on 08/20/2012).

OpenStreetMap (2012b). *Data Working Group - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Data_working_group` (visited on 08/05/2012).

OpenStreetMap (2012c). *Import/Catalogue - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Import/Catalogue` (visited on 08/20/2012).

OpenStreetMap (2012d). *Map Features - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Map_Features` (visited on 08/13/2012).

OpenStreetMap (2012e). *OpenStreetMap Statistics*. URL: `http://www.osm.org/stats/data_stats.html` (visited on 08/05/2012).

OpenStreetMap (2012f). *Osmosis - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Osmosis` (visited on 08/11/2012).

OpenStreetMap (2012g). *OSM User Blocks*. URL: `http://www.osm.org/user_blocks` (visited on 08/01/2012).

OpenStreetMap (2012h). *Planet OpenStreetMap*. URL: `http://planet.osm.org` (visited on 08/01/2012).

OpenStreetMap (2012i). *Vandalism - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Vandalism` (visited on 08/05/2012).

OpenStreetMap (2012j). *Vandalismus Zwijndrecht*. URL: http://wiki.osm.org/wiki/File:Vandalismus_Zwijndrecht.gif (visited on 08/11/2012).

OpenStreetMap (2012k). *Watch List - OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/OWL_%28OpenStreetMap_Watch_List%29 (visited on 08/05/2012).

OSMstats (2012). *Statistics of the Free Wiki World Map*. URL: http://osmstats.altogetherlost.com (visited on 08/11/2012).

Over, M., Schilling, A., Neubauer, S., and Zipf, A. (2010). Generating Web-based 3D City Models from OpenStreetMap: The Current Situation in Germany. *Computers, Environment and Urban Systems*, 34, 496–507.

Potthast, M. (2010). Crowdsourcing a Wikipedia Vandalism Corpus. In: *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. (July 19–23, 2010). Geneva, Switzerland.

Potthast, M., Stein, B., and Gerling, R. (2008). Automatic Vandalism Detection in Wikipedia. In: *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*. (Mar. 30–Apr. 3, 2008). Glasgow, Scotland.

Sui, D.Z. (2004). Tobler's first law of geography: A big idea for a small world? *Annals of the Association of American Geographers*, 94, 269–277.

Van den Berg, H., Coetzee, S., and Cooper, A.K. (2011). Analysing Commons to Improve the Design of Volunteered Geographic Information Repositories. In: *Proceedings of AfricaGEO 2011*. (May 31–June 2, 2011). Cape Town, South Africa.

West, A.G., Kannan, S., and Lee, I. (2010). Detecting Wikipedia Vandalism via Spatio-temporal Analysis of Revision Metadata? In: *Proceedings of the Third European Workshop on System Security (EUROSEC '10)*. (Apr. 13, 2010). Paris, France.

Wikipedia (2012a). *Huggle*. URL: http://en.wikipedia.org/wiki/Wikipedia:Huggle (visited on 09/24/2012).

Wikipedia (2012b). *Subtle Vandalism Taskforce*. URL: http://en.wikipedia.org/wiki/Wikipedia:Subtle_Vandalism_Taskforce (visited on 09/12/2012).

Zielstra, D. and Hochmair, H.H. (2011). A Comparative Study of Pedestrian Accessibility to Transit Stations using Free and Proprietary Network Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2217, 145–152.

9. A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis

Authors

Christopher Barron, Pascal Neis, and Alexander Zipf

Journal

Transactions in GIS

Status

Published: 23 December 2013 / Accepted: 28 October 2013 / Revised: 27 October 2013 / Submitted: 8 August 2013

Reference

Barron, C., Neis, P., and Zipf, A. (2013). A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. *Transactions in GIS*, doi: 10.1111/tgis.12073

Contribution statement

Christopher Barron mainly wrote the manuscript and implemented the comprehensive framework. Pascal Neis contributed the idea and design for the implementation of the framework. Furthermore, both co-authors supported this publication by continuing discussions about the methods and the results of the study.

.....	.....
Prof. Dr. Alexander Zipf	Christopher Barron

Abstract

OpenStreetMap (OSM) is one of the most popular examples of a *Volunteered Geographic Information* (VGI) project. In the past years it has become a serious alternative source for geodata. Since the quality of OSM data can vary strongly, different aspects have been investigated in several scientific studies. In most cases the data is compared with commercial or administrative datasets which, however, are not always accessible due to the lack of availability, contradictory licensing restrictions or high procurement costs. In this investigation a framework containing more than 25 methods and indicators is presented, allowing OSM quality assessments based solely on the data's history. Without the usage of a reference data set, approximate statements on OSM data quality are possible. For this purpose existing methods are taken up, developed further, and integrated into an extensible open source framework. This enables arbitrarily repeatable intrinsic OSM quality analyses for any part of the world.

Keywords: OpenStreetMap; volunteered geographic information; spatial data quality assessment; intrinsic approach.

9.1. Introduction

In the past decade a significant transition within the *World Wide Web* (WWW) was carried out leading to an altered usage of the WWW where users no longer act as sheer consumers of pre-defined content. Instead, they are more and more part of a contributing process, sharing knowledge and information (O'Reilly 2005). Popular examples are the Internet encyclopedia Wikipedia and content sharing platforms such as Flickr for photos and Youtube for videos. These platforms provide the opportunity of contributing various types of content, so-called *User-Generated Content* (UGC) (Brando and Bucher 2010, Chilton 2012, Goodchild 2009). Besides, *Volunteered Geographic Information* (VGI) can be thought of as a special case of UGC. VGI, also referred to as crowd-sourced geodata, is defined as the collaborative acquisition of geographical information and local knowledge by volunteers, amateurs or professionals (Goodchild 2007). Among others, e.g. Map Insight[1], Map Reporter[2]), Wikimapia[3] or Google Map Maker[4], *OpenStreetMap* (OSM) has evolved to one of the greatest and most famous VGI projects in the past years (Chilton 2012, Goodchild and Li 2012) with 1.3 million

[1]http://mapinsight.teleatlas.com
[2]http://mapreporter.navteq.com
[3]http://wikimapia.org
[4]http://www.google.com/mapmaker

users registered at August 2013 (OpenStreetMap 2013h). Since commonly no authorized instance examines the contributed information, data quality assurance plays a crucial role within the OSM project (Flanagin and Metzger 2008, Goodchild and Li 2012). This fact is becoming more and more important, not least because OSM turns out to be a serious geodata alternative for different applications and is used in a wide range of *geographic information systems* (GIS) and applications (Amelunxen 2010, Goetz and Zipf 2013, Hagenauer and Helbich 2012).

A commonly used way of assessing the OSM data quality is the comparison with ground truth reference datasets (Girres and Touya 2010, Haklay 2010, Helbich et al. 2012, Mooney et al. 2010b, Neis et al. 2012, Zielstra and Hochmair 2012). However, accessibility to high quality and commercial datasets for such extrinsic analyses is often limited due to costs and licensing restrictions (Mooney et al. 2010b). Therefore, suitable alternatives are necessary. The key motivation for this article is to investigate how OSM data can be evaluated without a reference for comparison purposes. One possible approach is the investigation of the data's history (Exel et al. 2010). From the data's history, so-called intrinsic indicators present one opportunity to supply information regarding the data quality. For this purpose, however, new methods, indicators and visualization types are needed to evaluate the quality of OSM data. To this end, a framework for intrinsic OSM data quality analyses, named *iOSMAnalyzer*, was developed. The framework was implemented as a tool using free and open source components. This allows anyone to generate information about OSM data quality for a freely selectable area using only OSM's data history. In the context of spatial data quality analysis, Devillers et al. (2002) discussed the limitations of metadata as an assessment factor to help users to evaluate if a dataset is usable or not. These findings resulted in the introduction of a system which proved that it is relevant to include data quality visualization issues in the communication between producers and users of the data (Devillers et al. 2002, Devillers et al. 2007). Thus an additional motivation was that the developed framework should facilitate the decision whether the quality of OSM data in a selected area of a user's choice is sufficient for her or his use case or not.

The remainder of this article is organized as follows. Section 2 gives an introduction to OSM and summarizes related scientific studies on OSM data quality. The main focus in this work is on Section 3. Before a short introduction to the developed framework, several methods and indicators for intrinsic OSM quality analyses are introduced. Section 4 evaluates the outcome using exemplary results of different regions. Finally, Section 5 summarizes the results of this investigation and discusses further research

needs.

9.2. The OpenStreetMap project: Introduction and related state of the art research

The goal of the OSM project is to create a free and editable world map (Ramm et al. 2010). Within the project volunteers, amateurs and professionals from different social worlds (Lin 2011) act as sensors (Flanagin and Metzger 2008) and collect geographic data. This bottom up process stands in contrast with the traditional centralized procedure of collecting geographic data (Goodchild 2007). The motivation for contributing to OSM varies heavily: it ranges from self-expression over manifestation and representation of people's online identity to a simple fun factor. Meaningful extracurricular activities, interesting technologies and a fascinating general project development are further motivational reasons (Budhathoki 2010). In general, data for OSM can be derived from multiple sources and edited and imported by means of different freely available editors. The most popular editors are the *Java OpenStreetMap Editor* (JOSM[5]), the online flash editor Potlatch[6] or the web-based JavaScript editor iD[7]. The classic approach is the collection of spatial data with portable and GPS enabled devices. In addition, several companies such as Aerowest, Microsoft Bing (Bing 2010) or Yahoo! released, at least temporarily, their aerial images for the OSM project. The community is allowed to use these images as a base layer for tracing geographic features, such as for example buildings, forests or lakes. The contributors' local knowledge is also a valuable source of geographic information. Furthermore, datasets, which fit the licensing restrictions, can also be imported to the OSM database. At best, this is done in close collaboration with the (local) community and respective mailing lists as the appropriateness of imports is discussed controversially (Zielstra et al. 2013).

All contributed data is stored according to the OSM data model wherein point features are represented by "Nodes" and linear features by "Ways". Polygonal objects are represented by "closed Ways". Additionally, features can be further specified semantically by key-value pairs, so-called tags. There are no restrictions to the usage of tags. Whereas traditional authoritative and commercial data sets usually follow the *Resource Description Framework* (RDF) notion, each OSM feature can hold multiple tags or no tags at all. Nevertheless, for the purpose of consistency, it is recommended to

[5]http://josm.openstreetmap.de/
[6]http://www.openstreetmap.org/edit?editor=potlatch2
[7]http://ideditor.com/

use commonly accepted key-value pairs from the OSM map features web page (OpenStreetMap 2013b). Finally, Relations are used to model logical relationships between the previously mentioned features (Ramm et al. 2010).

9.2.1. Parameters of geodata quality

Quality in general plays a key role when working with all kinds of geodata, especially in data production and assessment (Veregin 1999) or exchange (Goodchild 1995). This is especially the case with OSM data, as the contributors are not faced with any restrictions during the data collection and annotation process. In the field of geo-information, the principles of the "*International Organization for Standardization*" (ISO) can be taken into account for quality assessment. The ISO 19113[8] standard describes general principles of geodata quality and ISO 19114[9] contains procedures for quality evaluation of digital geographic datasets. The ISO 19157[10] "Geographic Information: Data Quality" standard, currently under development, aims to harmonize all standards related to data quality and revises the aforementioned standards. The quality of spatial data can be evaluated with the help of following elements of ISO 19113:

- "Completeness": describes how complete a dataset is. A surplus of data is referred to as "Error of Commission", a lack of data in contrast as "Error of Omission".
- "Logical Consistency": declares the accuracy of the relations manifested within a dataset. This element can be further subdivided into "intra-theme consistency" and "inter-theme consistency".
- "Positional Accuracy": defines the relative and absolute accuracy of coordinate values.
- "Temporal Accuracy": the historical evolution of the dataset.
- "Thematic Accuracy": describes the accuracy of the attributes assigned to a geometry.

However, OSM data quality heavily depends on the purpose for which the data will be deployed. We refer to this as "Fitness for Purpose" assessment, previously defined by Veregin (1999) as determining "fitness-for-use".

[8] http://www.iso.org/iso/catalogue_detail.htm?csnumber=26018
[9] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=26019
[10] http://www.iso.org/iso/catalogue_detail.htm?csnumber=32575

9.2.2. Quality assessment in OpenStreetMap – Overview of related scientific research

The increasing availability of voluntarily and collaboratively collected geodata, in particular OSM, led to numerous scientific studies with a focus on the evaluation of this data. In the beginning, investigations mainly focused on the OSM road network with the help of a ground truth reference dataset. For instance, Haklay (2010) compared the OSM road network with Ordnance Survey Meridian 2 for England and Kounady (2011) with the *Hellenic Military Geographical Service* (HMGS) dataset for Athens, Greece. For Germany, Zielstra and Zipf (2010) conducted a comparison with TeleAtlas-MultiNet, Ludwig et al. (2011) with Navteq and, for the period from 2007 until 2011, Neis et al. (2012) with TomTom MultiNet. Using a different method, Helbich et al. (2012) also investigated the positional accuracy of the OSM road network. Employing a spatial statistical comparison method, the authors compare identical road junctions with TomTom and official survey data as a reference. All previously mentioned studies on the OSM road network show, broadly speaking, one commonality: a high positional accuracy and a huge amount of details are found around urban areas with a high number of contributors. In contrast, more rural areas often show a lower level of OSM data quality. However, some urban areas, Istanbul for example, show a high number of contributions by mappers with their main activity area located more than 1,000 km away (Neis et al. 2013). The authors state that in some way this mapping behavior contradicts the original idea of VGI as projects where people contribute their local knowledge.

Beside the road network, other features of OSM have also been object of interest for quality investigations. Mooney et al. (2010a) compare OSM land cover features with the *Ordnance Survey Ireland* (OSI) dataset using shape similarity tests. Girres and Touya (2010) examine different quality aspects of the French OSM dataset according to quality principles stated in ISO 19113: 2002. The authors highlight the problem of heterogeneity in VGI datasets such as OSM which is, among others, caused by the contributors' freedom within the data collection process.

As mentioned, a considerable number of studies have evaluated different data quality aspects of OSM data. In most cases, data is evaluated and compared with authoritative datasets. On the other hand, very few studies target analyses conducted without a reference dataset. However, as stated by Batini and Scannapieco (2006), intrinsic data quality analyses capture the data's inherent quality and according to Wang and Strong (1996) and Batini and Scannapieco (2006) it includes the following dimensions:

accuracy, objectivity, believability and reputation. In the case of OSM, Mooney and Corcoran (2012) attempted to assess the quality of OSM features by analyzing objects with more than 15 different versions ("heavily edited objects") for several countries utilizing an OSM-Full-History-Dump. In another investigation the authors examine the tag assignment and the influence of the number of contributors on it (Mooney et al. 2010a). However, the results of both studies showed that the number of contributors to an object does not necessarily relate to the number of tags of an OSM feature. Furthermore, recent studies increasingly dwell on the contributors behind the submitted data. In this context the terminologies "feature quality", "user quality" and their "interdependency" are introduced (Exel et al. 2010). MVP-OSM, a tool for identifying areas of high quality contributions, also uses this approach. The tool's results are based on the contributors' local knowledge, identifying their experience and community recognition for a selected area (Napolitano and Mooney 2012). Moreover, a conceptual model to analyze contributor patterns in VGI projects in a more structured way is proposed (Rehrl et al. 2013). Following a data-orientated approach, a provenance vocabulary is presented by Kessler et al. (2011) allowing statements on the lineage of OSM data based on the (editing) history. Touya and Brando-Escobar (2013) proposed a method to infer the level of detail of OSM features based on several criteria such as the geometric resolution and the feature type. A model to estimate the uncertainty of geometric measurements of vector objects has been introduced by Girres (2011). For the investigation and estimation of length errors for different kinds of road samples the TOP100 road network of France have been utilized.

However, the overall activity of the contributors within the OSM project is analyzed where the authors illustrate a strong bias in the participation process (Neis and Zipf 2012). It should be noted that only 38% of all registered members have ever accomplished at least on edit and only 5% of all members have contributed in a significant manner. This behavior is closely linked to what is commonly known as the "Participation Inequality" in online communities (Nielsen 2006). The Participation Inequality follows a 90-9-1 pattern and is observed in several communities in the WWW, which rest on contributions of their members. Within those, 90% of all users solely consume information without contributing, 9% contribute from time to time and only 1% is actually responsible for the majority of the content.

9.2.3. OpenStreetMap data, history and tools

OSM data can be obtained from different sources in several file formats. A weekly updated OSM database dump-file (OpenStreetMap 2013g) containing a temporal snap-

shot of the entire world with a current compressed size of approximately 28 GB is available. Smaller extracts of specific regions are also offered by companies (Geofabrik 2013). These files are mainly provided as Esri-shapefiles (*.shp), in a compressed XML (*.osm.bz2) or a Protocolbuffer Binary Format (*.osm.pbf) which makes faster processing possible. Several tools are capable of processing OSM data. Probably the most prominent ones within the OSM software environment are the open source command-line Java tool OSMOSIS (OpenStreetMap 2013f) or the flexible C++/JavaScript framework Osmium (OpenStreetMap 2013e).

Beside the abovementioned snapshot of the recent database, the OSM-Full-History-Dump contains, with minor exceptions, the entire history of the OSM data. A new version of an OSM object is created whenever a feature's geometry is changed. The simple movement of an already existing Way's Node does not lead to a new version number. Moreover, adding, modifying, or deleting a tag also leads to the increase of a feature's version number. Regarding the versioning, a bug in the Potlatch 1 OSM editor up to 2011 has to be noted. This bug led to an erroneous increase of a feature's version number, although it was not edited, but lay within the spatial extent of an edited OSM changeset (Neis and Zipf 2012). The fact that not every change automatically leads to a new version of an OSM feature and vice versa has to be considered carefully.

Since the introduction of the OSM API 0.5 in October 2007, the recent OSM-Full-History-Dump includes every undertaken addition, modification and deletion within OSM. From contributions during or before OSM API 0.4, only a snapshot of the data which was actually visible at the changeover together with the history of their future changes are available. Moreover, as segments were removed with the introduction of the OSM API 0.5 they are also not included within the OSM-Full-History-Dump. Furthermore, it has to be noted that added or modified data of contributors who did not accept the "*Open Database License 1.0*" (ODbL) terms during the license change period are also not part of the current OSM-Full-History-Dump any more (OpenStreetMap 2013g).

9.3. Introducing the framework

In the following subsections, the developed *iOSMAnalyzer* framework will be delineated. After an introduction to the applied techniques, several methods and indicators for evaluating OSM data are presented. Finally, the structure of the developed framework is illustrated. The main focus is on the quality assessment for different "*Location Based Services*" (LBS) applications.

The approach proposed in this article differes significantly from previously conducted studies in several respects. An OSM-Full-History-Dump is used as a sole input for the analyses. Therefore, with the exception of the aforementioned particularities, the entire temporal dimension of the dataset can be taken into account. Furthermore, no ground truth reference dataset is deployed for OSM data quality evaluation. Therefore, specified areas within OSM can be evaluated regardless of whether a reference is available or not. Thus, a so-called intrinsic analysis approach is applied. For this intrinsic approach, new methods and indicators have been developed because traditional ones from extrinsic analyses are usually only suitable for comparison purposes. Excerpts of numerous techniques presented in this article, for example, are the investigation of the data's historical development, the comparison of features' characteristics at different timestamps or various spatial analyses. In some cases (e.g. feature completeness), however, an intrinsic approach does not allow absolute statements on data quality. Therefore, some results presented with this approach can only act as relative indicators making approximate statements about the possible data quality.

9.3.1. Defining a framework for intrinsic OSM quality assessment

As OSM data is used in a wide range of applications, the analyses have to be adjusted to different use cases and specific needs. Hence, in order to evaluate the OSM data, the finally calculated results of the *iOSMAnalyzer* are divided into the following categories which were selected according to the "Fitness for Purpose" approach: "General Information on the Study Area", "Routing and Navigation", "Geocoding", "Points of Interest-Search", "Map-Applications" and "User Information and Behavior". The fitness of VGI data always depends on the case and needs to be analyzed individually (Mondzech and Sester 2011, Neis et al. 2013). Therefore the framework's results can support the decision whether OSM data could be suitable for one of a great number of use cases. Overall, a set of more than 25 different intrinsic quality indicators (see Figure 9.1) is considered in the framework. Due to space restrictions a selected number of parameters are presented in the following.

9.3.2. General information on the study area

The evolution of OSM features over a specific period of time provides a first insight into the development and quality of an arbitrarily chosen area within OSM (Neis et al. 2013). For example the cumulated number of contributed points, lines and polygons per month gives a first general and more diverse impression of an area. Histograms

allow the visualization of these quantitative developments which have to be interpreted in very different ways: Ciepluch et al. (2011) allege that OSM datasets rise from the road network. Therefore, first peaks within line evolution histograms could indicate the beginning of general mapping activity. A significantly steeper growth within a few days or weeks is to be expected in case of bulk data imports or automated edits (bots). Both bulk imports and bots are usually documented in the wiki (OpenStreetMap 2012). Rating these automated edits in terms of good or bad quality heavily depends on the individual case (Zielstra et al. 2013).

Figure 9.1.: Overview of the *iOSMAnalyzer*'s intrinsic quality indicators.

People also collaboratively contribute to OSM in organized community events. These so-called "mapping parties" possibly lead to significant data increase in a region mainly

within a few days potentially enriching existing data (Hristova et al. 2013). The release of aerial images also has an impact on the quantity of data (Neis et al. 2012). For instance, significant peaks of contributed data after December 2010 and in spring 2011 are probably caused by the release of the Bing aerial images for the purpose of digitizing. However, to ensure high attribute quality roads as an example requires local knowledge. Only then their category, name or possible speed limitation can be identified correctly (Leeuw et al. 2011).

The quantitative calculations mentioned above allow only limited statements on data quality in some cases. As VGI projects such as OSM are mainly driven by their contributors, not only the data but also the behavior of the crowd can be analyzed to provide information about their contributions (Coleman et al. 2009, Exel et al. 2010, Neis and Zipf 2012, Rehrl et al. 2013). One general indicator is the overall number of (active) contributors within an area. Several investigations demonstrate that a high number of active contributors leads to a stable and good quality OSM dataset which is more probably kept up-to-date (Girres and Touya 2010, Haklay et al. 2010, Neis and Zipf 2012). As a consequence, a high and increasing number of people who have ever created or edited OSM data within an area indicates a possibly better data quality. In addition, the number of actually active contributors per month indicates whether those have contributed only once or in a more frequent way. The higher the number of monthly recurring and contributing mappers, the higher is the heterogeneity of mappers and consequently, the better is the overall data quality. A combination of the general evolution of points, lines or polygons together with the aforementioned information on contributor activity simplifies the interpretation of quantitative feature statistics. Imports and bots are usually carried out by a single registered member leading to a huge amount of created or edited data. By contrast, mapping parties, digitizing from aerial images and simple mapping activities by individuals usually are performed by a high number of contributors. Beside the overall number of contributors, the mappers' actual amount of created data can provide more in-depth information. In a global investigation four different member groups based on the number of created Nodes have been defiened by Neis and Zipf (2012): "Senior Mappers" (contributors with 1,000 and more created Nodes), "Junior Mappers" (contributors with at least 10 and less than 1,000 created Nodes), "Nonrecurring Mappers" (contributors with less than 10 created Nodes) and members with no edits. The more mappers with a high number of contributed Nodes can be identified, the more active contributors are present in an area. However, these measurements do not have to be true for a mapper's activity in general. A person identified as a "Nonrecurring Mapper" in one area could

be a very active mapper with high contribution rates in another area. Furthermore, there is no evidence that users with a high number of created Nodes also contribute high quality data. This could be expected due to their contribution experience in the selected area; however, this potential thesis requires further research. Concerning the distribution of created or edited features among the mappers, OSM shows an inequality in contributions (Neis and Zipf 2012). As stated by Nielsen (2006) this is referred to as the participation inequality in online communities. The less contributors that are responsible for the major proportion of the data the higher the dependence on those few. These contributors are therefore of particular importance for the OSM project. Moreover, a more uniform distribution shows that more people are contributing and this potentially leads, relatively speaking, to a better overall data quality because errors are more likely detected and fixed.

An important point in OSM is the currentness of data (Exel et al. 2010, Neis and Zipf 2012). After the initial collection process the further maintenance of OSM data is essential for a high quality and up-to-date dataset. Ideally the process of updating the OSM features' geometries and attributes is carried out continuously, homogeneously, throughout and is not limited to specific features. However, this is not the usual case within OSM. A possible way to analyze and represent the currentness is the visualization of the data's latest modification. It can be argued that the last editor of an OSM feature is responsible for its correctness and indirectly confirms this by uploading the modified data to the server. The case is problematic when a feature was already completely and accurately mapped in the past. These features can potentially be detected with the help of adjacent features (Exel et al. 2010) using a probabilistic approach. Features with an older timestamp surrounded by current features could represent an implicit peer review and attest to their currentness. The positional accuracy of the OSM data depends very much on the way the data was collected. Several factors such as GPS signal preciseness, displaced aerial images or bulk movements have an impact on data quality. A way to identify these possible positional inaccuracies without a reference dataset is the enhancement and modification of the method proposed by Helbich et al. (2012). Instead of comparing OSM with a ground truth reference dataset, the location of currently valid road junctions is compared with its previous location. As already mentioned, the latter must not necessarily be the last version of the Node representing the junction. By analyzing distance and the angle of two corresponding road junctions within a polar scatter plot, different conclusions regarding the positional accuracy of OSM data can be drawn: on the one hand, an accumulation of points within one angle segment of the diagram indicates possible corrections of the road network

caused by a potentially displaced editing basis (either aerial images or GPS traces). Yet a rectification could also be possible but is not clearly distinguishable from deterioration. However, if multiple road junctions show exactly the same distance and angle to their previous location, a bulk movement is very likely. On the other hand, a uniform distribution where all road junctions show an individual distance suggests no positional inaccuracies caused by the abovementioned issues. Within this proposed method, road junctions serve as an indicator. This means that beside the road network other features within the selected dataset could also be affected by positional inaccuracy. Referring to Touya and Brando-Escobar (2013), future research could investigate "source" tags or changeset comments of the corresponding features which could provide information about the method of acquisition or the reason for the displacement. The following list contains a summary of the relevant parameters of this section:

- Characterization of active mappers
- Currentness of data
- Evolution of OSM features
- Number of (active) contributors
- Positional accuracy of the OSM (road network) data

9.3.3. Geodata quality assessment for location based services

9.3.3.1. Routing and navigation

The completeness of the OSM road network plays a significant role in routing and navigation applications and has therefore been the subject of several investigations (Girres and Touya 2010, Haklay et al. 2010, Kounady 2011, Ludwig et al. 2011, Neis et al. 2012). In contrast to these thorough comparative studies, this investigation follows an intrinsic approach without the usage of any reference data. As illustrated in the example of Germany, roads in OSM are mapped completely, mainly in order of their hierarchy (Neis et al. 2012). In the beginning usually motorways are the first to be mapped completely. They are subsequently followed by municipal roads, streets in residential areas and all other roads such as forest tracks or smaller paths. Taking this information into account, a category of roads can be stated as "close to completion" if the monthly increase in length is very small or even close to zero. This assumption can be affirmed by a high number of active contributors along with an increasing length of mapped roads in other lower hierarchical road categories. In particular because it shows that contributors did not simply stop mapping, but instead, because

of the potential completeness of a road category, switched to a lower road category which is not completely mapped yet. However, this method can only be considered as a way to approximate the quality parameter completeness. Absolute statements on the completeness of the road network are only possible with the help of a ground truth reference dataset. However, a huge benefit of this indicator can be seen in its independency of a reference dataset which makes it applicable for any region in the world.

Beside the completeness, the logical consistency of the OSM road network is also one key element in routing applications. In accordance with Neis et al. (2012) three topological errors are taken into account by means of internal tests: (1) roads which are erroneously not connected to each other at junctions; (2) duplicate road geometries; and (3) intersecting roads without a common Node (3). The first inconsistency is identified by analyzing roads which do not share a common Node with another one and and lie within a radius of one meter. The second inconsistency is identified by calculating duplicate road geometries. The third inconsistency is detected by analyzing roads which intersect but do not share a common Node. This can also be caused by missing tags characterizing bridges or tunnels. These topological errors are calculated, quantified and subsequently visualized each on a single map. The relevant parameters of this section are:

- Completeness of the OSM road network
- Logical consistency of the OSM road network

9.3.3.2. Geocoding

The process of associating exact geographic locations with data such as street names or house numbers is generally referred to as geocoding (Amelunxen 2010) and plays a key role in many LBS applications. For this purpose, complete address information is necessary. By now, the OSM community has widely agreed on the so-called "Karlsruhe Schema" as a way to add addresses to OSM (OpenStreetMap 2013a). Within this schema house numbers are mapped either as single Nodes, as additional tags to existing features or as interpolation lines determining the start and end house number of a specified line (Ramm et al. 2010). As applications in LBS are not necessary capable of utilizing all these three methods, the overall distribution of house numbers over time gives a first impression of the fitness for one's needs. Other applications might be interested in the number of OSM features containing a complete address annotation. In comparison with the overall distribution of house numbers or house names the

cases with complete annotations demonstrate the attribute completeness of the existing address information according to the aforementioned "Karlsruhe Schema". This is of particular importance for LBS, which are not able to calculate parts of an address by means of spatial queries from administrative boundaries and, furthermore, shows the attribute completeness of the appropriate features.

Moreover, good routing and navigation applications are characterized by accurate geocoding results up to the level of single buildings. To this end, all buildings which are likely to contain a house number or house name are calculated. Doing this, not only are actually annotated building polygons considered but also information derived from spatially intersecting Nodes or interpolation lines with address information is taken into account. By now, no algorithm is known which can distinguish between buildings which should have a house number or house name or not. Therefore, all buildings with a smaller basis than 10 m^2 and with a specified list of tags are excluded (e.g. building=roof, building=garage, etc.).

Figure 9.2 visualizes this issue. The bottom right building is erroneously not annotated with a house number/name whereas the bottom left one, due to its size being less than 10 m^2, does not need a house number/name. The latter can be presumed to be a small hut without an official house number or house name.

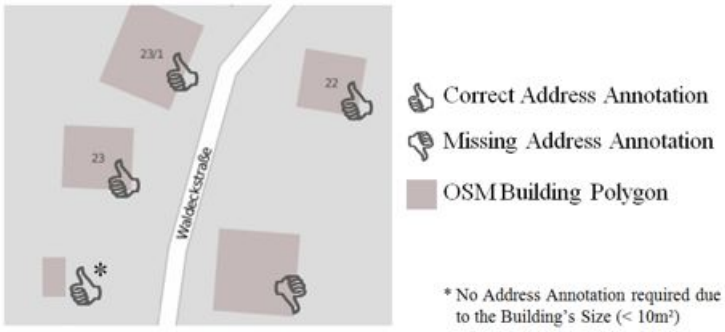


Figure 9.2.: Buildings which are likely to contain a house number/name (basemap: ©OpenStreetMap contributors).

Subsequently the development of the ratio between all buildings and those actually containing a number or name can be taken as an intrinsic indicator about the data's attribute completeness. Ideally, the cumulated number of house numbers and house names always corresponds to number of buildings, even if the number of buildings increases significantly (e.g. due to an import, a mapping party or better aerial images). The following list is a summary of the relevant parameters for geocoding:

- Buildings which should contain a house number/name
- Complete address annotation
- Overall distribution of house numbers/names

9.3.3.3. Points of interest-search

Points of Interests (POI) are important locations such as, for example, sights, restaurants or bus stops. In OSM, these are geographically represented by Nodes, Ways or Relations tagged with specific key-value pairs and loom large in several LBS applications. In this investigation all POIs are divided into the following nine thematic groups: accommodation and gastronomy, education, transport, finance, health care, art and culture, shop, tourism and others. Within these groups the quantitative development of all appropriate POIs can act as a first quantitative indicator. In general, an increasing number of POIs is a positive indicator, as the dataset is nearing completion. Beside the actual number of POIs, their detailed characterization by means of attributes has an impact on the feature's quality. In this investigation the authors hypothesize that a growing number of tags listed in the OSM map features in general increases the quality of the POIs because their characteristic values approximate reality more closely. This is especially true if the overall number of POIs is also increasing. However, it has to be stated that the development of the average number of tags can only act as an indicator because some OSM editors automatically assign tags to edited features which have to be filtered out.

Besides the quantitative development and the average number of attributes, the substantive differentiation within the POIs' attributes allows statements on the relative attribute completeness and therefore on the relative thematic accuracy. For this purpose a list of relevant keys selected from the OSM wiki is suitable (Kessler and Groot 2013) which describe the features of the aforementioned nine groups in more detail. The list consisting of the keys (e.g. name, opening_hours, operator, website, addr:housenumber, phone, wheelchair) is adapted to the individual case, as not all POI groups necessarily need to be annotated with all of them. The development of their percentage of relative completeness is an intrinsic measurement of attribute completeness, indicating how well a respective group of POIs is suited to specified use-cases in LBSs. An advantage of this procedure is that meaningful results can be achieved even if the POIs are not completely mapped. Using predefined lists of attributes which characterize qualitative completely attributed POIs is a promising approach to assessing attribute completeness without using a reference dataset for comparison purposes.

The following list contains a summary of the relevant parameters:

- Attributive Completeness of the POIs
- Average number of the POIs' tags
- Quantitative development of POIs

9.3.3.4. Map-applications

Beside the aforementioned use, OSM data is also widely used in map-applications. Earth surface characteristics within OSM are mainly represented by means of polygons, for instance tagged with a natural (e.g. glacier, wood or wetland) or a landuse (e.g. forest, residential area or vineyard) key. The accuracy of their geometric representation highly depends on the source (GPS traces, bulk imports or aerial images) and the acquisition scale of the contributed data (Mueller et al. 1995). A good way to determine the quality of these polygons is to calculate the equidistance between the polygons' adjacent vertices (Mooney et al. 2010a). In the proposed framework this approach is extended by the polygons' history. Comparing an initially created polygon with its currently valid one, the evolution of the equidistance serves as an intrinsic indicator for the relative quality development. The lower the equidistance of the currently valid polygon compared with its initially created version, the better the polygon's relative quality development, due to further editing which potentially led to a more precise geometric representation. However, several facts have to be considered: the algorithm of the tool (OpenStreetMap 2013d), that is used to split OSM data into smaller extracts, has an effect on polygons lying on the boundary of the selected bounding box (especially if polygons where moved there during their history). Using the hardcut algorithm, polygons are cropped at their last Node located within the bounding box. Furthermore, divided or merged polygons can also lead to biases within the calculated equidistance because of their significantly increased or decreased area. To exclude these outliers it was chosen iteratively to consider only polygons which do not differ in size between the two compared versions by more than 50%.

The intra-theme consistency as a part of the parameter logical consistency has a major influence on the quality of a spatial dataset. This is depicted by means of erroneously overlapping land use polygons. Within OSM, polygons attributed with a "landuse" tag represent the primary use of an area. Basically, these polygons should not overlap each other to avoid inconsistencies possibly leading to slivers, among other reasons. These overlaps are mainly caused by inaccurate digitizing or data imports, because in each case spatial integrity of the contributions is not necessarily examined

(Girres and Touya 2010). Nevertheless, sometimes a manifold land use of an area makes sense (e. g. militarily used forests). To take this fact into consideration, only overlaps with a size of less than 10% of the origin polygons are taken into account, because they more probably represent unintended overlaps. The lower the number of these detected cases, the better the intra-theme consistency concerning the land use polygons within the dataset. The following list summarizes the relevant parameters in this section:

- Erroneously overlapping land use polygons
- Evolution of the natural features' equidistance

9.4. Experimental analyses and results

In this section the results of the selected intrinsic quality indicators are outlined. For this purpose the cities of San Francisco (USA), Madrid (Spain), and Yaoundé (Cameroon) have been chosen. San Francisco is characterized by several bulk imports and a moderate-sized community. Representing a European metropolis, Madrid, as a counterpart to a US city equal in size, was chosen due to its moderate community activity without bigger imports. In contrast, the city of Yaoundé is a good example of a bulk import with no active mapping community. From all of the aforementioned indicators of the framework the following four are illustrated: road network completeness, the dataset's positional accuracy, house number completeness and the geometric representation of natural polygons.

9.4.1. Road network completeness

Figure 9.3 shows the total road network length for the selected cities. The results clearly illustrate differences concerning their possible completeness using the above-mentioned intrinsic indicators (cf. Section 3.4.1): San Francisco shows stable lengths for motorways/highways from May 2011 until today (491 km) whereas the length of secondary/tertiary roads (480 km) and residential roads (1,790 km) does not increase significantly from April 2012.

Except for the category "other roads" the road network therefore can be referred to as possibly close to completion. The strong increase of residential roads in October 2007 is accounted for by the TIGER/Line import. Madrid shows a similar pattern for secondary/tertiary roads which remain stable in length (814 km) from August 2012.

All other road categories are still being mapped, although with varying intensity. Motorways/highways show an increase of approximately 10 km within the last few months whereas the categories residential roads and other roads show a much higher average amount of contribution. Minor changes in length are not necessarily new roads but can also be caused by changing the value of the highway key. In contrast, the diagram of Yaoundé reveals a stepped contribution with longer periods of no contribution at all. This suggests a hardly active community and possible data imports. Taking the small amount of active contributors into account (see Figure 9.4) in the case of Yaoundé, no statements on the road network completeness are possible without using a reference dataset.

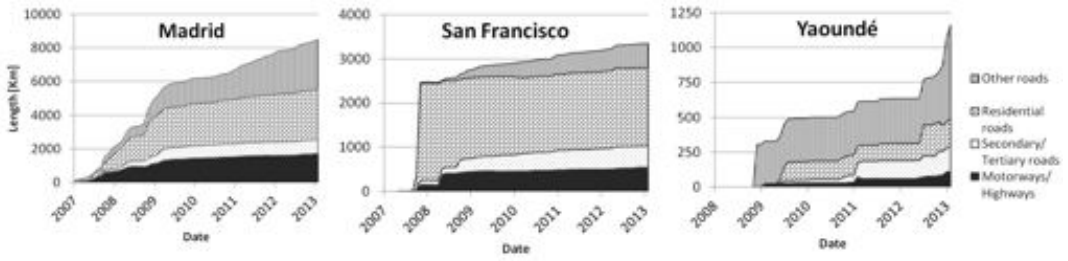


Figure 9.3.: Development of the OSM road network length by street category for the cities of Madrid, San Francisco and Yaoundé.

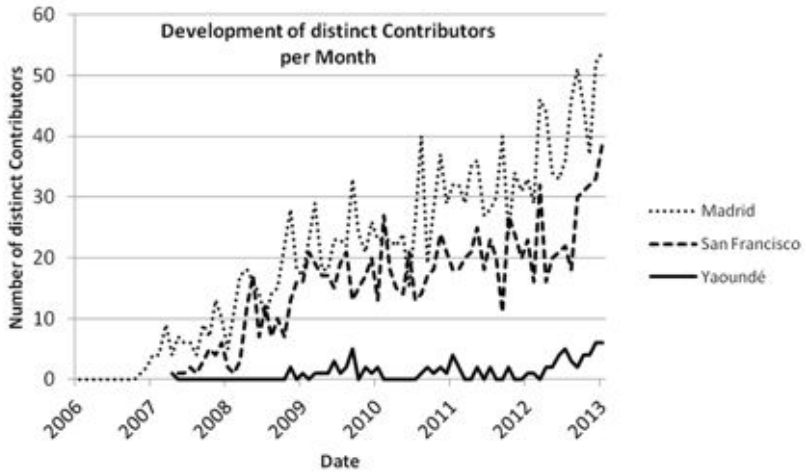


Figure 9.4.: Number of distinct contributors per month for the cities of Madrid, San Francisco and Yaoundé.

9.4.2. Positional accuracy of the dataset

As described in Section 9.3.2, comparing actual road junctions with the previous location before their last modification gives insights into possible positional inaccuracies, for instance triggered by displaced aerial images or bulk movements. Figure 9.5 shows a uniform distribution of points around the centre of Madrid and San Francisco. However, the city of Madrid shows some special characteristics: in five cases two road junctions show exactly the same distance and angle to each other, indicating a possible bulk movement of the OSM data. It has to be mentioned that within this method only junctions are selected which are clearly identifiable by means of their adjacent road names. Therefore it is possible that not only the identified roads, but also other roads or even other features situated nearby could be affected by a possible bulk movement. Furthermore, within Madrid's polar scatter plot a vast amount of points has been shifted by up to 15 m between 240° and 300°. This can mean either improvement or deterioration of the positional accuracy. A precise distinction can only be made if changesets, source tags or underlying aerial images are further investigated. However, this gives a hint where further analyses are needed, potentially carried out with a ground truth reference dataset.

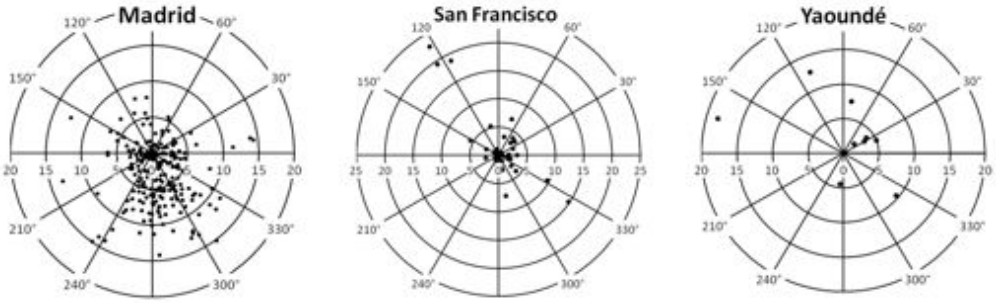


Figure 9.5.: Polar scatter plot of degree and distance between currently visible road junctions and their previous location for the cities of Madrid, San Francisco and Yaoundé.

9.4.3. Buildings with a house number/name

In Figure 9.6, the cities of Madrid and Yaoundé both show a significant increase of new buildings within one month, which, due to their vast amount, is probably caused by bulk imports. This specifically applies to the city of Yaoundé where an average of 1.2 users (max: 7 users; min: 0 users) are active per month and 123,204 building polygons

were imported in November 2012. In total, only four buildings are annotated with a house number/name indicating very low attribute completeness. In contrast, 1,146 (10.2%) of all buildings within San Francisco are tagged with a house number/name, whereas Madrid takes up a middle position with 1,024 (4.0%) tagged buildings. Nevertheless, in terms of attribute completeness all exemplarily investigated cities show a relatively low number of house numbers/names. Furthermore, in each of these three cases the number of created buildings does not increase proportionally.

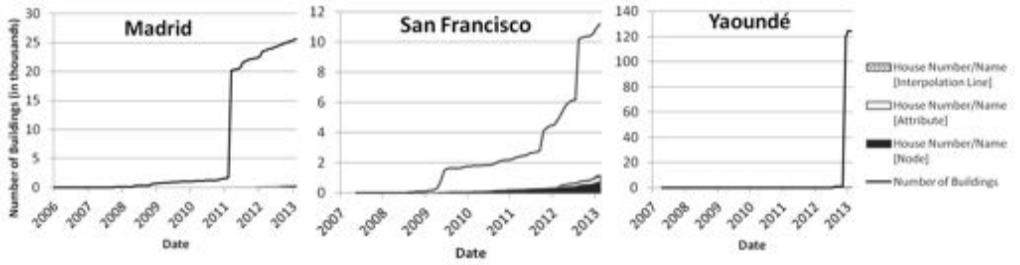


Figure 9.6.: Development of buildings (with a house number/name) for the cities of Madrid, San Francisco and Yaoundé.

9.4.4. Development of natural polygons' geometrical representation

Figure 9.7 shows the development of the equidistance of polygons tagged with a natural or landuse tag as described in Section 3.4.4. They are sorted by the equidistance in their first version in ascending order.

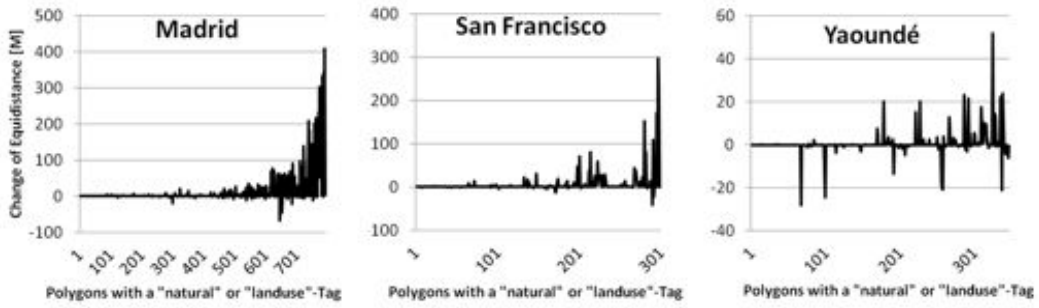


Figure 9.7.: Equidistance development of polygons tagged with a natural or landuse tag for the cities of Madrid, San Francisco and Yaoundé.

Examining the equidistance's development of the three selected cities several facts can be determined: with an improvement of the equidistance by an average of 11.9 m

Madrid shows the highest increase (San Francisco: 6.1 m; Yaoundé: 0.5 m). Furthermore, the geometric representation of 29.5% of all investigated polygons were improved (San Francisco: 25.6%; Yaoundé: 20.1%). As Figure 9.7 indicates within the city of Madrid in particular polygons with a high equidistance have been improved significantly during their history. However, the majority of the polygons' geometry in all three cities has not been changed (Madrid: 62.7%; San Francisco: 67.4%; Yaoundé: 62.9%).

9.4.5. Architecture framework

Figure 9.8 illustrates the entire architecture and workflow of the developed framework. The *iOSMAnalyzer* is implemented as a command line-based tool running on the Linux operating system. It is written in the Python programming language and based solely on open source components.

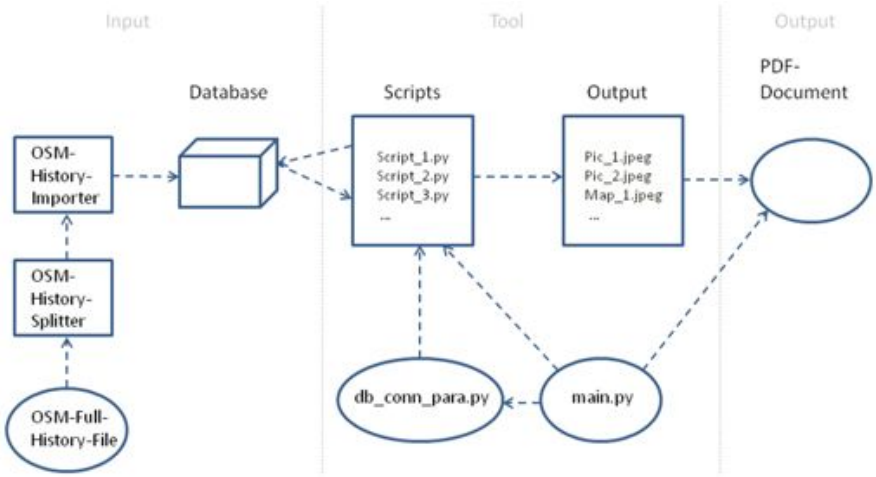


Figure 9.8.: Architecture of the *iOSMAnalyzer* framework.

As carefully evaluated beforehand, cropping features with the softcut algorithm (OSM-History-Splitter) leads to distorted statistics in some cases, especially when a single version of a feature's history lies mainly beyond the chosen bounding box. With the help of the OSM-History-Importer (OpenStreetMap 2013c) the clipped data is imported to a PostgreSQL/PostGIS database. Resting upon this database several previously developed Python scripts compute the results from the data which are plotted to a PDF file. This file contains diagrams, tables, results of statistical analyses and maps expressing and visualizing several computed intrinsic quality indicators. Finally,

it has to be mentioned that at the time of this research the OSM-History-Importer did not represent deleted Way features as such within the database. This potentially can lead to minor biases in some analyses.

9.5. Conclusions and future work

In this investigation a framework containing a broad range of more than 25 different methods and indicators is presented to evaluate the quality of an OSM dataset based on an OSM-Full-History-Dump. The holistic and thorough intrinsic approach carried out in this investigation allows data quality evaluations without a ground truth reference dataset. This is beneficial in many respects: accessibility to high quality and commercial datasets is often limited due to high costs and contradictory licensing restrictions. These facts allow OSM quality analyses without any regional or financial limitations. The calculated results are provided in the form of statistics, tables, diagrams and maps and give a compact quality overview of a freely selectable area. As quality heavily depends on the individual use case, the OSM data is evaluated in terms of "Fitness for Purpose" for LBSs concerning the categories "General Information on the Study Area", "Routing and Navigation", "Geocoding", "Points of Interest-Search", "Map Applications" and "User Information and Behavior". However, absolute statements on data quality are only possible with a high quality reference dataset as a basis for comparison. Nevertheless, in an intrinsic approach quality parameters, as for example the road network completeness can only be determined approximately; in this example by investigating the historical increase of mapped roads within different road categories. Furthermore, the contributor activity also has an effect on OSM data. This investigation revealed that the interpretation of some intrinsic quality indicators is facilitated and supported by means of contributor activity.

For future work Relations (e.g. turn restrictions, bus routes, etc.) should be taken into consideration. Currently, the OSM-History-Importer does not support the import of Relations to the database, therefore these were not considered in this research. Furthermore, as the quality of OSM data also depends on the project's contributors, more in-depth analyses regarding their experience, quality of contributions and reputation have to be integrated into the framework. Moreover some of the proposed methods could be evaluated by means of a ground truth reference dataset. The higher the conformity of the intrinsic results within this comparison, the better the proposed indicator is possibly suited. Additionally, a pre-calculated signature database containing different patterns of quality manifestations could serve as a reference for other OSM

areas with similar characteristics (e.g. community activity, size or spatial structure). Due to its modular structure, the implemented framework can easily be extended by further methods and indicators.

References

Amelunxen, C. (2010). An Approach to Geocoding Based on Volunteered Spatial Data. In: *Geoinformatik 2010*. (Mar. 17–17, 2010). Kiel, Germany.

Batini, C. and Scannapieco, M. (2006). *Sharing Imperfect Data*. Berlin, Germany: Springer.

Bing (2010). *Bing Engages Open Maps Community - Bing Maps Blog*. URL: `http://www.bing.com/community/site_blogs/b/maps/archive/2010/11/23/bing-engages-open-maps-community.aspx` (visited on 08/11/2012).

Brando, C. and Bucher, B. (2010). Quality in User-Generated Spatial Content: A Matter of Specifications. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

Budhathoki, N. (2010). Participants' Motivations to Contribute to Geographic Information in an Online Community. Ph.D. Dissertation. University of Illinois, Urbana-Champaign, Urbana, IL, USA.

Chilton, S. (2012). Crowdsourcing Is Radically Changing the Geodata Landscape: Case Study of OpenStreetMap. In: *Proceedings of the UK 24th International Cartography Conference*. (Nov. 15–21, 2009). Santiago, Chile.

Ciepluch, B., Mooney, P., and Winstanley, A. (2011). Building Generic Quality Indicators for OpenStreetMap. In: *Proceedings of the Nineteenth Annual GISRUK Conference*. (Apr. 27–29, 2011). Portsmouth, England.

Coleman, D., Georgiadou, Y., and Labonte, Y. (2009). Volunteered Geographic Information: The Nature and Motivation of Producers. *International Journal of Spatial Data Infrastructures Research*, 4, 332–358.

Devillers, R., Bédard, Y., Jeansoulin, R., and Moulin, B. (2007). Towards Spatial Data quality Information Analysis Tools for Experts Assessing the Fitness for Use of Spatial Data. *International Journal of Geographical Information Science*, 21, 261–w82.

Devillers, R., Gervais, M., Bédard, Y., and Jeansoulin, R. (2002). Spatial Data Quality: From Metadata to Quality Indicators and Contextual End-user Manual. In: *Proceedings of the OEEPE/ISPRS Joint Workshop on Spatial Data Quality Management*. (Mar. 21–22, 2002). Istanbul, Turkey.

Exel, M. van, Dias, E., and Fruijtier, S. (2010). The Impact of Crowdsourcing on Spatial Data Quality Indicators. In: *Proceedings of the Sixth International Conference on Geographic Information Science (GIScience 2010) Workshop on the Role of Volunteered Geographic Information in Advancing Science*. (Sept. 14–17, 2010). Zurich, Switzerland.

Flanagin, A. J. and Metzger, M. J. (2008). The Credibility of Volunteered Geographic Information. *GeoJournal*, 72, 137–148.

Geofabrik (2013). *OpenStreetMap Data Extracts*. URL: `http://download.geofabrik.de/`.

Girres, J.F. (2011). A Model to Estimate Length Measurements Uncertainty in Vector Databases. In: *Proceedings of the Seventh International Symposium on Spatial Data Quality (ISSDQ'11)*. (Oct. 12–14, 2011). Coimbra, Portugal.

Girres, J.F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS*, 14, 435–459.

Goetz, M. and Zipf, A. (2013). The Evolution of Geo-crowdsourcing: Bringing Volunteered Geographic Information to the Third Dimension. In: *Crowdsourcing Geographic Knowledge*. Berlin, Germany: Springerpp. 9–59.

Goodchild, M.F. (1995). Sharing Imperfect Data. In: *Sharing Geographic Information*. New Brunswick, NJ, USA: JRutgers University Press413–425.

Goodchild, M.F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211–221.

Goodchild, M.F. (2009). NeoGeography and the Nature of Geographic Expertise. *Journal of Location Based Services*, 3, 82–96.

Goodchild, M.F. and Li, L. (2012). Assuring the Quality of Volunteered Geographic Information. *Spatial Statistics*, 1, 110–120.

Hagenauer, J. and Helbich, M. (2012). Mining Urban Land Use Patterns from Volunteered Geographic Information by Means of Genetic Algorithms and Artificial Neural Networks. *International Journal of Geographical Information Science*, 26, 963–982.

Haklay, M. (2010). How good is OpenStreetMap Information: A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. *Environment and Planning B*, 37, 682–703.

Haklay, M., Basiouka, S., Antoniou, V., and Ather, A. (2010). How many Volunteers does it take to Map an Area well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47, 315–322.

Helbich, M., Amelunxen, C., Neis, P., and Zipf, A. (2012). Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. In: *Proceed-*

ings of GI Forum 2012: Geovisualization, Society and Learning. (July 4–6, 2012). Salzburg, Austria.

Hristova, D., Quattrone, G., Mashhadi, A., and Capra, L. (2013). The Life of the Party: Impact of Social Mapping in OpenStreetMap. In: *Proceedings of the 7th International Conference on Weblogs and Social Media*. (July 8–11, 2013). Cambridge, Massachusetts, USA.

Kessler, C. and Groot, R.T.A. de (2013). Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap.

Kessler, C., Trame, J., and Kauppinen, T. (2011). Tracking Editing Processes in Volunteered Geographic Information: The Case of OpenStreetMap. In: *Proceedings of the COSIT'11 Workshop: Identifying Objects, Processes and Events in Spatio-Temporally Distributed Data*. (Sept. 12–16, 2011). Belfast, Maine, USA.

Kounady, O. (2011). Assessing the Quality of OpenStreetMap Data. M.Sc. Thesis. Department of Civil, Environmental and Geomatic Engineering, University College of London, UK.

Leeuw, J. de, Said, M., Ortegah, L., Nagda, S., Georgiadou, Y., and DeBlois, M. (2011). An Assessment of the Accuracy of Volunteered Road Map Production in Western Kenya. *Remote Sensing*, 3, 247–256.

Lin, Y. (2011). A Qualitative Enquiry into OpenStreetMap Making. *New Review of Hypermedia and Multimedia*, 17, 53–71.

Ludwig, I., Voss, A., and Krause-Traudes, M. (2011). A Comparison of the Street Networks of Navteq and OSM in Germany. *Advancing Geoinformation Science for a Changing World*, 1, 65–84.

Mondzech, J. and Sester, M. (2011). Quality Analysis of OpenStreetMap Data based on Application need. *Cartographica*, 46, 115–125.

Mooney, P. and Corcoran, P. (2012). Characteristics of heavily edited Objects in OpenStreetMap. *Future Internet*, 4, 285–305.

Mooney, P., Corcoran, P., and Winstanley, A. (2010a). A Study of Data Representation of Natural Features in OpenStreetMap. In: *Proceedings of the 6th GIScience International Conference on Geographic Information Science*. (Sept. 14–17, 2010). Zurich, Switzerland.

Mooney, P., Corcoran, P., and Winstanley, A. (2010b). Towards Quality Metrics for OpenStreetMap. In: *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. (Nov. 2–5, 2010). San Jose, CA, USA.

Mueller, J.C., Lagrange, J.P., and Weibel, R. (1995). London, UK: Taylor and Francis.

Napolitano, M. and Mooney, P. (2012). *MVP OSM: A Tool to Identify Areas of High Quality Contributor Activity in OpenStreetMap*. URL: https://github.com/napo/mvp-osm.

Neis, P., Zielstra, D., and Zipf, A. (2012). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4, 1–21.

Neis, P., Zielstra, D., and Zipf, A. (2013). Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Future Internet*, 5, 282–300.

Neis, P. and Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project - The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 146–165.

Nielsen, J. (2006). *Participation Inequality: Encouraging More Users to Contribute*. URL: http://www.nngroup.com/articles/participation-inequality/ (visited on 09/23/2013).

OpenStreetMap (2012). *Import/Catalogue - OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Import/Catalogue (visited on 08/20/2012).

OpenStreetMap (2013a). *Karlsruhe Schema - OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Proposed_features/House_numbers/Karlsruhe_Schema.

OpenStreetMap (2013b). *Map Features - OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Map_Features.

OpenStreetMap (2013c). *MaZderMind/OSM-History-Renderer*. URL: https://github.com/MaZderMind/osm-history-renderer/tree/master/importer.

OpenStreetMap (2013d). *MaZderMind/OSM-History-Splitter*. URL: https://github.com/MaZderMind/osm-history-splitter.

OpenStreetMap (2013e). *Osmium - OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Osmium.

OpenStreetMap (2013f). *Osmosis - OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Osmosis.

OpenStreetMap (2013g). *Planet.osm/full - OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Planet.osm/full.

OpenStreetMap (2013h). *Registered Users - OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Stats#Registered_users.

O'Reilly, T. (2005). *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. URL: http://oreilly.com/web2/archive/what-is-web-20.html (visited on 09/21/2013).

Ramm, F., Topf, J., and Chilton, S. (2010). Cambridge, UK: UIT.

Rehrl, K., Groechenig, S., Hochmair, H., Leitinger, S., Steinmann, R., and Wagner, A. (2013). A Conceptual Model for Analyzing Contribution Patterns in the Context of VGI. In: *Progress in Location-Based Services*. Berlin, Germany: Springer373–388.

Touya, G. and Brando-Escobar, C. (2013). Detecting Level-of-Detail Inconsistencies in Volunteered Geographic Information Data Sets. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 48, 134–143.

Veregin, H. (1999). Data Quality Parameters. In: *Geographical Information Systems: Principles and Technical Issues*. New York, USA: John Wiley and Son177–189.

Wang, R. and Strong, D. (1996). Beyond Accuracy: What Data Quality means to Data Consumers. *Journal of Management Information Systems*, 12, 5–33.

Zielstra, D. and Hochmair, H.H. (2012). Comparison of Shortest Path Lengths for Pedestrian Routing in Street Networks Using Free and Proprietary Data. In: *Proceedings of the Transportation Research Board - 91st Annual Meeting*. (Jan. 22–26, 2012). Washington, DC, USA.

Zielstra, D., Hochmair, H.H., and Neis, P. (2013). Assessing the Effect of Data Imports on the Completeness of OpenStreetMap – A United States Case Study. *Transaction in GIS*, 17, 315–334.

Zielstra, D. and Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

10. Generation of a Tailored Routing Network for Disabled People Based on Collaboratively Collected Geodata

Authors

Pascal Neis and Dennis Zielstra

Journal

Applied Geography

Status

Published: 27 December 2013 / Accepted: 5 December 2013 / Revised: 4 December 2013 / Submitted: 6 October 2013

Reference

Neis, P. and Zielstra, D. (2014). Generation of a tailored routing network for disabled people based on collaboratively collected geodata. *Applied Geography*, 47:70–77.

Contribution statement

Pascal Neis developed the methods and applications for this study and wrote the majority of the manuscript. Dennis Zielstra supported this publication through continuing discussions about the methods and the results of the study. Furthermore, extensive proof-reading by the co-author led to substantial improvements to the manuscript.

.....	.....
Prof. Dr. Alexander Zipf	Dennis Zielstra

Abstract

The generation of a routing network for disabled people inherits a number of prerequisites that need special consideration. Widespread routing applications that rely on commercial or governmental geodata sources are not feasible for this specific task, due to the lack of detailed information about features such as sidewalks, surface conditions or road incline. In recent years the research community has experienced a strong increase in studies related to routing applications tailored to disabled people in which the lack of a sophisticated dataset played a major role. This study proposes an algorithm for the generation of a disabled people friendly routing network, based on collaboratively collected geodata provided by the *OpenStreetMap* (OSM) project. This new representation of a routing graph can be used in numerous applications and maps dedicated to people with disabilities. The algorithm is tested and evaluated for selected areas in Europe, resulting in newly generated extended networks that include sidewalk information. The results have shown that the success of the final implementation of the introduced algorithm depends highly on the attribute quality of the OSM dataset.

Keywords: Route generation; Routing graph; Disabled people; Collaborative mapping.

10.1. Introduction

Routing and navigation applications on the Internet, in cars or on personal smartphones are omnipresent. Most common devices and applications rely on geodata provided by one of the well-known proprietary data providers such as Navteq™or TomTom™. These providers offer routing network data which is suitable for motorized and (for selected cities) non-motorized path finding applications. People with special needs, however, who rely on a more specialized dataset, cannot utilize the provided commercial geo-information and require highly detailed ground-truth data. Commercial geodata providers do not offer this detailed information due to the high costs that arise during the collection and the maintenance of the data.

In the past few years the number of freely available and open source geo-information platforms on the Internet has increased tremendously. These new data sources are oftentimes referred to as *Volunteered Geographic Information* (VGI; Goodchild 2007). As the name implies, most of these platforms rely on the contributions of non-professional volunteers that collaboratively collect geodata. A number of possible motivational factors that trigger VGI project contributions have been identified in a recent study, including the desire to make geospatial information freely available to everyone, learn-

ing new technologies, relaxation and recreation, self-expression or just pure fun (Budhathoki and Haythornthwaite 2013). The contribution patterns found in VGI projects tend to be more casual in comparison to the contributions made to *Public Participation Geographic Information Systems* (PPGIS) in which volunteers collect geodata for a particular purpose, such as to improve land use planning or discuss policy issues and decision making (Brown 2012). One of the biggest and most established projects in the realm of VGI is *OpenStreetMap* (OSM[1]). In contrast to the aforementioned proprietary data providers, the OSM project data is distributed under an *Open Data Commons Open Database License* (ODbL[2]). This particular license allows interested Internet users to download, copy, distribute, transmit and adapt the collected geodata, free of charge, as long as OSM and its contributors are credited in the final project.

Despite early concerns about the credibility and reliability of VGI (Flanagin and Metzger 2008) several studies demonstrated the potential of OSM in a variety of applications in recent years. OSM data has been utilized to develop a number of *Location Based Services* (LBS; Neis and Zipf 2008), to evaluate the urban accessibility in the aftermath of an earthquake (Bono and Gutiérrez 2011) and to simulate future urban growth patterns in Mumbai (India) (Moghadam and Helbich 2013). At the time of writing, the project had more than 1.4 million registered members who contributed with varying intensity to the project. In a number of major cities the volunteers collect information about sidewalks, road surfaces, road incline, pedestrian crossings, and tactile paving[3]. This level of detail is essential when considering the creation of a suitable routing graph for disabled people, such as wheelchair users or elderly people.

The terminology used to describe the target user group for the developed algorithm can vary and will be discussed in more detail in the section 2 of this paper. However, the main research question of this study is: How can freely available, collaboratively mapped geodata be utilized to generate a routing network for disabled people with special navigation information needs? The benefit and advantage of this newly generated routing network lies in its multipurpose character. This allows the network to be used in route-planning, real-time navigation or for both online and print maps, by providing detailed information about the "best" individual route based on the user's limitations. The open approach to data collection efforts in OSM lead to high object densities and details in selected urban areas, at times illustrating barriers for disabled people. For areas that do not provide this level of detail, the map can easily be edited

[1] http://www.openstreetmap.org (visited on 5 October 2013)
[2] http://opendatacommons.org/licenses/odbl/ (visited on 27 November 2013)
[3] http://www.blind.accessiblemaps.org/index2.html (visited on 5 October 2013)

to serve the individual purpose.

The remainder of this article is structured as follows: Section 2 presents some background information and related research in the field of routing networks and wayfinding for disabled people. Section 2 also contains detailed information about the requirements and parameters that the generated network should inherit and the routing algorithm should take into account when computing a route. Additionally, the OSM project and research related to the project will be briefly introduced. In section 3, the methodology including data preparation and the generation of the tailored routing network is described. Section 4 includes the evaluation of the presented algorithm by testing the generated sidewalk networks for selected areas in Europe. The article concludes with a discussion of potential algorithm limitations, a summary of the findings and an outlook on future research.

10.2. Background and related work

Routing applications on mobile devices and desktop computers are oftentimes used when planning a trip or during a visit of an unfamiliar place such as a new city. While the local knowledge of an individual helps to find the shortest or fastest path in familiar places on a day to day basis, routing applications can help to experience a similar situation in unfamiliar areas. Disabled people rely on very detailed information about potential obstacles in their neighborhood or in areas in which their daily life takes place. However, when visiting unknown places regular routing applications tailored to motorized traffic or pedestrians do not provide the detailed information needed. Depending on the requirements of the user, information about sidewalks, steps, surface conditions, crossings or tactile paving could be essential and heavily improve the routing experience of a disabled person.

Research that focuses on routing specifications and applications for disabled people, such as wheelchair users, blind, deaf or elderly people, has experienced a strong increase in recent years (Sobek and Miller 2006, Kasemsuppakorn and Karimi 2009, Kammoun et al. 2010). The most important finding that needs to be considered in any related analysis is that geodata requirements vary significantly depending on the project's purpose. Routing applications for non-motorized traffic, such as pedestrians, have different geodata requirements than applications tailored to motorized traffic and vice versa (Corona and Winter 2001, Walter et al. 2006). Similarly, patterns between geodata implemented in these widely used applications and the geodata requirements for applications tailored to disabled people need to be evaluated.

10.2.1. Routing network requirements for disabled people

Several studies in the past have highlighted the prerequisites that the geodata source of choice has to fulfil to be considered for a potential navigation system for pedestrians (Gaisbauer and Frank 2008), wheelchair users (Charles et al. 2002, Kasemsuppakorn and Karimi 2009) or blind people (Kammoun et al. 2010). Oftentimes the customized system and its corresponding data are created through extensive surveys. A specification by the German Institute for Standardization (*Deutsches Institut für Normung* (DIN)) provides a foundation for this particular type of information. DIN 18024-1 describes the accessibility requirements for disabled people in public transit infrastructure and buildings. The standards include a number of recommendations for different handicap types, which also help to define the target user group for which our study was conducted (Source: DIN 18024-1):

- Wheelchair users
- Blind and visually impaired people
- Deaf and hearing impaired people
- Walking impaired people
- People with other handicaps
- Elderly people
- Children and people of short or tall statue

Based on the specification, some of recommended parameters that need to be implemented in the final dataset can be surface information, incline and width of a street segment. However, based on a number of different studies, other parameters for a disabled friendly routing network have been determined (Matthews et al. 2003, Beale et al. 2006, Sobek and Miller 2006, Ding et al. 2007, Kasemsuppakorn and Karimi 2009, Menkens et al. 2011). Table 10.1 summarizes all parameters based on the findings of the studies, the DIN 18024-1 and some newly defined parameters based on our research. In some of the studies the desired geodata was traced from satellite imagery (Kasemsuppakorn and Karimi 2008, Kasemsuppakorn and Karimi 2009), while others developed tools that generated a network by utilizing a buffer method (Karimi and Kasemsuppakorn 2012), implementing pedestrian GPS traces (Kasemsuppakorn and Karimi 2013), developing a binary image processing method to retrieve a pedestrian network (Gaisbauer and Frank 2008, Kim et al. 2009) or presented an automated method to generate a sidewalk network from building blocks (Ballester et al. 2011).

Table 10.1.: Summary of required parameters for the generation of a routing network for disabled people.

Parameter	Description	Reference
Type of street	Ways which can be used for a routing network for disabled people	[8]
Sidewalk	Has the street a sidewalk, and if yes on which side?	[1,2,3,4,5,6,7,8]
(Sidewalk) Width	Width of the street/sidewalk	[1,2,3,4,5,6,7,8]
(Sidewalk) Surface	Surface of the street/sidewalk	[1,2,4,5,6,7,8]
(Sidewalk) Smoothness	Smoothness of the street/sidewalk	[1,4,6,7,8]
(Sidewalk) Slope/Incline	Incline of the street/sidewalk	[1,2,4,5,6,7,8]
(Sidewalk) Camber	Camber of the street/sidewalk	[1,2,4,7]
(Sidewalk) Curb/Kerb	Sloped curb (height)	[1,2,3,4,6,7,8]
(Sidewalk) Curvature	Curvature of the street/sidewalk	[2,7]
Lighting	Is the street lighted?	[4,7,8]
Tactile Paving	Is tactile paving available?	[7,8]
Steps	Number of steps	[1,2,3,5,6,7,8]
Step height	Height of the individual steps	[3,7]
Ramp	Is a ramp (at the steps) available?	[1,2,3,6,7,8]
Handrail	Is a handrail railing (at the steps/ramp) available?	[7]
Crossing	Crossing (with/without traffic signals)	[1,2,7,8]
General Access	General access information of the street/sidewalk	[3,8]

Notes: [1]Matthews et al. (2003) [2]Beale et al. (2006) [3]Sobek and Miller (2006) [4]Ding et al. (2007) [5]Kasemsuppakorn and Karimi (2009) [6]Menkens et al. (2011) [7]DIN 18024-1 [8]Our research

10.2.2. Collaboratively collected geodata: The OpenStreetMap project

User Generated Content (UGC; Anderson 2007) and particularly *Volunteered Geographic Information* (VGI; Goodchild 2007) have become a widely known Internet phenomenon in recent years. The OSM project, initiated in 2004, is the most successful VGI project based on collaboratively collected and freely available geodata (Mooney et al. 2010, Neis et al. 2012b, Goetz and Zipf 2012). Most contributors collect the geodata by utilizing GPS handhelds, such as smartphones or by tracing satellite imagery available to the project (e.g. Yahoo until 2011 or Microsoft Bing since 2010). Neis and Zipf (2012) have shown that the largest and most active community of the project is located in Germany and that almost three-quarters of the members who ever made a contribution to the project are from Europe. However, Neis and Zipf (2012) also proved that only a small number of OSM members have contributed at least one object to the database (almost 33% of all members). At the time of writing, less than 2% of all members actively collect information each month (OSMstats 2013), a pattern that

can be found in similar online community-based projects such as Wikipedia, defined as "Participation Inequality" (Nielsen 2006).

A wide range of recent studies have shown that for selected regions the collaboratively collected geodata of the OSM project can be an alternative to commercial or administrative datasets (Haklay 2010, Girres and Touya 2010, Zielstra and Hochmair 2011b, Neis et al. 2012b). Hagenauer and Helbich (2012) criticize that oftentimes the empirical studies in prior publications only consider objects of certain types (e.g. roads) for descriptive measurements. However, it was also stated that urban areas are better mapped than rural counter parts, a pattern that was also described as "urban bias" (Mooney et al. 2013), which means that the data concentration and quality correlates in most cases with the population density. Mooney et al. (2013) similarly denoted that differences in representation and coverage between urban areas can be found in OSM.

A comprehensive analysis by Neis et al. (2013) showed that urban areas in a worldwide comparison can differ in terms of data quality and number of active community members. These factors highly influence the fitness of the OSM dataset for different purposes. Each purpose and end-user application has different requirements to the dataset and needs to be treated and evaluated individually (Mondzech and Sester 2011, Mooney et al. 2013). First analyses by Zielstra and Hochmair 2011a, Zielstra and Hochmair 2012 and Neis et al. 2012b have shown that the OSM project provides a comprehensive network for pedestrians in comparison to commercial or governmental dataset distributors.

One of the main reasons for the development of more advanced applications such as Location Based Services or 3D applications based on VGI is the increased data collection efforts by the OSM community, which is not solely limited to streets, landuse information or buildings anymore. More details are being added to the map every day, including public transportation information, address-data such as house numbers, or detailed information that can be used for an adequate route planning application for people with disabilities. The Wheelmap[4] project is tailored for this particular purpose and allows volunteers to mark locations on a map which provide wheelchair friendly environments or accessibility. The information provided by the contributors is then saved to the OSM database. This project shows some of the advantages of collaboratively collected geodata. In contrast to other VGI projects, such as Google Map Maker or TomTom's Map Share, contributors can easily create and add new objects or features to the database while the entire geodata collection of the project is freely available.

[4]http://wheelmap.org (visited on 5 October 2013)

10.3. Methodology

The generation of the proposed routing network consists of two processing steps. Each individual step can be summarized as follows:

1. Data preparation (Section 10.3.1): In the first step a regular routing network based on the available OSM dataset is generated. It is important to evaluate in this step whether a street segment has additional parameters which are relevant to the generation of the final network (e.g. sidewalk or surface information).
2. Generation (Section 10.3.2): After the initial data preparation, the second step involves the creation of the disabled friendly routing network, utilizing all relevant information that was retrieved from original OSM dataset.

10.3.1. Data preparation

The OSM project has three different object types that allow the active contributors to map features of the real world (Ramm et al. 2010). A Node object represents a point feature with its latitude and longitude coordinates, whereas a Way object is utilized to represent streets or closed line areas (i.e. polygons) such as landuse information or buildings. The Relation object contains information on how two or more objects are related to each other (e.g. a bus or tram line of the public transportation network). Attribute information about objects are added by applying Tags consisting of a key-value pair. A comprehensive list of OSM key-value pairs for a large number of map features is available on one of the OSM related wiki pages (OpenStreetMap 2013a). However, it needs to be noted that this list does not represent a strict specification or standardization, which means that each contributor can assign keys or values based on her/his own understanding and preference. Girres and Touya (2010) and Brando and Bucher (2010) criticized this tagging procedure in OSM and suggested that the data quality can be improved by using predefined specifications for objects and their corresponding tags. Nevertheless, the current tagging implementation is an essential part of the open approach to data contributions in OSM (Neis et al. 2012a).

The default OSM dataset is not applicable for routing or navigation purposes. Schmitz et al. (2008) and Renz and Wölfl (2010) introduced different methods on how to generate a routing network based on OSM data. These initial concepts were implemented in the first processing step of the disabled friendly routing network generation. The creation of the routing graph is followed by the identification of the relevant OSM tags. Nearly all of the aforementioned special requirements for disabled people

(Section 10.2.1) are mapped in OSM in some way or another. The representation of sidewalks in the OSM database plays a major role in this particular case. A sidewalk is only mapped as a separate feature if the sidewalk is not in close proximity to the street (Ramm et al. 2010). In all other cases the information of the sidewalk is part of the street object, e.g. sidewalk:left:surface=good. There are multiple OSM values with different key combinations that can be utilized for our purpose. Table 10.2 matches the prerequisites of a disabled friendly routing network (Table 10.1) with the corresponding OSM Tags. Overall only two parameters shown in Table 10.1 cannot be found in the OSM mapping schema: the camber and curvature of a sidewalk.

Table 10.2.: Generated routing network parameters and corresponding OSM tags.

Parameter	OSM coding (key=value ; if several values possible, they are separated by a "|" or by a note)	Unit
Type of street	highway=living_street[1]	-
Sidewalk	footway=left|right|yes|no|both sidewalk=left|right|yes|no|both	-
Sidewalk width	sidewalk(:left|:right):width=*	[m]
Sidewalk surface	sidewalk(:left|:right):surface=paved[2]	-
Sidewalk smoothness	sidewalk(:left|:right):smoothness=good[3]	-
Sidewalk slope/incline	sidewalk(:left|:right):incline=*	[%]
Sidewalk curb/kerb	sidewalk(:left|:right):sloped_curb(:start|:end)=*	[m]
Lighting	lit=yes|no	-
Tactile paving	tactile_paving=yes	-
Steps	step_count=*	-
Step height	step:height=*[4]	[cm]
Ramp	highway=steps ramp=yes ramp:wheelchair=yes ramp:stroller=yes	-
Handrail	handrail(:left|:right|:center)=yes|no|left|right|both|center	-
Crossing	highway=crossing or footway=crossing crossing=traffic_signals|uncontrolled|island traffic_signals:sound=yes/no traffic_signals:vibration=yes/no supervised=yes|no	-
General access	foot=yes|no, wheelchair=yes|no	-

Notes: [1]Additional highway-values: primary*, primary_link*, secondary*, secondary_link*, tertiary*, tertiary_link*, unclassified*, living_street, pedestrian, residential, service, track, footway, cycleway, bridleway, steps (*only if accessible for pedestrians/wheelchairs) [2]Additional surface-values: paved, asphalt, concrete, paving_stones, cobblestone, concrete_plates [3]Additional smoothness-values: excellent, good, intermediate, bad, very_bad [4]Currently a proposed OSM tag

10.3.2. Generation

The generation of the sidewalk routing network consists of several geometric processes. Figure 10.1 illustrates the individual steps of the algorithm. In Step 1 junctions are created, which consist of three ways and one node. Each way has a sidewalk declaration in the OSM database. In Step 2 a temporary line running parallel to each way segment is generated for each side at which a sidewalk exists. The newly generated lines represent the temporary paths for pedestrians and wheelchair users. During the generation of these temporary paths the way type, documented in the OSM database, is taken into account too. For instance, the temporary line for a tertiary road will be created with a distance of 5m, while in the case of a residential road a distance of 3.5m will be applied. The distances are based on guidelines provided by the German "*Forschungsgesellschaft für Straßen- und Verkehrswesen* (FGSV)", which include detailed information about the construction of roads and other infrastructure. Furthermore each sidewalk parameter (e.g. surface or width) is transferred from the initial line to the temporarily generated sidewalk line. In Step 3 the final sidewalk geometries are connected to their corresponding junction node. If a connection between two sidewalks crosses a way of the initial OSM network, a crossing between the two sidewalks will be created (see Figure 10.1, Step 3). The last step (Step 4) removes all ways of the initial network that have a newly generated sidewalk representation. The final image in Figure 10.1 shows the routing network generated by the algorithm as an overlay on an OSM basemap.

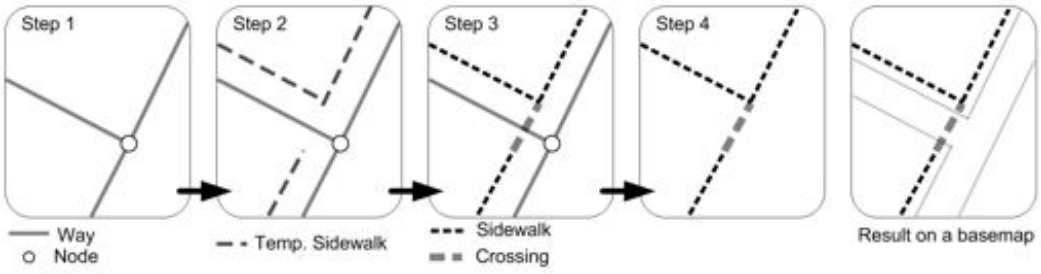


Figure 10.1.: Generation of routing network for disabled people.

10.4. Evaluation

The prototype of the algorithm was tested for all capital cities of the 50 sovereign European states. For each city a test region was extracted from the OSM dataset using a circular polygon with a radius of 2 km around each city center. The position

of the city center was determined by utilizing the geocoding tool of the Nominatim (OpenStreetMap 2013b) software. The OSM raw data was downloaded as a planet database dump file (OpenStreetMap 2012b). The clipping process was accomplished with the help of the OSMOSIS (OpenStreetMap 2012a) tool, followed by the generation of the sidewalk routing graph for each city. The comparison of the selected areas showed that the networks for 36 out of 50 cities have less than 1% of the required sidewalk parameter information to create a representative graph, whereas eleven cities have less than 10% of the required information (Table 10.3). Only the networks for the city centers of Berlin (Germany), London (United Kingdom) and Riga (Latvia) proved to have more than 30% of the required information and were selected for the following evaluation.

Table 10.3.: Percentage of sidewalk information included in OSM networks (OSM data date: July 13th, 2013).

Capital city (Country)	Percentage
Berlin (Germany)	61%
Riga (Latvia)	36%
London (United Kingdom)	34%
Athens (Greece), Belgrade (Serbia), Bern (Switzerland), Copenhagen (Denmark), Ljubljana (Slovenia), Luxembourg (Luxembourg), Podgorica (Montenegro), Sarajevo (Bosnia and Herzegovina), Tbilisi (Georgia), Vienna (Austria), Vilnius (Lithuania)	>1% and <10%
Amsterdam (Netherlands), Andorra la Vella (Andorra), Ankara (Turkey), Astana (Kazakhstan), Baku (Azerbaijan), Bratislava (Slovakei), Brussels (Belgium), Budapest (Hungary), Bucharest (Romania), Chisinău (Moldova), San Marino (San Marino), Dublin (Ireland), Helsinki (Finland), Kiew (Ukraine), Lisboa (Portugal), Madrid (Spain), Minsk (Belarus), Monaco (Monaco), Moscow (Russia), Nicosia (Cyprus), Osla (Norway), Paris (France), Prague (Czech Republic), Reykjavik (Iceland), Rome (Italy), Skopje (Republic of Macedonia), Sofia (Bulgaria), Stockholm (Sweden), Tallinn (Estonia), Tirana (Albania), Vaduz (Liechtenstein), Valetta (Malta), Vatican City (Vatican City), Warsaw (Poland), Yerevan (Armenia), Zagreb (Croatia)	<1%

The parsing, processing and generation of the sidewalk network, was implemented in JAVA programming language and took less than eight seconds for each city. Table 10.4 contains more information and general statistics for each of the three test areas. The values provided in the “Generated Sidewalk Network Length” column contain the total length of all features with at least one sidewalk Tag in the OSM dataset. If a

street has a sidewalk on both sides the length of the feature is only counted once.

Table 10.4.: Network lengths of tested areas.

	Berlin	Riga	London
Total network length	322 km	271 km	393 km
Network length for pedestrians	176 km	160 km	170 km
Network length which could contain sidewalk information	146 km	111 km	223 km
Generated sidewalk network length	89 km	40 km	76 km
Parsing, processing & creating network	<7 s	<5 s	<8 s
Errors during the processing (e.g. due to duplicate ways)	5	0	12
Warnings during the processing (e.g. due to crossing unconnected ways)	22	5	48

Figure 10.2 shows the individual ways (black lines) that are tagged with sidewalk information in the three test areas. The center of Berlin proved to have good sidewalk information coverage with a decline in information concentration when moving away from the center, especially in the Northeast and Southwest areas (Figure 10.2(a)). Most sidewalk information in Riga (Figure 10.2(b)) lies in one city district east of the Daugava River, whereas in London (Figure 10.2(c)) the majority of the required information is only distributed along the main roads.

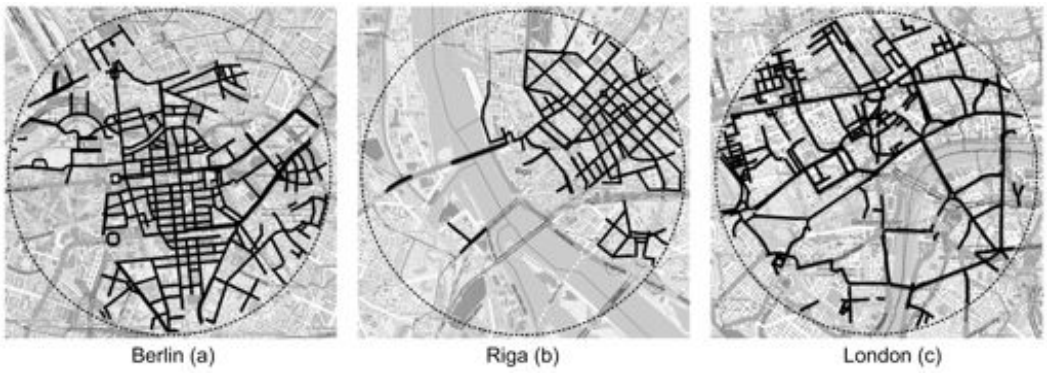


Figure 10.2.: Streets (black) that contain sidewalk information.

To evaluate the efficiency of the presented algorithm, 100 shortest paths between random start and end points in each test area were calculated. For comparison purposes two paths were generated for each city. The first path was computed on the regular street network graph, whereas the computation of the second path was based on the newly generated sidewalk graph (Table 10.5). Next to the total length comparison between both paths, indicating potential detours due to errors of omission or

commission, a buffer comparison method introduced by Goodchild and Hunter (1997) was applied to test if the computed route geometries of the sidewalk graph differ from the routes of the regular street network. A buffer of 10 meters on each side of the generated routes was applied and the percentage of overlap between the buffers was determined. The results showed that the largest total length difference can be found in London, combined with the lowest polygon overlap value, indicating slightly different routes between the two generated networks.

Table 10.5.: Comparison of 100 tested shortest-path calculations.

	Berlin	Riga	London
Total length of tested routes for street network graph	210.576 km	229.612 km	225.129 km
Total length of tested routes for sidewalk network graph	211.876 km	230.386 km	227.445 km
Difference	+1.300 km	+0.774 km	+2.236 km
Average percentage overlap between the result of the street network and sidewalk route (10 m buffer method)	90%	89%	78%

Next to the aforementioned factors, it is important to evaluate whether the computed path, based on the newly presented approach, exists only along major street types, such as primary or secondary roads, or if it also contains footways or sidewalks, i.e. ways that are not accessible to motorized traffic. Figure 10.3 illustrates the number of road features that were utilized during the generation of the routes based on the regular road network in each city.

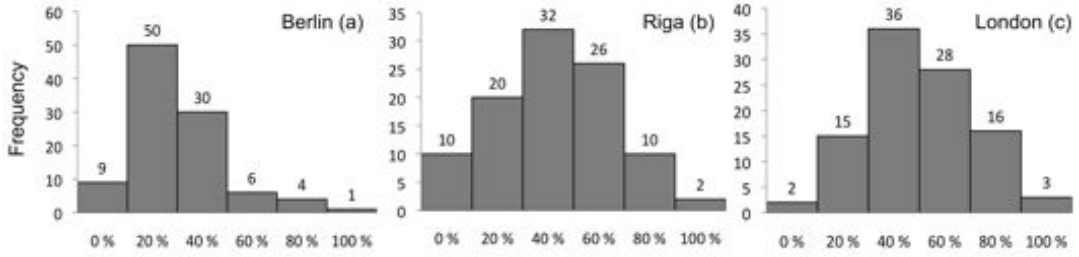


Figure 10.3.: Percentage of footway feature lengths.

Additionally, the corresponding percentage of footway information that was implemented in the total route length was computed. The results show that the generated routes for Riga and London have a higher percentage while Berlin reveals the lowest

value in this comparison. These results can be compared to the percentage of footway information utilized during generation of the tested routes based on the newly generated sidewalk network (Figure 10.4). All three diagrams show an improvement in the number of footway features. Although London includes less sidewalk information in the OSM dataset in comparison to Berlin, the tested area in London still shows a similar or slightly better result. Similarly good results can be reported for Riga, where the test dataset only contains about 36% of the sidewalk information that is needed. However, the majority (89%) of the calculated routes implement more than 60% of footway or sidewalk information.

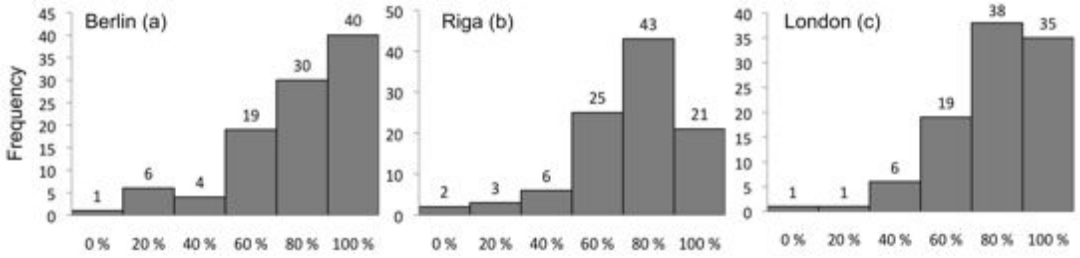


Figure 10.4.: Percentages of footway and sidewalk information in routes with sidewalks.

Further, the quantity of the previously introduced crucial tags for a disabled friendly routing network was evaluated (cf. Table 10.1 & 10.2). Table 10.6 shows the percentages of features that were tagged with the additional information that is needed to create the desired network. Some of the introduced tags shown in Table 10.2 were missing entirely in the three tested areas.

Table 10.6.: Completeness of disabled routing related sidewalk information.

Percentage in mapped sidewalks	Berlin	Riga	London
Lighting	20.6%	74.2%	91%
Smoothness	1.6%	0%	0%
Surface	28.2%	44.4%	9.8%
Width	0%	0%	2.1%
Number of mapped crossings	79	253	458

However, an additional visual inspection of the selected 50 datasets showed that some special cases occur in Reykjavik (Iceland) or Helsinki (Finland). Reykjavik experienced a data import of sidewalk information which was incorrectly tagged as footways, which contain width and surface information but cannot be utilized with the current erroneous tags. A similar situation can be found in Helsinki, where many sidewalks

were mapped as separated footway objects which do not contain the required tags to create a sophisticated routing network for disabled people. Furthermore, it seems that many sidewalks were mapped as separated footways only for map rendering purposes, one of the caveats of the open approach to data collections in OSM. Contributors tend do make these changes to the dataset so that each object is illustrated and rendered in the actual map by the default OSM map engines, instead of just being linked as additional tags somewhere in the database.

10.5. Limitations

During the development of the network and the testing process of the algorithm several problems occurred when utilizing the OSM dataset. As the evaluation of the algorithm has shown, the geodata quality has to be tested for the individual use case. This means that the algorithm can only generate an adequate network if the corresponding sidewalk information is available in the area of interest. A second major issue is the completeness and variety of keys and values that the OSM contributors can apply to the individual objects. The collected information in the tested areas for instance showed, that some contributors use a point as decimal mark while others prefer to use a comma. Others switch between meter and centimeter units when collecting information about the width or the sloped curb of a sidewalk. Other contributors again attach the units of their measurements directly to the value of the object. Besides these errors in naming conventions when tagging an object in OSM, other information in the database is sometimes not interpretable. For instance the key incline, which describes the slope of a street, was used for about 78,000 ways (according to an OSM tag information webpage[5]). 42% of the values of this particular key include information such as "up" and 26% are tagged with "down". This additional information, whether the slope value was taken when going "down" or "up" the road, renders useless when generating a routing network for wheelchair users. This means that almost 68% of the information retrieved from the incline tag uses a temporary value such as "up" or "down" which indicates that further information is needed[6].

A similar issue can be detected when utilizing the key "sloped_curb". The OSM wiki contains detailed information about how the kerb of a sideway should be tagged. For our analysis the key "sloped_curb" was implemented due to its importance on the wheelchair routing webpage (OpenStreetMap 2013c). Several other documenta-

[5]http://taginfo.openstreetmap.org/keys/incline#values (visited on 5 October 2013)
[6]http://wiki.openstreetmap.org/wiki/Key:incline (visited on 5 October 2013)

tions also recommend using the key "kerb"[7], sometimes also referred to as "curb". Next to the different naming conventions, a second ambiguity with this particular tag arises when determining the exact location of the kerb information. Where should the contributor add this information? Should a node be added to the start and the end of a way or should it be added as a tag to the way (e.g. "sidewalk:start:kerb" and "sidewalk:end:kerb")? A standardized tagging convention in this particular case would improve the OSM quality significantly.

However, one of the main questions that arise is: Do contributors map this detailed information worldwide although it is not being rendered in the OSM standard maps? At least in recent years the volunteers started collecting detailed information beyond the scope of regular streets or buildings. A few years ago, the OSM dataset did not provide any turn restriction or detailed address information for navigation applications. After the community was introduced to applications that utilize this information, there was an increase in mapping and tagging efforts for these particular attributes.

10.6. Conclusions and future work

In this article we introduced a newly developed algorithm that generates a routing network for disabled people from a freely available and collaboratively collected geodataset, provided by the OSM project. The newly created network proved to have several advantages over traditional routing networks and is highly adaptable. The variety of supported attributes during the network generation allows the algorithm to be used for different use cases such as routeplaners or personal navigation assistants for people with disabilities. Furthermore, the new representation of a sidewalk network can be implemented in several types of online, offline and printed maps.

During the development of the prototype of the algorithm several issues occurred with the applied VGI dataset. In some cases the provided information proved to be unfeasible due to contributor collection errors or the lack of information in the selected test area. Therefore it needs to be noted that the preferred type of information and its corresponding quality have to be tested for each individual case where OSM data will be utilized (c.f. Mooney et al. 2013, Mondzech and Sester 2011). However, the proposed algorithm and its generated network for pedestrians and disabled people provide room for new research projects based on the current findings, such as the combination with OSM 3D city models (Goetz 2012a or indoor (Goetz 2012b), blind (Kammoun et al. 2010) and tactile (Pielot and Boll 2010) routing applications.

[7] http://wiki.openstreetmap.org/wiki/Proposed_features/kerb (visited on 5 October 2013)

Furthermore, several improvements to the algorithm are feasible. During the generation of the sidewalk network it could be useful to consider building information, which is also available in the OSM project database, to position the sidewalks correctly between the road and a row of houses, similar to the work introduced by Ballester et al. (2011). Some required tags, such as the incline of a road, are currently not widely mapped by the volunteers of the OSM project. In this particular case, the combination of the 2D way geometry from OSM together with a *Digital Elevation Model* (DEM) could result in a strong improvement (cf. Beale et al. 2006).

Lastly, combining the suggested generated network with the original OSM data topology would allow the development of a multi modal routing graph that implements sidewalk and public transportation network information, e.g. to plan a route for wheelchair users. Also, barriers such as street lamps or road signs in the middle of a sidewalk should be taken into account during the creation of the new sidewalk network.

References

Anderson, P. (2007). What is Web 2.0? Ideas, Technologies and Implications for Education. In: *JISC*.

Ballester, M.G., Pérez, M.R., and Stuiver, H.J. (2011). Automatic Pedestrian Network Generation. In: *Proceedings of the 14th AGILE International Conference on Geographic Information Science*. (Apr. 18–21, 2011). Utrecht, The Netherlands.

Beale, L., Field, K., Briggs, D., Picton, P., and Matthews, H. (2006). Mapping for Wheelchair Users: Route Navigation in Urban Spaces. *The Cartographic Journal*, 43(1), 68–81.

Bono, F. and Gutiérrez, E. (2011). A network-based analysis of the impact of structural damage on urban accessibility following a disaster: the case of the seismically damaged Port Au Prince and Carrefour urban road networks. *Journal of Transport Geography*, 19, 1443–1455.

Brando, C. and Bucher, B. (2010). Quality in User-Generated Spatial Content: A Matter of Specifications. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

Brown, G. (2012). An Empirical Evaluation of the Spatial Accuracy of Public Participation GIS (PPGIS) Data. *Applied Geography*, 34, 289–294.

Budhathoki, N. and Haythornthwaite, C. (2013). Motivation for Open Collaboration: Crowd and Community Models and The Case of OpenStreetMap. *American Behavioral Scientist*, 57, 548–575.

Charles, S. H., Kincho, H. L., Jean-Claude, L., and John, C. K. (2002). A Performance-based Approach to Wheelchair Accessible Route Analysis. *Advanced Engineering Informatics - Artificial*, 16(1), 53–71.

Corona, B. and Winter, S. (2001). Datasets for Pedestrian Navigation Services. In: *Proceedings of the AGIT Symposium*. (July 4–6, 2001). Salzburg, Austria.

Ding, D., Parmanto, B., Karimi, H.A., Roongpiboonsopit, D., Pramana, G., Conahan, T., and Kasemsuppakorn, P. (2007). Design Considerations for a Personalized Wheelchair Navigation System. In: *Proceedings of the 29th Annual International Conference of the IEEE EMBS Cité Internationale*. (Aug. 23–26, 2007). Lyon, France.

Flanagin, A. J. and Metzger, M. J. (2008). The Credibility of Volunteered Geographic Information. *GeoJournal*, 72, 137–148.

Gaisbauer, C. and Frank, A.U. (2008). Wayfinding Model for Pedestrian Navigation. In: *Proceedings of the 11th AGILE International Conference on Geographic Information Science 2008*. (May 5–8, 2008). Rome, Italy.

Girres, J.F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS*, 14, 435–459.

Goetz, M. (2012a). Towards Generating Highly Detailed 3D CityGML Models from OpenStreetMap. *International Journal of Geographical Information Science*, 27(5), 1–21.

Goetz, M. (2012b). Using Crowd-sourced Indoor Geodata for the Creation of a Three-dimensional Indoor Routing Web Application. *Future Internet*, 4, 575–591.

Goetz, M. and Zipf, A. (2012). Using Crowdsourced Indoor Geodata for Agent-Based Indoor Evacuation Simulations. *ISPRS International Journal of Geo-Information*, 1, 186–208.

Goodchild, M.F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211–221.

Goodchild, M.F. and Hunter, G.J. (1997). A Simple Positional Accuracy Measure for Linear Features. *International Journal of Geographical Information Science*, 11, 299–306.

Hagenauer, J. and Helbich, M. (2012). Mining Urban Land Use Patterns from Volunteered Geographic Information by Means of Genetic Algorithms and Artificial Neural Networks. *International Journal of Geographical Information Science*, 26, 963–982.

Haklay, M. (2010). How good is OpenStreetMap Information: A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. *Environment and Planning B*, 37, 682–703.

Kammoun, S., Dramas, F., Oriola, B., and Jouffrais, C. (2010). Route Selection Algorithm for Blind Pedestrian. In: *Proceedings of the International Conference on Control, Automation and Systems*. (Oct. 27–30, 2010). KINTEX, Gyeonggi-do, Korea.

Karimi, H.A. and Kasemsuppakorn, P. (2012). Pedestrian Network Map Generation Approaches and Recommendation. *International Journal of Geographical Information Science*, 27(5), 947–962.

Kasemsuppakorn, P. and Karimi, H.A. (2008). Data Requirements and Spatial Database for Personalized Wheelchair Navigation. In: *Proceedings of the 2nd International Convention on Rehabilitation Engineering and Assistive Technology*. (May 13–15, 2008). Bangkok, Thailand.

Kasemsuppakorn, P. and Karimi, H.A. (2009). Personalised Routing for Wheelchair Navigation. *Journal of Location Based Services*, 3(1), 24–54.

Kasemsuppakorn, P. and Karimi, H.A. (2013). A Pedestrian Network Construction Algorithm Based on Multiple GPS Traces. *Transportation Research Part C: Emerging Technologies*, 26, 285–300.

Kim, J., Park, S.Y., Bang, Y., and Yu, K. (2009). Automatic Derivation Of A Pedestrian Network Based On Existing Spatial Datasets. In: *Proceedings of the ASPRS /MAPPS 2009 Fall Conference*. (Nov. 16–19, 2009). San Antonio, Texas.

Matthews, H., Beale, L., Picton, P., and Briggs, D. (2003). Modelling Access with GIS in Urban Systems (MAGUS): Capturing the Experiences of Wheelchair Users. *Area*, 35(1), 34–45.

Menkens, C., Sussmann, J., Al-Ali, M., Breitsameter, E., Frtunik, J., Nendel, T., and Schneiderbauer, T. (2011). EasyWheel - A Mobile Social Navigation and Support System for Wheelchair Users. In: *Proceedings of the 8th International Conference on Information Technology: New Generations*. (Apr. 11–13, 2011). Las Vegas, Nevada, USA.

Moghadam, H.S. and Helbich, M. (2013). Spatiotemporal urbanization processes in the megacity of Mumbai, India: A Markov chains-cellular automata urban growth model. *Applied Geography*, 40, 140–149.

Mondzech, J. and Sester, M. (2011). Quality Analysis of OpenStreetMap Data based on Application need. *Cartographica*, 46, 115–125.

Mooney, P., Corcoran, P., and Ciepluch, B. (2013). The Potential for using Volunteered Geographic Information in Pervasive Health Computing Applications. *Journal of Ambient Intelligence and Humanized Computing*, 4(6), 731–745.

Mooney, P., Corcoran, P., and Winstanley, A. (2010). Towards Quality Metrics for OpenStreetMap. In: *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. (Nov. 2–5, 2010). San Jose, CA, USA.

Neis, P., Goetz, M., and Zipf, A. (2012a). Towards Automatic Vandalism Detection in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 315–332.

Neis, P., Zielstra, D., and Zipf, A. (2012b). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4, 1–21.

Neis, P., Zielstra, D., and Zipf, A. (2013). Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Future Internet*, 5, 282–300.

Neis, P. and Zipf, A. (2008). OpenRouteService.org is Three Times "Open": Combining OpenSource, OpenLS and OpenStreetMaps. In: *Proceedings of the GIS Research UK 16th Annual conference GISRUK 2008*. (Apr. 2–4, 2008). Manchester, UK.

Neis, P. and Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project - The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 146–165.

Nielsen, J. (2006). *Participation Inequality: Encouraging More Users to Contribute*. URL: `http://www.nngroup.com/articles/participation-inequality/` (visited on 09/23/2013).

OpenStreetMap (2012a). *Osmosis - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Osmosis` (visited on 08/11/2012).

OpenStreetMap (2012b). *Planet OpenStreetMap*. URL: `http://planet.osm.org` (visited on 08/01/2012).

OpenStreetMap (2013a). *Map Features - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Map_Features` (visited on 10/05/2013).

OpenStreetMap (2013b). *Nominatim - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Nominatim` (visited on 10/05/2013).

OpenStreetMap (2013c). *Wheelchair Routing - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Wheelchair_routing` (visited on 10/05/2013).

OSMstats (2013). *Statistics of the free wiki world map*. URL: `http://osmstats.altogetherlost.com` (visited on 10/05/2013).

Pielot, M. and Boll, S. (2010). Tactile Wayfinder: Comparison of Tactile Waypoint Navigation with Commercial Pedestrian Navigation Systems. *Pervasive Computing. Lecture Notes in Computer Science*, 6030, 76–93.

Ramm, F., Topf, J., and Chilton, S. (2010). Cambridge, UK: UIT.

Renz, J. and Wölfl, S. (2010). A Qualitative Representation of Route Networks. In: *Proceedings of the 19th European Conference on Artificial Intelligence*. (Aug. 16–20, 2010). Lisbon, Portugal.

Schmitz, S., Neis, P., and Zipf, A. (2008). New Applications Based on Collaborative Geodata - the Case of Routing. In: *Proceedings of XXVIII INCA International Congress on Collaborative Mapping and Space Technology*. (Nov. 4–6, 2008). Gandhinagar, Gujarat, India.

Sobek, A. and Miller, H. (2006). U-Access: A Web-based System for Routing Pedestrians of Differing Abilities. *Journal of Geographical Systems*, 8(3), 269–287.

Walter, V., Kada, M., and Chen, H. (2006). Shortest Path Analyses in Raster Maps for Pedestrian Navigation in Location based Systems. In: *Proceedings of the International Symposium on Geospatial Databases for Sustainable Development*. (Sept. 27–30, 2006). Goa, India.

Zielstra, D. and Hochmair, H.H. (2011a). A Comparative Study of Pedestrian Accessibility to Transit Stations using Free and Proprietary Network Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2217, 145–152.

Zielstra, D. and Hochmair, H.H. (2011b). Digital Street Data: Free versus Proprietary. *GIM International*, 25, 29–33.

Zielstra, D. and Hochmair, H.H. (2012). Comparison of Shortest Path Lengths for Pedestrian Routing in Street Networks Using Free and Proprietary Data. In: *Proceedings of the Transportation Research Board - 91st Annual Meeting*. (Jan. 22–26, 2012). Washington, DC, USA.

11. Measuring the Reliability of Wheelchair User Route Planning based on Volunteered Geographic Information

Authors

Pascal Neis

Journal

Transactions in GIS

Status

Accepted: 16 February 2014 / Revised: 13 February 2014 / Revised: 14 December 2013 / Submitted: 2 October 2013

Reference

Neis, P. (2014). Measuring the Reliability of Wheelchair User Route Planning based on Volunteered Geographic Information. *Transactions in GIS*, (accepted).

Contribution statement

Pascal Neis developed the methods and applications for this study and wrote the manuscript.

.....

Prof. Dr. Alexander Zipf

Abstract

The development of a wheelchair user friendly route planning application inherits a number of special requirements and details that need to be considered during the generation of the routing graph and the corresponding algorithm, making this task much more complex in comparison to car or pedestrian related applications. Each wheelchair type and more importantly each individual user might have different needs with regards to the way condition or other criteria. This study proposes a new approach to route planning for wheelchair users tailored for individual and personal requirements provided by the user and the calculation of a reliability factor of the computed wheelchair path. The routing graph is based on the freely available *Volunteered Geographic Information* (VGI) retrieved from the *OpenStreetMap* (OSM) project. The newly created algorithm is evaluated and tested for a selected area in Bonn, Germany. A new reliability factor is introduced that gives a direct feedback about the quality of the generated path. Similar factor estimations can also be utilized for multiple route planning applications where VGI or other commercial or administrative data is implemented and more detailed factors than a simple geometric representation of a street network are of importance.

Keywords: route planning; wheelchair users; volunteered geographic information; reliability.

11.1. Introduction

Tailored routing applications for people with special needs have experienced an increased interest by researchers and developers in recent years. Most commonly used route planners do not consider special routing cases, due to the lack of detailed information in the utilized datasets. One of the main differences in comparison to widely used route planners for cars or pedestrians is that the surface or the incline of a way or path can strongly affect the usability of the particular way during the determination of a wheelchair path. Other major factors that might influence the results can be different wheelchair types, such as electric or power wheelchairs, in comparison to manual wheelchairs, or if a wheelchair user is alone or supervised by an additional person. Each setting has its own individual characteristics (Matthews et al. 2003) and therefore special requirements that need to be included in the utilized geodata (Beale et al. 2006). Proprietary and administrative data usually lack the detail of information that is needed for an adequate pedestrian or wheelchair route planning application

(Holone et al. 2008, Neis and Zielstra 2014b). During the generation of the route it is essential to give each way segment (with its corresponding parameters) a weight to assure a safe, accurate and efficient route to the desired destination for each individual (Kasemsuppakorn and Karimi 2009).

A number of extensive surveys in recent research studies allowed for the determination of several important characteristics that the data source of interest needs to inherit. The methods introduced in these studies ranged from simple digitizing tasks and pre-processing steps of datasets (Beale et al. 2006, Kasemsuppakorn and Karimi 2008, Kammoun et al. 2010, Neis and Zielstra 2014b), over newly developed tools that derive a network through buffering (Karimi and Kasemsuppakorn 2012) or by implementing pedestrian GPS traces (Kasemsuppakorn and Karimi 2013), to the creation of an adequate network based on binary image processing procedures (Gaisbauer and Frank 2008, Kim et al. 2009).

Next to these new developments, several research studies in recent years (Holone et al. 2007, Neis and Zipf 2008, Neis et al. 2012b, Zielstra and Hochmair 2012, Neis and Zielstra 2014a) and projects such as OpenRouteService (2013) or MapQuest (2013) have shown that collaboratively collected geo-information from volunteers, also known as *Volunteered Geographic Information* (VGI; Goodchild 2007), in particular data from the *OpenStreetMap* (OSM) project, can be a reliable data source for car, bicycle and pedestrian routing or navigation applications. In comparison to the datasets utilized in the aforementioned research studies with individual, none interoperable solutions that are sometimes solely based on use case generated wheelchair routing networks, the OSM project has the potential to be the only and central database that allows the collection of the desired wheelchair network data with its corresponding attributes. For instance the WheelMap (2013) project collects information about the accessibility of locations for wheelchair users based on OSM data and proves that the community is willing and able to collect such type of detailed information for these particular users with special requirements. Furthermore, in comparison to proprietary dataset providers, the OSM project has the significant advantage that contributors can simply add new objects to the database tailored to their needs.

Thus, the objective of this article is to take these latest trends into consideration and present a new approach to route generation for wheelchair users based on the data of the OSM project. However, due to the heterogeneous pattern of the data quality of this VGI source, it is crucial to provide additional information about the quality of the generated route. For this purpose an additional method is introduced which focuses on the calculation of a reliability value for the computed path. The article

is structured as follows: In the second section related research work, including new developments in the determination of the requirements of a wheelchair route network and their corresponding routing algorithm and cost functions will be introduced. This section also contains a brief introduction to OSM as a relevant dataset. Section three illustrates which OSM information is required to create a wheelchair routing network. The fourth section describes the newly developed weighting function for the wheelchair routing algorithm and explains how the reliability of the generated path is evaluated. This part is followed by a section that evaluates the newly presented approach. The last section summarizes the findings and gives an overview of potential future research.

11.2. Related work

Several studies about the development of route planning applications for wheelchair users have been published in the past ten years (Matthews et al. 2003, Beale et al. 2006, Holone et al. 2008, Völkel and Weber 2008). The common solution in most of these approaches is the implementation of the Dijkstra (2010) algorithm. A wheelchair routing algorithm requires a multi-criteria network which makes more advanced methods such as contraction hierarchies (Geisberger et al. 2008) inapplicable. Wheelchair route lengths are also usually less than 10 km long, which means a directional Dijkstra algorithm is sufficient and oftentimes the main goal of former research projects was to proof the feasibility of the used algorithm, instead of the development of a fast route planning application for wheelchair users. In general the optimisation of a wheelchair navigation algorithm is more complex than for car or pedestrian related navigation purposes (Kasemsuppakorn and Karimi 2009). The multi-criteria parameters are essential to find an adequate route for each wheelchair type and of course for the specific user's needs. The properties of a way, such as the surface texture, width or incline, can play a major role in each individual case. Several detailed weighting methods were introduced by Matthews et al. (2003), Beale et al. (2006) and Kasemsuppakorn and Karimi (2009) that show how these requirements can be used during the determination of a wheelchair user's route.

The network graph takes a major role during the development of a wheelchair routing algorithm and the corresponding weighting functions. A number of research studies investigated in detail which parameters are generally required for indoor (Charles et al. 2002) and outdoor applications (Matthews et al. 2003, Sobek and Miller 2006). Kasemsuppakorn and Karimi (2009) compared different studies and projects in which specific parameters were implemented and tested them for their applicability for hand-

icapped people. They distinguished only parameters that are essential for wheelchair related routing and concluded that the following parameters are most important for each segment of the route network: length, width, slope, sidewalk surface, steps, sidewalk conditions and sidewalk traffic. Nearly all of these parameters can be gathered by conducting a sophisticated survey; only the traffic detection on a sidewalk can be complicated and will not be considered in the presented algorithm of this paper. Instead it was decided to implement other parameters such as the height of the kerb (Beale et al. 2006), if a crossing was handicap friendly or not (Beale et al. 2006) and if the route was equipped with some sort of lighting.

The length of a way, i.e. the distance between the start and end point, is usually applied as a weight in most routing algorithms. The width of a route is an important factor for wheelchair routing and has been defined with a minimum width of 90 to 91.5 cm by the DIN 18024-1 (DIN 18024-1 1998), a specification by the German Institute for Standardization (in German: *Deutsches Institut für Normung* (DIN)). A similar value can be found in the United States' *Americans with Disabilities Act* (ADA) standard for Accessible Design (ADA 2010). The slope indentifies the incline of the way which limits the accessibility for wheelchair users when reaching a certain threshold. This parameter is mostly defined in percent and should not lie above 3-6% (DIN 18024-1 1998) or 5% (ADA 2010). The surface parameter describes the surface type of the way. The DIN 18024-1 and ADA 2010 standards do not suggest detailed values for this parameter and classify ways as "shall be easy, with low-vibration and safely accessible in each weather condition " (DIN 18024-1 1998) and "ground surfaces shall be stable, firm, and slip resistant" (ADA 2010). These definitions do not allow the surface of a way for wheelchairs to be of loose grit/gravel or grass types which needs to be considered when developing the algorithm. Similar to the surface condition, the smoothness of a way can play a major role for wheelchair users. Steps are of crucial importance and constitute an insurmountable obstacle for non-electric or un-supervised wheelchair users and will not be considered as shortcuts during the way finding process. Lastly, the maximum height of a kerb to access a sidewalk is defined with less than 3cm (DIN 18024-1 1998) for wheelchair users.

The introduced parameters are of high importance when developing a wheelchair user friendly routing algorithm. Holone et al. (2008), Rashid et al. (2010) and Menkens et al. (2011) proved in their work that the user generated geodata, also known as VGI, of the *OpenStreetMap* (OSM) project can be utilized for this particular case to assist in the generation of wheelchair navigation or routing applications.

Since its initiation in 2004 the OSM project has attracted more than 1.3 million

registered members (OSMstats 2013) who collaboratively collect, edit and update geodata (Mooney and Corcoran 2013). The majority (almost 70%) of the members are located and active in Europe (Neis and Zipf 2012). However, at the time of writing less than 2% of all registered members were actively contributing geodata to the project on a monthly basis (OSMstats 2013). The collected information covers a plethora of objects of the real world, such as streets, buildings, points of interests, landuse, public transportation or railway information and is represented in the OSM project through Nodes, Ways and Relations. For all objects the volunteer can add attributes, in OSM referred to as Tags (OpenStreetMap 2013f) or key-value-pairs, to describe the feature in more detail. The Relation object can be utilized to map relations between the aforementioned OSM objects such as bus- or tram-routes.

Several research studies in recent years have shown that the geometric representation of the real world in OSM can be highly accurate and complete for different cities or countries in the world, sometimes performing even better than commercial or administrative datasets (Haklay 2010, Girres and Touya 2010, Ludwig et al. 2011, Zielstra and Hochmair 2011, Neis et al. 2012b, Zielstra et al. 2013, Fairbairn and Al-Bakri 2013, Neis and Zielstra 2014a). It needs to be noted that most analyses only show this pattern in urban areas while rural areas do not show the same detail in coverage (Hagenauer and Helbich 2012, Koukoletsos et al. 2012). Neis et al. (2013) also revealed that when comparing selected world regions, the data completeness does not show the same quality in all urban areas. Thus, only for selected regions the OSM dataset can be a replacement or at least an alternative to other proprietary data sources (Ludwig et al. 2011, Neis et al. 2012b, Neis et al. 2013). One of the largest caveats of collaboratively mapped street network geodata is the lack of attribute information (Ludwig et al. 2011, Mondzech and Sester 2011, Neis et al. 2012b). For instance, missing street names, turn restrictions or address information (Neis et al. 2012b), particularly for this study, missing sidewalk surface or width information, can have a major impact on the final product. Mooney et al. (2013) and Mondzech and Sester (2011) summarized that the best approach to answer the question whether OSM data should be implemented or not, is to evaluate the quality of the OSM dataset for the selected area of interest and its particular purpose or role in the final project.

11.3. Preparing a wheelchair network based on VGI

When applying an OSM dataset for routing purposes, the existing topology of the data can be used to create a traditional routing graph with vertices and edges (Schmitz et al.

2008, Renz and Wölfl 2010). For any type of specialization in the routing application the parameter matching between the desired requirements and the OSM object tags is important. Table 11.1 summarizes the OSM tags that were utilized during the creation of a wheelchair-user-friendly network based on the designated OSM wiki page (OpenStreetMap 2013g). Additionally, the tags "sidewalk" for sidewalk information and "lit" for lighting information of ways were included during the network generation.

Table 11.1.: OSM tags relevant for the creation of a wheelchair routing network.

Parameter Description	OSM Tag (if several values available, separated by "|")	
	Key	Value
Type of street	highway	*[1]
Sidewalk(s) of the way	sidewalk|footway	left|right|yes|no|both
Width of the way	width	* in [m]
Surface of the way	surface	*[2]
Smoothness of the way	smoothness	*[3]
Slope/incline of the way	incline	* in [%]
Curb/kerb of the way	sloped_curb	* in [m]
Lighting	lit	yes|no
General access of a way	foot|wheelchair	yes|no

Notes: [1]Additional highway-values: trunk, trunk_link, primary, primary_link, secondary, secondary_link, tertiary, tertiary_link, unclassified, living_street, pedestrian, residential, service, track, footway, cycleway, bridleway, steps, access_ramp, crossing [2]Additional surface-values: paved, asphalt, concrete, paving_stones, cobblestone, concrete_plates [3]Additional smoothness-values: excellent, good, intermediate, bad, very_bad

The creation of the wheelchair user network was tested for a designated area in Bonn (Germany) with an overall area size of 2.8 km^2. The raw OSM data was retrieved in form of a planet dump file (OpenStreetMap 2013d) and clipped to the size of the desired test area with the Java based OSMOSIS tool (OpenStreetMap 2013c). The final transformation of the OSM street network to a more sophisticated sidewalk representation was accomplished by utilizing the algorithm introduced by Neis and Zielstra 2014b. The authors tested the newly developed algorithm and applied it to OSM data for several selected regions in Europe. They revealed that in areas with good data density and the required detailed information, the collaboratively collected geodata can be utilized for disabled people friendly routing applications.

The following Figure 11.1 shows the entire wheelchair network of the tested area in the central business district of Bonn (Germany) with a total graph length of 56 km.

The figure clearly shows that not all streets in the tested area contain sidewalk or street condition information.

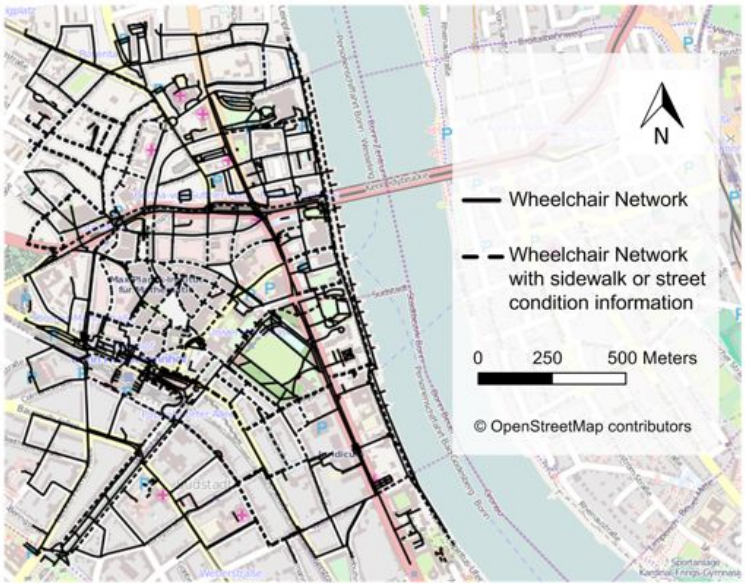


Figure 11.1.: Wheelchair routing network of the test area in Bonn (Germany) (data date August 18th, 2013).

11.4. Weighting and reliability

The generation of a personal wheelchair routing network highly depends on the specific user parameters and preferences that need to be taken into account. The routing process consists of two steps accordingly: "route preference quantification and route calculation" (Kasemsuppakorn and Karimi 2009). This means that based on the user's requirements each way segment in the routing graph has to receive its own weight. Matthews et al. (2003) and Beale et al. (2006) presented an approach to this matter where the impedance of each requirement was determined by a survey in which wheelchair users indicated impassable way segments that, for instance, had a poor surface texture such as gravel. Kasemsuppakorn and Karimi (2009) used a different approach and described three methods on how to weigh the impedance level for each computation based on the individual user requirements. The "*Absolute Restrictions Method*" (ARM) computed suitable routes by eliminating steps and ways with low width values. Inapplicable segments such as parts of ways with a poor surface texture receive a high cost during the route computation in ARM.

11.4.1. Prioritizing user requirements and determining individual impedance

The newly developed routing algorithm will only utilize segments in its computation which are conforming with the user requirements and more specifically only allows particular maximum thresholds for each parameter. For instance, if a user specifies that she/he only wants to use a route that includes segments with sloped curb values below 3 cm, the path determined by the algorithm will not contain any way segments that have curbs with a height above this value. Table 11.2 illustrates each parameter (left) with its values (middle) and importance/weight (top). This means that the wheelchair user can easily identify which parameter is most important for her or his planned route, based on each individual specific parameter. The corresponding values for each parameter shown in Table 11.2 are utilized during the routing computation to include or exclude individual way segments. A parameter can also be marked as "Equal", which means that this particular parameter and its values will not be considered as an obstacle.

Table 11.2.: Personalized weight-parameters and score values.

			Weighting		
	Extremely Important	Very Important	Important	Less Important	Equal
			Score Value		
Parameter	1	2	3	4	0
1. Slope	$< 3\%$	$< 4\%$	$< 5\%$	$< 6\%$	*
2. Width	> 120 cm	> 110 cm	> 100 cm	> 91.5 cm	-[1]
3. Surface	Concentre	Asphalt	Paving Stones	Cobblestone	*
4. Smoothness	Excellent	Good	Intermediate	Bad	*
5. Sloped Curb	< 3 cm	< 5 cm	< 7 cm	< 9 cm	*
6. Lighting	100%	75%	50%	25%	*

Notes: [1]A way for a wheelchair user must have at least a width of 91.5cm

Based on the user's choice for each parameter, Score Values (at the top) are assigned (Table 11.2). A low Score Value represents a high importance, whereas 0 means the parameter is of no importance. For instance, a way segment with a slope of < 3 % will receive a Score Value of 1 and a segment with a slope of < 5 % a Score Value of 3 accordingly. The Score Values are separated into different classes which are based on the range of values that can be found in the DIN/ADA standards and OSM tags (Table 11.1). Overall the Score Values are crucial for the computation of the resulting parameter weights. An example scenario for a user's selection based on his or her

individual preferences is shown in Table 11.3.

Table 11.3.: Example of user selection and resulting parameter weights.

Weighting	Extremely Important	Very Important	Important	Less Important	Equal		
Score Value	1	2	3	4	0		
			Weight-Percentage			Weight	Individual
Parameter	100%	75%	50%	25%	0%		Weight
1. Slope	1	0	0	0	0	100%	30.8%
2. Width	0	0	0	1	0	25%	7.7%
3. Surface	0	1	0	0	0	75%	23.1%
4. Smoothness	0	0	1	0	0	50%	15.4%
5. Sloped Curb	0	1	0	0	0	75%	23.1%
6. Light	0	0	0	0	0	0%	0%
					Sum:	325%	100%

The Score Values and the corresponding Weight-Percentages reflect the importance of a parameter. Thus the sum of all individual weight-percentages of the parameters and the relative parameter weights for each parameter (at the right side of Table 11.3) can be calculated. These final parameter weights are essential for the cost calculation of each segment during the routing process. The final *Impedance-Score* (IS) of each segment can then be computed based on the aggregation of each parameter Score Value and the user's individual parameter weight (Equation 11.1).

$$ImpedanceScore(IS) = \sum_{i=1}^{6} ScoreValue_i * IndividualWeight_i \tag{11.1}$$

11.4.2. Path and reliability determination

A routing graph is traditionally defined as $G = \{V, E\}$ where V is a set of vertices and E a set of edges between those vertices. In our particular case each E^n has a variety of attributes gathered from the OSM dataset, such as the surface or smoothness parameter. Based on the aforementioned IS and the length of the segment, a *Cost* for each E^n can be calculated (Equation 11.2).

$$Cost = ImpedanceScore * length \tag{11.2}$$

During the calculation of the *Cost* value, each parameter of the segment is checked against the individual requirements provided by the user. The IS of a segment is '0' when a parameter has a higher Score Value than the one determined from the user

input. The *Cost* of a segment will be '0' accordingly, which indicates that the segment is impassable during the personal route determination. For instance, the following situation may occur: The width of a way segments is very important to a wheelchair user (>110cm, Table 11.2). Thus the Score Value is set to 2 by the user and the Weight-Percentage of 75% for the width parameter is applied (cf. Table 11.3). Based on these requirements, the IS will be set to '0' if a way segment proves to have a width of 105cm, which results in a Score Value of 3.

This step differs from the *Cost*-function proposed by Kasemsuppakorn and Karimi (2009), who implemented all segments of their generated network, during the path computation (except steps and narrow ways). Our approach guarantees that only segments will be utilized during the route generation that have an equal or better Score Value than the user's requirements. The computation of each individual and personal route highly depends on the quality and quantity of the utilized geodata. In this study the wheelchair routing network is based on the VGI data provided by the OSM project. Due to the open approach to data contributions in OSM the collaboratively mapped geodata can show inequalities in quality depending on the particular country and area of interest. If a way segment of the routing graph is missing one of the introduced parameter information, for example surface condition, the impedance score function will use the worst value during the computation.

Next to the generation of the routing graph and the computation of the best route for the user, it is essential to provide some sort of quality statement about the calculated path. Thus, the objective of the last method introduced in this paper is to provide a reliability value which evaluates the fitness of the generated route to the user's requirements. Therefore the presented weight-parameters (Table 11.2) and the corresponding personal individual weights (Table 11.3) are implemented in a new Equation 11.3. For each parameter that has an individual weight, the total length of all segments of the route that contain a value for the related attribute and the total length of all segments of the route will be aggregated. The ratio of the lengths is multiplied by the individual weight. The sum of all products represents the *Reliability Factor* (RF) of the computed route based on the used data during the route computation.

$$ReliabilityFactor(RF) = \sum_{i=1}^{6} IndividualWeight_i * \frac{AvailableLength_i}{TotalLength_i} \qquad (11.3)$$

The RF provides additional information about the quality of the generated route, influenced by the attribute availability during the route computation, which can be

crucial for the individual wheelchair user.

11.5. Evaluation

The evaluation of the proposed algorithm is based on a comparison between a set of randomly sampled routes, generated with a regular wheelchair weighting function, similar to the ARM introduced by Kasemsuppakorn and Karimi (2009) and our developed weighting function. The objective of the evaluation is to show how routing results can differ between the two approaches and where potential strengths and weaknesses can be found. The origin and destination points of a total number of 40 paths were randomly selected in the generated OSM wheelchair routing network of Bonn. The route generations were computed with the Java GeoTools (2013) framework and the prototype for each weighting function and the final reliability factor were implemented accordingly. All test routes were simulated for a wheelchair user who has the following requirements:

1. The width of the sidewalk or way is equal to or larger than 1 m.
2. The surface condition is better or equal to the 'paving stones' class.
3. The smoothness of the section is better or equal to 'good'.

The introduced RF for the utilized OSM wheelchair network plays a major role in the evaluation due to the irregular data quality pattern of the OSM dataset and provides useful information about the reliability of the determined route.

The results of the analysis in Figure 11.2 show that the RFs (based on Equation 11.3) of the 40 calculated routes range between 5% and 83%. The comparison of the passability factors of the two line charts highlights the difference between both weighting functions. The regular wheelchair routing weighting function (Figure 11.2a) does not exclude impassable segments during the path computation, thus it is possible that the determined route contains streets that are not passable by the wheelchair user. This caveat is also represented in the RF which only provides information whether the computed paths contain sidewalk or street condition information, but does not identify if all parts of the path are passable. An opposite result can be seen for our introduced weighting function (Figure 11.2b) where the passability factor and the RF are equal for all routes.

However, the following Figure 11.3a shows a comparison of the sorted RFs of all calculated routes for both wheelchair weighting functions. Overall the regular wheelchair weighting function shows a higher RF than the newly introduced function. It needs

to be noted though that, similarly to the results shown in Figure 11.2a, the aforementioned issue that this RF does not consider impassable road sections also influences the results shown in Figure 11.3a.

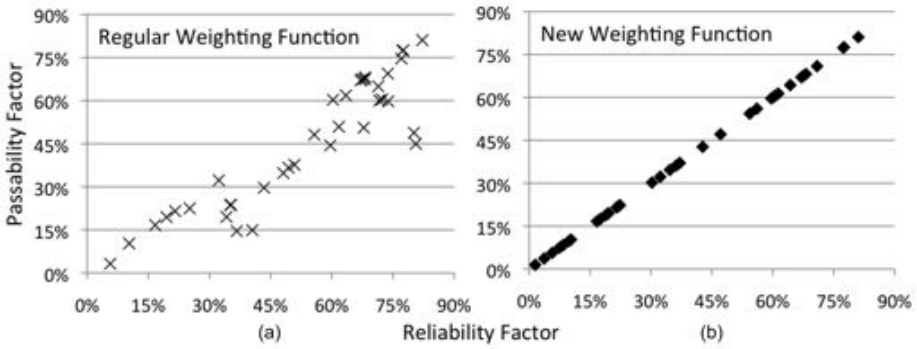


Figure 11.2.: Distribution of reliability and passability factor of 40 sample routes for two weighting functions.

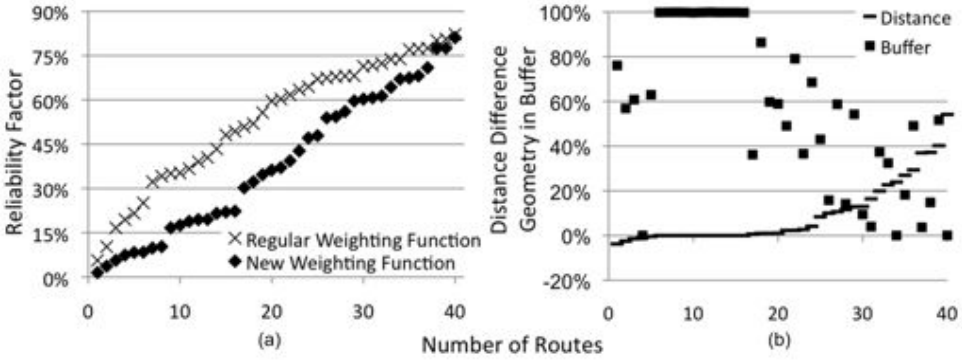


Figure 11.3.: (a) Comparison between RFs and (b) Distance and geometry differences between generated routes with regular and new weighting function.

Additionally, the differences between the route lengths of the two weighting functions were computed (Figure 11.3b) and tested for their corresponding similarities or differences in geometries by applying a buffer comparison method introduced by Goodchild and Hunter (1997). A buffer of 1 m on each side of the generated routes was applied and the percentage of overlap between the buffers was determined. The graph shows that the newly introduced weighting function generates longer routes in seven cases with length differences between 20% and 55% and geometry overlaps between 0% and 50%. While these longer routes are not necessarily desirable in a routing application,

the algorithm guarantees that the calculated route does not contain any impassable way segments for the wheelchair user. Eleven of the generated routes are of identical length and geometry when applying each function and in five cases the new approach results in a shorter route. The shorter routes contain way segments which have worse way conditions, but still fall into the category of the predefined user settings.

Several factors can influence the path generation and routing result accordingly. As described in a prior section of the paper and shown in Figure 11.1, the availability of the detailed attributes in the OSM dataset is crucial to retrieve a sophisticated result. Additionally the routing results can be largely impacted, both positively and negatively, by the requirements of the individual wheelchair user.

In a second analysis we evaluated 40 routes with the same origin and destination points as utilized in the prior comparison but with a slightly different requirement: the smoothness value was changed from good to intermediate. The results showed that this minor change caused for 34 of the 40 routes to have an almost equal total route length and similar route geometry with both weighting functions, indicating that several streets are tagged with intermediate instead of good smoothness values in the tested area.

During the development and evaluation of the proposed algorithm, several problems occurred with the VGI dataset that was utilized. Many cities in selected countries show a good geometric representation of the corresponding street network but these networks usually lack the detail of information that is needed for this use case. In addition to the more obvious errors of omission in the geometries of the dataset and missing attribute information of existing data, it is also essential to note that oftentimes the information that is needed exists in OSM but is annotated in an incorrect or unusual data format. The open approach to data collection in OSM is mostly considered as one of the major advantages of the project (Neis et al. 2012a). Interested data contributors simply register before adding their selected information without considering any type of geodata standards. The disadvantage that lies within this approach is that the data tends to be hard to interpret at times or even turns out to be useless due to inconsistencies in the attribution process. OSM provides the contributors with the opportunity to propose new standards which are usually more tailored towards a specific use of the dataset (OpenStreetMap 2013e). Unfortunately, this also resulted in several duplicate proposed Tags that describe one and the same object or parameter. Additionally it has to be taken into account that several units can be used when measuring one parameter (such as centimetre, meter, inch, feet or percent). These characteristics of the OSM dataset and its resulting difficulties and effects were analyzed and criticized in

several studies in recent years (Brando and Bucher 2010, Girres and Touya 2010, Neis et al. 2012b). The findings prove that it is unavoidable to conduct a manual and visual fitness for purpose check of the OSM geodata and that the program that creates the routing graph needs to be highly adjustable and robust against such inconsistencies in attributes and values. For these reasons the presented approach in this paper, i.e. the calculation of the reliability factor of the computed wheelchair path, is essential. The calculated value will always give immediate feedback to the user to what degree she or he can trust the generated path.

However, an important aspect of the utilized OSM dataset is that every user can also add missing or edit incorrect objects. For a general quality assurance of the collected information, several websites are available such as OSM Inspector (OpenStreetMap 2013b) or Keepright (OpenStreetMap 2013a). An interested contributor or user can find more detailed information on these websites about particular types of errors in the data for a selected area of interest. It needs to be noted, however, that the information provided by these websites highly depends on the corresponding tool that was developed. This means that it is possible that a particular tag (key-value pair) is not considered by the webpage and consequently cannot be tested.

11.6. Conclusions and future work

In this study two novel approaches for the assessment and evaluation of the feasibility of a wheelchair user friendly routing algorithm and its generated path were introduced. The first method computes a tailored, individual path based on specific user requirements, while the second method evaluates the generated path by providing a reliability factor based on the utilized data. The computation of the wheelchair route is highly influenced by the user's restrictions regarding way conditions or other obstacles that can or cannot be passed. Based on these predefined user requirements individual weights for each way segment of the routing graph are computed. In comparison to prior research studies the resulting graph does not include way segments that have way conditions that are worse than the predefined settings made by the user. The results gathered from the evaluation of the introduced algorithm has proven that the newly created weighting function computes reliable wheelchair paths but also highlighted that the results strongly depend on wheelchair user requirements and the provided geodata. For certain cases the limitations defined by the user can lead to detours or unsuccessful route generations, due to the missing information in the dataset or poor road or pavement quality. Further research is needed, for instance in form of an extensive survey,

to determine the maximum detour length a wheelchair user is willing to travel.

The generated wheelchair routing network is based on the freely available VGI dataset of the OSM project and the quality of this data can be problematic at times. On the other hand no large area data alternatives are available for the development of a wheelchair user routing application. The current situation in OSM regarding this special type of information is similar to the situation some years ago when the project was lacking general routing information until first applications sparked the interest of the community to add this information. Maybe a similar development can be seen in the near future with applications that require detailed attributes for wheelchair routing and the OSM project will turn into a central database for applications tailored to people with special needs. Additionally, the development of dedicated online tools that illustrate different quality parameters for the presented wheelchair routing network could improve the situation. Despite the latest developments in OSM data contributions, there is no guarantee that certain object types or attribute information will ever be mapped entirely in OSM for the area of interest. However, with the introduced reliability factor of the route, the user can get a direct feedback if the required information is available and to what degree the generated path can be trusted. This is an important advantage especially when considering the aforementioned heterogeneity of the OSM data quality. Furthermore the reliability factor function can also be used for any type of routing or navigation purpose that is based on other VGI, commercial or administrative datasets in which not only the geometric representation of the world is important, but also other attributes and related metadata about the objects.

The VGI dataset provided by the OSM project proved to be a valuable source for wheelchair route planning as long as the detailed wheelchair related tags are included. Future research focusing on the timeliness of the data needs to be conducted to insure that the OSM dataset is undergoing certain update processes to maintain and improve the currently available data. Prior research studies have shown that the community of the OSM project is contributing and updating the general information (Neis and Zipf 2012, Mooney and Corcoran 2013) but no results have been published about the very detailed attribute information that is needed for wheelchair routing.

Acknowledgments

The author would like to thank Dennis Zielstra and Prof. Alexander Zipf for their valuable comments towards the improvement of this paper.

References

ADA (2010). *Americans with Disabilities Act (ADA) Standards for Accessible Design*. URL: `http://www.ada.gov/2010ADAstandards_index.htm` (visited on 01/01/2014).

Beale, L., Field, K., Briggs, D., Picton, P., and Matthews, H. (2006). Mapping for Wheelchair Users: Route Navigation in Urban Spaces. *The Cartographic Journal*, 43(1), 68–81.

Brando, C. and Bucher, B. (2010). Quality in User-Generated Spatial Content: A Matter of Specifications. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

Charles, S. H., Kincho, H. L., Jean-Claude, L., and John, C. K. (2002). A Performance-based Approach to Wheelchair Accessible Route Analysis. *Advanced Engineering Informatics - Artificial*, 16(1), 53–71.

Dijkstra, E.W. (2010). A Note on two Problems in Connexion with Graphs. *Numerische Mathematik*, 1(1), 269–27.

DIN 18024-1 (1998). *Standard specification by the German Institute for Standardization (Deutsches Institut für Normung (DIN)) 18024-1: "Barrierefreies Bauen - Teil 1: Straßen, Plätze, Wege, öffentliche Verkehrs- und Grünanlagen sowie Spielplätze; Planungsgrundlagen" / Barrier-Free Design - Part 1: Streets, Places, Roads and Recreational Areas; Planning Basics*. URL: `http://nullbarriere.de/din18024-1.htm` (visited on 01/03/2014).

Fairbairn, D. and Al-Bakri, M. (2013). Using Geometric Properties to Evaluate Possible Integration of Authoritative and Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, 2, 349–370.

Gaisbauer, C. and Frank, A.U. (2008). Wayfinding Model for Pedestrian Navigation. In: *Proceedings of the 11th AGILE International Conference on Geographic Information Science 2008*. (May 5–8, 2008). Rome, Italy.

Geisberger, R., Sanders, P., Schultes, D., and Delling, D. (2008). Contraction Hierarchies: Faster and Simpler Hierarchical Routing in Road Networks. In: *Proceedings of the 7th Workshop on Experimental Algorithms, Volume 5038 of Lecture Notes in Computer Science*. Springer.

GeoTools (2013). *The Open Source Java GIS Toolkit*. URL: `http://geotools.org`.

Girres, J.F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS*, 14, 435–459.

Goodchild, M.F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211–221.

Goodchild, M.F. and Hunter, G.J. (1997). A Simple Positional Accuracy Measure for Linear Features. *International Journal of Geographical Information Science*, 11, 299–306.

Hagenauer, J. and Helbich, M. (2012). Mining Urban Land Use Patterns from Volunteered Geographic Information by Means of Genetic Algorithms and Artificial Neural Networks. *International Journal of Geographical Information Science*, 26, 963–982.

Haklay, M. (2010). How good is OpenStreetMap Information: A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. *Environment and Planning B*, 37, 682–703.

Holone, H., Misund, G., and Holmstedt, H. (2007). Users Are Doing It For Themselves: Pedestrian Navigation with User Generated Content. In: *Proceedings of the 2007 International Conference on Next Generation Mobile Applications, Services and Technologies*. (Sept. 12–14, 2007). Cardiff, Wales, UK.

Holone, H., Misund, G., Tolsby, H., and Kristoffersen, S. (2008). Aspects of Personal Navigation with Collaborative User Feedback. In: *Proceedings of the Fifth Nordic Conference on Human-Computer Interaction*. (Oct. 20–22, 2010). Lund, Sweden.

Kammoun, S., Dramas, F., Oriola, B., and Jouffrais, C. (2010). Route Selection Algorithm for Blind Pedestrian. In: *Proceedings of the International Conference on Control, Automation and Systems*. (Oct. 27–30, 2010). KINTEX, Gyeonggi-do, Korea.

Karimi, H.A. and Kasemsuppakorn, P. (2012). Pedestrian Network Map Generation Approaches and Recommendation. *International Journal of Geographical Information Science*, 27(5), 947–962.

Kasemsuppakorn, P. and Karimi, H.A. (2008). Data Requirements and Spatial Database for Personalized Wheelchair Navigation. In: *Proceedings of the 2nd International Convention on Rehabilitation Engineering and Assistive Technology*. (May 13–15, 2008). Bangkok, Thailand.

Kasemsuppakorn, P. and Karimi, H.A. (2009). Personalised Routing for Wheelchair Navigation. *Journal of Location Based Services*, 3(1), 24–54.

Kasemsuppakorn, P. and Karimi, H.A. (2013). A Pedestrian Network Construction Algorithm Based on Multiple GPS Traces. *Transportation Research Part C: Emerging Technologies*, 26, 285–300.

Kim, J., Park, S.Y., Bang, Y., and Yu, K. (2009). Automatic Derivation Of A Pedestrian Network Based On Existing Spatial Datasets. In: *Proceedings of the ASPRS /MAPPS 2009 Fall Conference*. (Nov. 16–19, 2009). San Antonio, Texas.

Koukoletsos, T., Haklay, M., and Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transaction in GIS*, 16, 477–498.

Ludwig, I., Voss, A., and Krause-Traudes, M. (2011). A Comparison of the Street Networks of Navteq and OSM in Germany. *Advancing Geoinformation Science for a Changing World*, 1, 65–84.

MapQuest (2013). *Map Editor, Maker and OSM - MapQuest Open*. URL: `http://open.mapquest.com`.

Matthews, H., Beale, L., Picton, P., and Briggs, D. (2003). Modelling Access with GIS in Urban Systems (MAGUS): Capturing the Experiences of Wheelchair Users. *Area*, 35(1), 34–45.

Menkens, C., Sussmann, J., Al-Ali, M., Breitsameter, E., Frtunik, J., Nendel, T., and Schneiderbauer, T. (2011). EasyWheel - A Mobile Social Navigation and Support System for Wheelchair Users. In: *Proceedings of the 8th International Conference on Information Technology: New Generations*. (Apr. 11–13, 2011). Las Vegas, Nevada, USA.

Mondzech, J. and Sester, M. (2011). Quality Analysis of OpenStreetMap Data based on Application need. *Cartographica*, 46, 115–125.

Mooney, P. and Corcoran, P. (2013). Has OpenStreetMap a role in Digital Earth Applications? *International Journal of Digital Earth*. DOI: `10.1080/17538947.2013.781688`.

Mooney, P., Corcoran, P., and Ciepluch, B. (2013). The Potential for using Volunteered Geographic Information in Pervasive Health Computing Applications. *Journal of Ambient Intelligence and Humanized Computing*, 4(6), 731–745.

Neis, P., Goetz, M., and Zipf, A. (2012a). Towards Automatic Vandalism Detection in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 315–332.

Neis, P. and Zielstra, D. (2014a). Current Developments and Future Trends in VGI Research: The Case of OpenStreetMap. *Future Internet*, 6, 76–106.

Neis, P. and Zielstra, D. (2014b). Generation of a Tailored Routing Network for Disabled People based on Collaboratively Collected Geodata. *Applied Geography*, 47, 70–77.

Neis, P., Zielstra, D., and Zipf, A. (2012b). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4, 1–21.

Neis, P., Zielstra, D., and Zipf, A. (2013). Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Future Internet*, 5, 282–300.

Neis, P. and Zipf, A. (2008). OpenRouteService.org is Three Times "Open": Combining OpenSource, OpenLS and OpenStreetMaps. In: *Proceedings of the GIS Research UK 16th Annual conference GISRUK 2008*. (Apr. 2–4, 2008). Manchester, UK.

Neis, P. and Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project - The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 146–165.

OpenRouteService (2013). *Online route planner*. URL: `http://openrouteservice.org`.

OpenStreetMap (2013a). *Keep Right - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Keep_Right`.

OpenStreetMap (2013b). *OSM Inspector - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/OSM_Inspector`.

OpenStreetMap (2013c). *Osmosis - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Osmosis`.

OpenStreetMap (2013d). *Planet OSM Files*. URL: `http://planet.osm.org`.

OpenStreetMap (2013e). *Proposed features - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Proposed_features`.

OpenStreetMap (2013f). *Tags - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Tags`.

OpenStreetMap (2013g). *Wheelchair routing - OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Wheelchair_routing`.

OSMstats (2013). *Statistics of the free wiki world map*. URL: `http://osmstats.altogetherlost.com` (visited on 10/05/2013).

Rashid, O., Dunbar, A., Fisher, S., and Rutherford, J. (2010). Users Helping Users - User generated content to assist wheelchair users in an urban environment. In: *Proceedings of the 9th International Conference on Mobile Business and 2010 Ninth Global Mobility Roundtable*. (June 13–15, 2010). Athens, Greece.

Renz, J. and Wölfl, S. (2010). A Qualitative Representation of Route Networks. In: *Proceedings of the 19th European Conference on Artificial Intelligence*. (Aug. 16–20, 2010). Lisbon, Portugal.

Schmitz, S., Neis, P., and Zipf, A. (2008). New Applications Based on Collaborative Geodata - the Case of Routing. In: *Proceedings of XXVIII INCA International Congress on Collaborative Mapping and Space Technology*. (Nov. 4–6, 2008). Gandhinagar, Gujarat, India.

Sobek, A. and Miller, H. (2006). U-Access: A Web-based System for Routing Pedestrians of Differing Abilities. *Journal of Geographical Systems*, 8(3), 269–287.

Völkel, T. and Weber, G. (2008). RouteCheckr: Personalized Multicriteria Routing for Mobility Impaired Pedestrians. In: *Proceedings of the 10th international ACM SIGACCESS Conference on Computers and Accessibility*. (Oct. 13–15, 2008). Halifax, Nova Scotia, Canada.

WheelMap (2013). *Wheelmap.org is an online map to search, find and mark wheelchair-accessible places*. URL: `http://wheelmap.org/en`.

Zielstra, D. and Hochmair, H.H. (2011). A Comparative Study of Pedestrian Accessibility to Transit Stations using Free and Proprietary Network Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2217, 145–152.

Zielstra, D. and Hochmair, H.H. (2012). Comparison of Shortest Path Lengths for Pedestrian Routing in Street Networks Using Free and Proprietary Data. In: *Proceedings of the Transportation Research Board - 91st Annual Meeting*. (Jan. 22–26, 2012). Washington, DC, USA.

Zielstra, D., Hochmair, H.H., and Neis, P. (2013). Assessing the Effect of Data Imports on the Completeness of OpenStreetMap – A United States Case Study. *Transaction in GIS*, 17, 315–334.

12. Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap

Authors

Pascal Neis and Dennis Zielstra

Journal

Future Internet

Status

Published: 27 January 2014 / Accepted: 13 January 2014 / Revised: 10 January 2014 / Submitted: 10 December 2013

Reference

Neis, P. and Zielstra, D. (2014). Recent Developments and Future Trends in Volunteered Geographic Information Research: The Case of OpenStreetMap. *Future Internet*, 6(1):76-106.

Contribution statement

Pascal Neis wrote the majority of the manuscript. Dennis Zielstra supported this publication by continuing discussions about potential contributions to the study. Furthermore, extensive proof-reading by the co-author led to substantial improvements to

the manuscript.

.....	.....
Prof. Dr. Alexander Zipf	Dennis Zielstra

Abstract

User-Generated Content (UGC) platforms on the Internet have experienced a steep increase in data contributions in recent years. The ubiquitous usage of location enabled devices, such as smartphones, allows contributors to share their geographic information on a number of selected online portals. The collected information is oftentimes referred to as *Volunteered Geographic Information* (VGI). One of the most utilized, analyzed and cited VGI-platforms, with an increasing popularity over the past few years, is *OpenStreetMap* (OSM), whose main goal it is to create a freely available geographic database of the world. This paper presents a comprehensive overview of the latest developments in VGI research, focusing on its collaboratively collected geodata and corresponding contributor patterns. Additionally, trends in the realm of OSM research are discussed, highlighting which aspects need to be investigated more closely in the near future.

Keywords: Volunteered Geographic Information; OpenStreetMap, User Generated Content, Future Trends.

12.1. Introduction

Since the discontinuation of the selective availability of the *Global Positioning System* (GPS) in 2000, allowing users to receive a non-degraded signal, the increase of GPS enabled devices and new technological developments, allowed more and more people to use their location information for designated *Location Based Services* (LBS), to position their photographs or other information on a world map or to use it for spare time activities such as geocaching (Goodchild 2007, Coleman 2010a, Mooney et al. 2013, Roick and Heuser 2012). In a similar timeframe and as a consequence of the Web 2.0 phenomenon (O'Reilly 2005), Internet users began not only to passively consume information, but also started to create or edit web content based on their individual requirements, preferences or interests. Tapscott (1997) described these participants as "prosumers", a portmanteau of producer and consumer, whereas Coleman et al. (2009) titled them "produsers". As a result of this development, projects, such as Wikipedia, or image and video sharing platforms, such as Flickr or YouTube, experienced a significant increase in user numbers and contributors over the past few years.

Several terms were introduced in mainstream media and research alike to describe this new pattern in data development. General data contributions, such as Wikipedia entries or blog posts, are oftentimes summarized as *User-Generated Content* (UGC;

Anderson 2007) or *User-Created Content* (Wunsch-Vincent and Vickery 2007). The additional geographic component, usually represented by latitude and longitude values, separates a special data type from these contributions, oftentimes termed *crowd-sourced geodata* (Hudson-Smith et al. 2009, Heipke 2010), *collaborative GI* (Bishr and Kuhn 2007, Bishr and Mantelas 2008), or more commonly known as *Volunteered Geographic Information* (VGI; Goodchild 2007). Related concepts that do not solely focus on the collection of information have also been termed in many different ways, such as *collaborative mapping* (Rouse et al. 2007), *wikification of Geographic Information Systems* (GIS) (Sui 2008), *participatory GIS* (Elwood 2006), *public participation GIS* (Sieber 2006) or *web mapping 2.0* (Haklay et al. 2008).

VGI has become a widespread phenomenon in media and academia alike. A number of research projects in recent years have analyzed the advantages and disadvantages related to VGI. In many cases, the researchers investigated the data and contributor information of the *OpenStreetMap* (OSM) project, one of the most successful VGI projects in recent years (Haklay 2010, Mooney et al. 2010a, Neis et al. 2012b, Goetz 2012), which has also been frequently cited in the GIS community (Goodchild 2007, Budhathoki and Haythornthwaite 2013).

The objective of this paper is to provide an overview of current developments in VGI research, focusing on the different methods that were applied to analyze the members and their corresponding data contributions. After discussing the most essential results from the selected studies, different lessons that can be drawn from the presented research and potential future trends that lie in the empirical analysis of VGI datasets will be discussed.

The remainder of this article is structured as follows: The next Section briefly introduces VGI, followed by a comparison of different VGI projects. Section 3 provides an overview and summarizes the findings of previously conducted OSM research. Future trends and questions are discussed in Section 4, followed by a conclusion in Section 5.

12.2. Volunteered geographic information

The term *Volunteered Geographic Information* (VGI), coined by Goodchild (2007) in 2007, describes the process of collecting spatial data by individuals, most times on a voluntary basis. In most cases, the contributed VGI is collected in a database or file system structure and sometimes freely available to other interested Internet users. To be able to contribute to one of the VGI platforms, the information has to match a geographic position. This can either be achieved by tracing it from georeferenced aerial

imagery or by actively collecting GPS tracks with a designated device. Furthermore, a broadband Internet connection and additional hardware in the form of a smartphone or personal computer are needed. Although these prerequisites seem to be trivial in most modern societies, we will discuss at a later point that they tend to explain certain patterns in the global distribution of VGI projects.

The steep increases in VGI contributions lead to a number of diverse platforms and projects utilizing the data and technologies in spatial decision making, participatory planning and citizen science (Elwood 2010). VGI data also experienced more attention, due to its successful implementation for humanitarian or crisis mapping purposes (Roick and Heuser 2012). In this context, the *Humanitarian OpenStreetMap Team* (HOT; OpenStreetMap 2013e) obtained an important role. Since 2009, it has coordinated the creation, production and distribution of free mapping resources to support humanitarian relief efforts in many places around the world. Ushahidi (2013), a different platform that collects humanitarian crisis information, was initially developed to map reports of violence in Kenya in 2008 and has evolved into a valuable tool during a number of projects in recent years. The potential of VGI has also been proven for urban management purposes (Song and Sun 2010), flood damage estimation (Poser and Dransch 2010), wild fire evacuation (Pultar et al. 2009) or other important cases of risk, crisis and natural disaster management (Ostermann and Spinsanti 2011, Manfré et al. 2012, Horita and De Albuquerque 2013) or responses (Goodchild and Glennon 2010, Neis et al. 2010, Bono and Gutiérrez 2011).

The reason why VGI is implemented in many of these scenarios is its open approach to data collection. VGI is sometimes the cheapest and oftentimes the only source of geo-information, particularly in areas where access to geographic information is considered an issue of national security (Goodchild 2007). The aforementioned Internet connection, essential for data contributions to VGI platforms, can be a serious caveat in developing countries, which, in combination with language issues based on the fact that many VGI services only support the Roman alphabet in the English language and large analphabetic population rates, can hinder the contributions to a VGI project in these areas (Goodchild 2007). However, despite these facts, the OSM project has developed into one of the largest and well-known VGI projects of the past seven years. Many different OSM-based maps have been created in recent years, tailored to different purposes, such as skiing, hiking or public transportation, by rendering the collected information in a particular way. More advanced projects, such as OpenRouteService.org (Neis and Zipf 2008, Schmitz et al. 2008) or OSRM (Luxen and Vetter 2011), have shown that collaboratively collected geo-information by volunteers can be a reli-

able data source for car, bicycle, pedestrian and possibly wheelchair or haptic-feedback routing or navigation applications. Based on the increasing success of the OSM project, several other companies implemented the idea of collaboratively collected or corrected geographic information for their own business solutions or data sources.

12.2.1. Comparison of recent VGI projects

In this section, we compare six different VGI projects that collect geodata in a collaborative way (Table 12.1). The comparison focuses on platforms that are providing more advanced geographical information, such as real-world features, in comparison to other VGI portals that solely share geolocated tweets or images. Furthermore, the comparison does not include other widely used LBS platforms, such as Google Maps, Telenav or Apple Maps, due to their limited VGI functionality.

The oldest VGI project in our comparison is the aforementioned OSM project. A similar project, Wikimapia, was founded in 2006 and is not related to Wikipedia in any way. The four other projects listed in Table 12.1 (Map Maker, Here Map Creator, Map Share and Waze) were established several years after OSM claimed more popularity and are partially owned by well-known proprietary geodata providers. In the past, these platforms had a limited functionality that only allowed volunteers to report an error in the map data in the form of a note (e.g. Map Share). Nowadays, registered members can create or make corrections to existing street data on almost all of the aforementioned commercial platforms. The data license and the availability of the collected information is a major difference between the compared VGI projects. In contrast to the aforementioned proprietary data providers, the OSM dataset is available under the *Open Data Commons Open Database License* (ODbL), which allows users to copy, distribute, transmit and adapt the data, as long as OSM and its contributors are credited. More importantly, if the user alters or builds upon the OSM dataset, the results can be distributed under the same license. The collected data of the OSM and Wikimapia projects can be downloaded through a designated *Application Programming Interface* (API) or as a complete database dump file.

Although the Waze project claims to have the largest contributor base, we will discuss, at a later point, that a large number of registered members does not always imply better data quantity or quality.

Table 12.1.: Comparison of VGI projects.

Attribute	Map Maker (Google)[1]	HERE Map Creator (Nokia)[2]	Map Share (Tom Tom)[3]	Waze[4]	Wiki-mapia[5]	Open Street Map[6]
Initiated in	2008	2012	2007	2008	2006	2004
Number of users or registered members [in million]	N/A	N/A	60[7]	45[9]	1.9[5]	1.3[8]
Active contributors per month in 2013	40,000[10]	N/A	N/A	12-13 million[9]	N/A	20,000[8]
Coverage (number of countries) in 2013	>220	>120	>90[11]	World	World	World
Licence	Property of Google	Property of Nokia	Property of Tom Tom	Property of Waze	CC BY-SA[12]	ODbL[13]
Data downloadable	No	No	No	No	Yes[14]	Yes

Notes: [1]http://www.google.com/mapmaker [2]http://here.com/mapcreator
[3]http://www.tomtom.com/mapshare/tools [4]http://waze.com [5]http://wikimapia.org
[6]http://www.osm.org [7]the number of enabled devices [8]http://osmstats.altogetherlost.com
[9]http://www.huffingtonpost.com/2013/05/09/facebook-waze-purchase_n_3249070.html
[10]http://google-latlong.blogspot.de/2013/04/welcoming-united-kingdom-to-google-map.html
[11]the number of countries for which map data is available [12]Creative Commons Share-Alike license [13]Open Data Commons Open Database License http://opendatacommons.org/licenses/odbl/ [14]only via a web-API

12.2.2. The OSM project

The OSM project was initiated in 2004. Its main database and web services are hosted on a number of servers at University College London. Additional server infrastructure was established through several donation rounds. All servers and interfaces to create and share OSM data are mainly developed and administered by volunteers (Elwood 2008). The project's goal is sometimes described as "building a global map" (Elwood et al. 2012); however, the main aim of the project is to build a freely available database with geographic information, which can, of course, be used for mapping purposes, but also for navigation or other applications. The *OSM Foundation* (OSMF), an international not-for-profit organization, has been established to encourage the growth, development and distribution of free geospatial data and to provide geospatial data for anyone to use and share (OpenStreetMap 2013d). Additionally, the OSMF is divided into several working groups (OSMF 2013) that support the project in specific areas of interest. For instance, the Operations Working Group plans and maintains the OSM

API and servers.

At the time of writing, the OSM project had almost 1.4 million registered members (OSMstats 2013), who contributed almost 2.1 billion points and 220 million lines, which are partially based on 3.6 billion GPS points that have been uploaded. Similar to other online communities, such as Wikipedia, only a small percentage of those volunteers actively contribute data on a regular basis, as we will discuss in more detail at a later point in this article. To be able to contribute data to the OSM project, the potential member has to register and create an account. In contrast to other VGI projects, the newly registered member can add, modify or delete geographic objects in the OSM database right after the registration process, whereas, for instance, in Google Map Maker, the edits made by new members are reviewed first. This relatively open approach to data contributions in OSM is described as one of the major benefits of the project (Neis et al. 2012a).

The collection of geo-information by online communities is oftentimes described as a bottom-up approach (Bishr and Kuhn 2007). In the case of the OSM project, however, different data contribution types need to be distinguished. In the first few years of the project, most contributors collected the geo-information by utilizing GPS-enabled handhelds. However, between 2007 and 2011, the Internet company Yahoo! (OpenStreetMap 2013k), allowed the OSM project to trace data from their satellite imagery, and a similar agreement could be established with Microsoft Bing Aerial Imagery (OpenStreetMap 2013b) in November 2010. The availability of both imagery platforms had a large impact on the collection of new objects in OSM. Particularly the release of the Bing imagery datasets resulted in a strong increase in building information (Goetz and Zipf 2012a). Additionally, several countries achieved a large data collection in OSM by importing commercial or governmental road network datasets that comply with the OSM license. Examples can be found in the Netherlands, Austria and the United States. For Spain and France, cadastral building information was successfully imported to the OSM database.

OSM contributors can use multiple ways to communicate with each other. Most of the project-related information is collected and shared in the official OSM wiki, which covers a plethora of subjects, such as detailed information about tutorials for beginners, usable software or documentation about how objects should be mapped. Additionally, contributors use a variety of *Internet Relay Chats* (IRCs; OpenStreetMap 2013f) or mailing lists (OpenStreetMap 2013g) to ask questions regarding tagging conventions in OSM, software development, data imports or other topics. Figure 12.1 illustrates the digital infrastructure of the OSM project and its community.

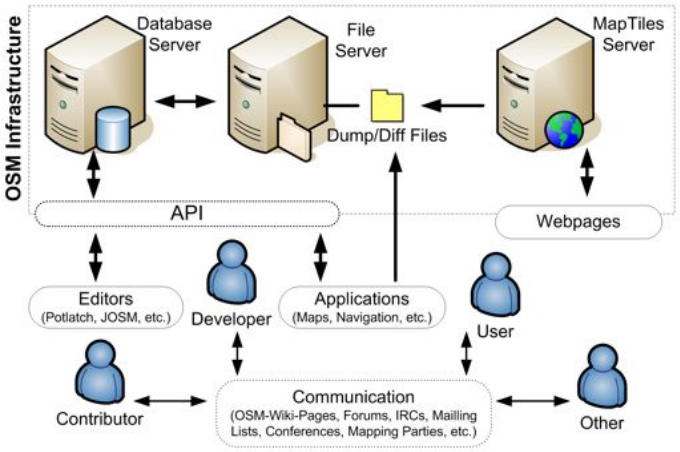


Figure 12.1.: OSM Project digital infrastructure and its community.

Many active OSM contributors also participate in so-called "Mapping Parties", during which the contributors meet at a certain location, get to know each other, share experiences about OSM and spend some time exploring and mapping the community (Budhathoki and Haythornthwaite 2013). Sometimes, these events can also take place at previously unmapped areas to improve the data collection efforts in those regions (Elwood et al. 2012). The main events in the OSM community that attract most participants are the yearly "State of the Map" conferences, which are held at different locations in the world.

To make modifications to the OSM database, the contributors can use multiple editor types. The newly developed iD editor makes it easier for new contributors to add information to the map, while the long-established editors Potlatch or JOSM (Java OpenStreetMap Editor) are preferably used by more advanced members (OpenStreetMap 2013c). There are also a large number of different editors available for multiple mobile devices, such as smartphones, and different operating systems.

The data contribution patterns of the active OSM community have changed over the past few years for different world regions. During the first few years of the project, most volunteers focused their data collection efforts on road network data. Nowadays, other real world features such as buildings, land use or public transportation information are being added in many regions to provide more details to the users. When a volunteer creates an object in the OSM database, representing a real-world feature, she/he can use three different object types (Ramm et al. 2010). Point information is represented by a Node object in OSM, whereas a Way object is utilized when mapping lines or polygons (latter, in the form of a closed line feature). If a number of Node and/or

Way objects are related to each other, the Relation object can be utilized to map this particular information (e.g. turn restrictions of the street network or tram/bus lines for public transportation). Any modifications or contributions made to the OSM database by a single member are stored in a changeset, and its extent covers the entire area within which a contributor made her or his changes. Each object in OSM can be annotated by a variety of attribute information, also referred to as tags, which consist of a key-value pair. Any contributor is free to propose and discuss new tags to describe real-world features (Haklay and Weber 2008), resulting in a bottom-up tagging approach, indicating that there is no traditionally, enforced tagging limitation with which mappers have to comply. However, a large number of suggested key-value combinations that are widely used in the community are provided in a wiki (OpenStreetMap 2013h) that helps to standardize certain objects in OSM. It also needs to be noted that a variety of map render-engines influences the creation of map "standards", due to their specific rendering functions that only allow certain features with particular tags to be shown on the map.

The collected data and created changesets can be retrieved via the OSM web-API (only for a limited extent) or as a complete database dump for the entire world. Websites, such as Geofabrik's OSM Data Extracts, also provide data downloads for a specific area of interest. It needs to be noted that these pre-processed downloads only include the latest object versions, representing the current state of the features in the map. For analytical purposes, full history dump files are available (OpenStreetMap 2013i), which include all versions of all features and allow for more advanced methods to test for potential changes between different versions in the dataset. Traditionally, the OSM datasets were provided in *Extensible Markup Language* (XML) format. To improve performance and to allow for faster processing, a binary data format, i.e. *Protocol Buffer Binary Format* (PBF), has become more common in recent years.

The success of the OSM project has been increasing in recent years, and several companies, such as Apple (OpenStreetMap 2013a), Flickr (Flickr 2013) and foursquare (Foursquare 2013), switched their mapping applications entirely or partially to OSM. Others created start-up companies that are building their solutions around OSM.

12.3. Current developments

In 2007, early questions were raised about the usefulness of VGI in science (Kuhn 2007). While a number of publications highlighted that the credibility and reliability of VGI could be questionable (Elwood 2008, Flanagin and Metzger 2008), more recently,

researchers focused on the actual quality analyses of the VGI datasets (Haklay 2010, Neis et al. 2012b, Girres and Touya 2010). These first research contributions were followed by studies about trust (Bishr and Janowicz 2010, Kessler and Groot 2013), contributor behavior (Budhathoki and Haythornthwaite 2013, Neis and Zipf 2012) and gender distributions in VGI projects (Stephens 2013, Steinmann et al. 2013b). The goal of the following three sections is to give a comprehensive and detailed overview about VGI research progress in recent years with the focus on OSM.

12.3.1. Data quality analysis

The quality assessment of geographical information follows a predefined set of quality measures and criteria. A variety of publications are available that are related to the definition of these characteristics (Brassel et al. 1995, Van Oort 2006). In 2002 the International Organization for Standardization (known as ISO) released a standard that defines the quality attributes of geodata in ISO 19113:2002 (principles for describing the quality of geographic data) and ISO 19114:2003 (framework for procedures for determining and evaluating quality). According to ISO 19113:2002, the following five parameters define the quality of geodata: completeness, logical consistency, positional accuracy, temporal accuracy, and thematic accuracy. By the end of 2013, both ISO standards (19114 and 19113) have been aggregated to one single standard: ISO 19157:2013 (geographic information data quality). In the next sections we will present several OSM research projects dedicated to one or multiple of the aforementioned spatial data quality parameters.

12.3.1.1. Road network evaluation

Over the past five years, most OSM quality analyses evaluated the OSM road network in comparison to administrative or commercial datasets. In the first analysis, Haklay (2010) compared the 2008 OSM data for Great Britain with the *Ordnance Survey* (OS) dataset, Meridian 2. To evaluate the positional accuracy, he used a buffer comparison method previously introduced by Goodchild and Hunter (1997) and Hunter (1999). The completeness of the dataset was evaluated by conducting a grid-based length comparison of the road networks. The result revealed that the OSM dataset can provide an adequate coverage for 29.3% of the area of England. One year later, in 2009, the analysis was repeated, and OSM improved its coverage to 65% (Haklay and Ellul 2011). The quality and coverage for OSM in England showed a heterogeneous pattern, with stronger road network concentrations in urban areas, but a lack

of details, such as street names, whereas rural areas at times showed a complete lack of coverage. Zielstra and Zipf (2010) utilized a similar methodology to compare the commercial TomTom Multinet dataset with OSM for Germany. They concluded that the OSM data in Germany have a similar heterogeneity as previously found between OSM and OS data in the UK in terms of its completeness, highlighting that this particular VGI source can be an alternative to commercial providers in densely populated urban areas. Rural areas, however, tend to show less coverage in OSM and are not sufficient for the creation of more advanced applications, such as route planners. A different study for Germany compared OSM data with the commercial road network dataset distributed by Navteq (Ludwig et al. 2011). The study implemented a feature matching method, which was previously introduced by Devogele et al. (1998) and Walter and Fritsch (1999). Similar conclusions as the ones previously found by Zielstra and Zipf (2010) for the OSM dataset in Germany could be made. Girres and Touya (2010) conducted a study for France by extending the previously introduced analysis by Haklay (2010), which focused on the road network, to other features, such as *Points of Interest* (POIs), waterways and coastlines. The study showed a similar heterogeneity of the OSM dataset for France as previously revealed by other researchers for other countries. However, in this particular case, some of the discrepancies can be explained by imports of different datasets, a variety in data collection methods and the participation of contributors in designated projects focusing on selected features or a predefined area. Koukoletsos et al. (2012) introduced additional methods, similar to Ludwig et al. (2011), to match a VGI source to a reference dataset, improving the data quality evaluation. The results showed that their matching procedure for the two tested areas, utilizing OSM and OS transportation network information, is efficient and that the mismatch error was below 4%.

All aforementioned studies had a strong focus on geometrical accuracy and completeness. In the following years after these initial studies, different research projects shifted this focus to other geodata quality measures. The evaluation of attribute information revealed that the removal of topological errors in OSM for Great Britain was not keeping up with newly introduced data errors in the database (Pourabdollah et al. 2013). Canavosio-Zuzelski et al. (2013) introduced a photogrammetric approach to assess and enhance the positional accuracy of the OSM street network data using stereo imagery and a vector adjustment model. In their method, they compared the road centerlines with referenced satellite imagery in the U.S. Based on several test areas, their proposed approach was able to improve the positional point accuracy and to recover the positional street displacement of OSM data. In a different study (Fairbairn

and Al-Bakri 2013), a variety of methods were applied to evaluate positional and linear geometric accuracy and area shape similarity among datasets for integration purposes in different study areas for the UK and Iraq. The researchers concluded that the integration of OSM into the official dataset caused several issues from the geometrical matching perspective. Major differences can be accredited to the varying data collection procedures in OSM. In their test areas, some of the data was remotely mapped by contributors from different countries with little to no local knowledge. Hagenauer and Helbich (2012) presented an algorithm that allowed the mining of land-use patterns from the OSM street network. This was the first approach that actively enhanced the existing or generated a new dataset based on the collected VGI data. Additionally, Helbich et al. (2012) presented a spatial statistical method to compute the positional accuracy of road junctions by extracting and comparing these particular features in OSM to a proprietary dataset.

The temporal development of the OSM dataset in Germany was analyzed in a comprehensive study for the years 2007 to 2011 (Neis et al. 2012b). The results showed that the total difference between the OSM street network for motorized traffic and a comparable proprietary dataset, i.e. TomTom, was only 9%, indicating missing data in OSM. However, when considering the entire German OSM street network, including pedestrian paths and small trails, the VGI data source exceeded the proprietary information by 27%. The same study and Scheider and Possin (2012) revealed that other important information for navigation purposes, such as turn restrictions, included in most proprietary datasets, are oftentimes missing in the OSM dataset.

In 2011, Zielstra and Hochmair (2011b) conducted one of the first OSM studies outside of Europe. Based on similar methods introduced in prior studies (Haklay 2010), they compared the OSM dataset with proprietary data from TomTom and Navteq for the entire state of Florida (USA). In contrast to prior findings in Europe, the results of this study showed an opposite trend with stronger coverage in OSM for rural areas in Florida, whereas urban areas showed better coverage in TomTom and Navteq. However, the researchers accredited this pattern to the U.S. Census TIGER/Line street data import for OSM in 2008/2009. Data imports are a highly discussed topic within the OSM community with strong opinions for and against the import of license conform datasets. Zielstra et al. (2013) analyzed the import of the TIGER/Line dataset for the entire U.S. in more detail and summarized that the community is not focusing on improving the imported dataset. This statement could be made for rural, as well as urban areas. Instead, the OSM community rather focuses on adding more detailed information, such as pedestrian trails, after a data import of all major roads took

place.

More detailed analyses for the OSM US dataset were conducted with the focus on the road networks for motorized and non-motorized traffic (Zielstra and Hochmair 2011a, Zielstra and Hochmair 2012). The analyses included computations of pedestrian routes for different data sources in selected cities in Europe and the U.S., as well as the assessment of pedestrian accessibility to transit stations. Based on the total length comparisons of the generated routes, errors of commission and omission were identified in the datasets. Due to the dense coverage of pedestrian data in German cities, better results were found for European cities in comparison to U.S. cities (Zielstra and Hochmair 2011a, Zielstra and Hochmair 2012). In a different study for the US, the development of cycling-related features in OSM and Google were investigated and compared. Results revealed a high heterogeneity with regards to completeness between analyzed cities and showed that off-road trails were more completely mapped than on-street bicycle lanes (Hochmair et al. 2013).

12.3.1.2. Evaluation of Points of Interest (POI) and other features

When considering OSM for navigation and routing purposes, it is important to implement an exact transformation of an address or textual description of a place into a geographic location. This process, referred to as geocoding, was investigated in more detail by Amelunxen (2010), who compared the results of the geocoding functionality in Google Maps with the results gathered from OSM. Nearly all requests on the municipal, street and, in particular, house number level were classified as not sufficient for detailed spatial analysis purposes in OSM, highlighting one of the most profound caveats of the project. Jackson et al. (2013) showed similar results with regards to address information in their data comparison analysis of point features in OSM and other datasets.

Other studies compared OSM land cover features, such as land use or natural areas, for several countries to pseudo ground-truth data (Mooney et al. 2010a, Mooney et al. 2010b). The analysis investigated spatial sample point characteristics of the polygons to retrieve results about cases where features are under- or over-represented (in terms of the number of points used to represent a polygon feature). Furthermore, the distance between the polygons' adjacent points was computed for quality measurements. Ciepluch et al. (2010) manually compared the spatial coverage currency and ground-truth positional accuracy of OSM in comparison to Google Maps and Microsoft Bing Maps, revealing no clear pattern in favor of any of the tested sources.

POI features take another major role in VGI data sources, such as OSM. The collec-

tion of untraditional places of interest by volunteers, not available in governmental or proprietary datasets, drives the interest of many researchers in this domain. Mashhadi et al. (2012) compared POI from Navteq and Yelp with the data collected in OSM for London (UK) and Rome (Italy) and found a highly accurate correlation in terms of geographic position. Hristova et al. (2012) used a different approach to POI analysis and tried to map different community engagements based on the contributed POI data in OSM. The results showed that spatially clustered communities produce a higher quality of coverage than those with looser geographic affinity. In a different study, the spatial-semantic interaction of OSM POI was investigated in more detail (Mülligann et al. 2011). The authors presented a semantic similarity measure that can be used to support tools and contributors in collecting and cleaning up POI data. Lastly, a study that investigated the completeness of POI in OSM in the U.S. showed that, in contrast to prior findings about the road network, the imported data from the *Geographic Names Information System* (GNIS) database was subsequently updated by the OSM community (Hochmair and Zielstra 2013).

12.3.1.3. Data trust and vandalism

Similarly to other open source related projects, Linus' Law (Raymond 1999) can have a major impact on the success of VGI. The assumption behind Linus' Law is that with the number of contributors, the quality of the product increases, an assumption which has been proven for Wikipedia, where the quality of an article increases with the number of contributors who work on it (Anthony et al. 2007, Wilkinson and Huberman 2007, Nemoto et al. 2011). While Haklay et al. (2010) found that the law generally applies to OSM positional accuracy, Mooney and Corcoran (2012c) and Mooney and Corcoran (2012a) could not identify a similar pattern when analyzing heavily edited objects in OSM.

Nearly all presented studies discussed thus far showed indications of similar data completeness or improved data quality for densely populated areas in OSM in comparison to proprietary and governmental datasets (Schilling et al. 2009, Haklay 2010, Neis et al. 2012b). Mooney et al. (2013) summarize that the OSM project proves to be heterogeneous with an urban bias and chances are that: "When one moves away from large urban centers the major issue for quality becomes one of coverage - in many rural areas there is little or no OSM coverage at all" (Mooney and Corcoran 2012c). While most conducted analyses focused on different areas in Europe and the U.S., Neis et al. (2013) investigated the aforementioned general pattern of OSM data on a larger scale. The study revealed that when comparing selected world regions, the data quality and

contributor activity does not necessarily always show the same pattern in all urban areas, particularly outside of Europe. In prior research, it had already been shown that the number of contributors can strongly influence the geometric data quality and spatial concentration of OSM data in different areas (Haklay 2010, Girres and Touya 2010); additionally, it was also determined that the temporal dataset quality is highly affected by the same criteria (Neis et al. 2013).

It has been criticized that most prior studies about OSM with the objective of a data quality analysis only consider certain object types (e.g. roads) for descriptive measurements (Hagenauer and Helbich 2012). However, other studies also highlighted the lack of attribute information, such as turn restrictions, speed limits or street names (Haklay 2010, Neis et al. 2012b, Ludwig et al. 2011), the lack of a well-defined data standard (Bishr and Kuhn 2007, Girres and Touya 2010, Brando and Bucher 2010) or some formal quality control process (Jackson et al. 2013) in OSM and VGI data in general. All of these factors lead to the questionable statement by Fairbairn and Al-Bakri (2013) that "it is probably better to have no mapping at all, rather than some inaccurate, possibly incomplete, user generated content". The best approach to answer whether OSM or any other VGI source should be utilized or not is to assess the OSM dataset quality for the selected area of interest and its particular role or purpose in the project (Haklay and Weber 2008, Goodchild 2008a, Mondzech and Sester 2011). Therefore, it is important not to look only at the completeness of the map data, but also to review the collected information in more detail, especially in areas where data imports or automated scripts took place and no active contributor community is available. Furthermore, it needs to be noted that the availability of aerial imagery in OSM editors introduces "armchair-mapping" patterns, in which case, the contributors of the project only trace objects from the satellite images and no local knowledge is needed (Neis et al. 2013). In most cases, areas with lower OSM community member numbers tend to have higher contributions based on armchair mapping, which stands in contrast to the "local knowledge" most people identify with when they refer to VGI (Goodchild 2007, Mooney and Corcoran 2013).

To simplify the evaluation of the OSM dataset for the users' areas of interest, many free and online quality assessment and assurance tools are available to get detailed quality information. Interested users are also able to report errors in the map by using OSM Notes or OpenStreetBugs. Other tools, such as Keep Right, Osmose or OSM Inspector, can be used to visualize detected errors in the map data. However, establishing some sort of trust in the collected VGI dataset is a really important factor. Several researchers presented the first approaches on how the volunteers could act as

a sort of quality measure. Bishr and Janowicz (2010) discussed a possible solution for VGI projects based on trust ratings in a social network that acts as an indicator for user reputation. In the case of the OSM project, this approach would not be suitable, due to the lack of a social network structure (Mooney and Corcoran 2012b). Kessler and Groot (2013) specified different patterns that can be used to determine trust values based on the history of contributed objects. In a second study, it was shown that for a test area in Germany, the approach can provide useful information for potential data users, even without a reference dataset, and the researchers pointed out that for further analysis, the reputation of each contributor should be considered as an important factor (Kessler et al. 2011).

Although the OSM project showed some promising developments in recent years, the increasing popularity also comes with a number of caveats, especially in the form of cases of vandalism, similar to developments seen in Wikipedia (Potthast et al. 2008). While Coleman (2010b) summarized some of the first methods on how to validate contributors and their spatial information, Neis et al. (2012a) developed the first prototype to automatically detect vandalism in VGI projects and revealed that within a timeframe of one week, at least one case of vandalism or accidentally destroyed objects by new or inexperienced members can be detected in the OSM database each day.

12.3.2. Contributor analysis

A second large spectrum of VGI research that experienced a strong increase of interest in the research community in recent years is dedicated to the contributor behavior of projects, such as Wikipedia or OSM. In many publications the voluntary members are titled as "users" (Stephens 2013, Mooney and Corcoran 2013). However, in the context of this paper, we want to distinguish between users (who use the data or online information), registered members (who have an account with the VGI project) and contributors (who actively contribute to a VGI project). The reason for this precise classification lies in the fact that the number of users does not reflect the actual number of active contributors and neither does the number of passive members that are only registered to the project, but never actively contribute. It is nearly impossible to determine the actual number of OSM users, since not every single user that implements the OSM data in a project or uses it in an application on his or her handheld device needs to be registered with the project. Merely the number of registered members can be determined from the OSM database and further processed for analysis.

12.3.2.1. Participation inequality

Similarly to other open source-related or online community-based projects, VGI platforms experience a so-called "participation inequality". Nielsen (2006) describes this phenomenon with a 90-9-1 rule, representing the 90% of users who never contribute to the project and merely function as "lurkers", the 9% of contributors that add information on an irregular basis and the 1% of contributors that account for almost all the collected information of the project. This phenomenon has been identified for Wikipedia (Wilkinson and Huberman 2007, Javanmardi et al. 2009), as well as for OSM (Neis and Zipf 2012, Budhathoki 2010) in similar ways. The activity of the project's community has a major impact not only on the collection of new geographic data, but also on timeliness of existing datasets. In the context of UGC platforms, the widely used online encyclopedia, Wikipedia, had almost 20 million registered members at the beginning of October, 2013, of which, a total of 1.7 million members (9%) had edited at least one article, but only 125,000 members (0.7%) had performed an action to articles in September, 2013 (Wikipedia 2013). It needs to be noted, however, that these numbers do not include changes made by unregistered, anonymous contributors.

Similar patterns can be found in VGI projects, such as OSM. In 2008, about 10% of the 30,000 registered members actively contributed to OSM (Ramm and Stark 2008). This positive trend continued in the following year (2009) for which a study had shown that, in total, about 33,400 (28%) of the 120,000 registered members edited data for the project (Budhathoki 2010). In 2010 the number of registered members increased to 300,000, of which almost 5% (16,500) actively contributed to the project on a monthly basis and only 3.5% of all members (12,000) accounted for 98% of the data volume (Neis et al. 2012b). In a more recent study, it was shown that 38% of the 500,000 registered OSM members edited at least one object of the projects dataset in 2011 (Neis and Zipf 2012). Figure 12.2 illustrates the growth of registered members and their corresponding activity over the past eight years. It also highlights the strong discrepancy between the number of registered members and the active contributors who created a changeset. Additionally, Figure 12.2 illustrates the significant difference between the number of "one-time-only" contributors (Coleman et al. 2009) who created only one changeset and contributors who performed several edits in the OSM database.

The conducted research by Neis and Zipf (2012) also analyzed the contributor activities by day of week and time of day. Almost all weekdays showed similar contribution patterns; only on Sundays did the project prove to have a slightly larger number of data edits, while the afternoon and evening hours were identified as time ranges with the highest activity in OSM for each day.

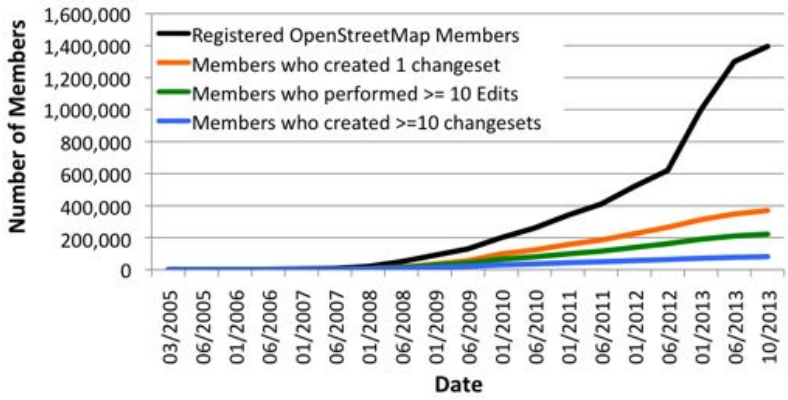


Figure 12.2.: Growth of OpenStreetMap membership numbers between 2005 and 2013.

Overall, the discussed results of the OSM project match similar patterns previously found in Wikipedia (Yasseri et al. 2012) or mobile phone communication behavior (Jo et al. 2012). Neis and Zipf (2012) also identified different member groups in OSM based on the number of contributions the members made to the project and revealed that only around 5% of all members contributed in a significant way. Although the absolute number of registered members is still increasing, the relative number of active members has been decreasing in the past two years. In 2012, only around 3% of all registered members made a contribution each month; at the end of 2012, only 18,000 (1.8%) of the one million registered members actively contributed any data. This negative trend continued in 2013. As of October, 2013, the OSM project had almost 1.4 million registered members and the number of active contributors in that month was only around 22,000 (1.6%) (OSMstats 2013). The negative trend in the relative contribution share is mainly influenced by the high amount of newly registered members in 2013 (Figure 12.2).

Based on these prior findings, a number of studies questioned the long-term motivation of the contributors of the project (Coleman et al. 2009, Neis et al. 2012b, Mooney and Corcoran 2013). When analyzing all created changesets of the OSM project, the increase in monthly volunteer numbers over the past few years and the consistencies in data contributions can be visualized as shown in Figure 12.3. Only half of the active members in OSM that contributed in the month of October (2013) are also long-term contributors and registered before or during 2012. A clear pattern can also be seen for the years 2008 to 2012, with almost 70% of contributor loss over the following years, stopping significant data edits and object changes in OSM.

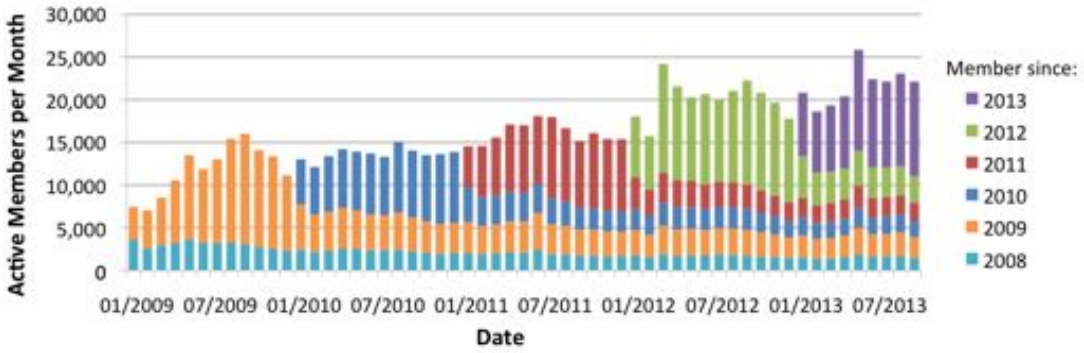
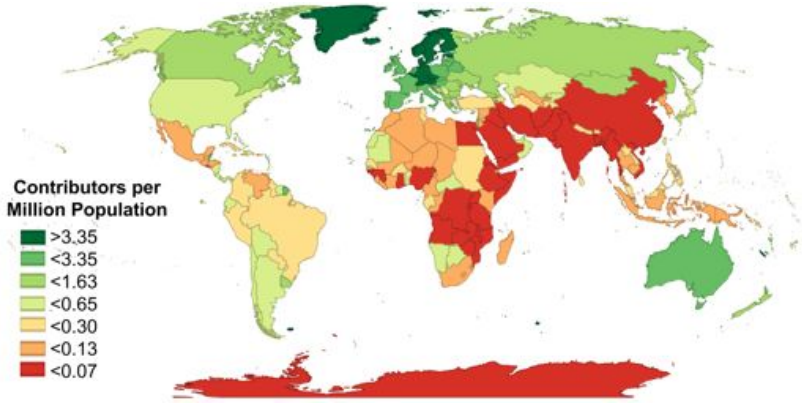


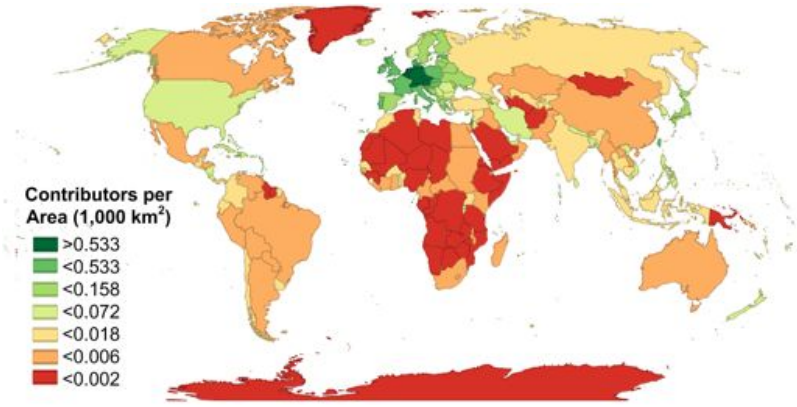
Figure 12.3.: Active contributors per month between 2009 and 2013.

12.3.2.2. Areal distribution

The areal distribution of the active OSM community members shows a similar heterogeneous pattern as the aforementioned data quality and quantity analyses. Since an OSM member does not have to provide his or her location information when registering to the project, Budhathoki (2010) and Neis and Zipf (2012) localized the members by utilizing different approaches. Budhathoki (2010) analyzed the number of added Nodes per country for each contributor, whereas Neis and Zipf (2012) focused on the first edit of a contributor or the area which shows the most activity, to identify the origin of the project members. Both studies showed similar results for the years 2010 and 2012 in which three-quarters of the contributors were located in Europe. The remaining quarter was distributed over North America and Asia. South America, Africa and Oceania proved to have only a small contributor number. When considering the population density of the different countries, it is surprising to see that the USA, China or India only show relatively small project contributor numbers, which can be caused by a number of reasons, such as other freely available datasets, such as the TIGER/Line data for the U.S., or governmental restrictions that make the collection of geodata illegal in certain countries. Additionally, Neis et al. (2013) illustrated that other factors next to population density or income must have an influence on contributor growth. The highest concentration of active contributors in OSM can be found in Germany. Out of the 2,500 daily contributors, around 550 members (22%) edited data in Germany (OSMstats 2013). Thus, it is not surprising to see that the German OSM dataset also shows a higher quality. Figure 12.4 illustrates the distribution of active OSM contributors per day related to population in millions (a) and area (b) for each country, highlighting the strong concentration of OSM contributors in Europe (b).



(a)



(b)

Figure 12.4.: Distribution of active OSM contributors per day and per population in million (a) and per area (1,000 km^2) (b) (August 1 - October 31, 2013).

The OSM contributors, however, are not just limiting their data collection efforts to their home regions. Budhathoki (2010) and Neis and Zipf (2012) revealed that a smaller number of highly active OSM members collect data in at least two or more countries.

12.3.2.3. Motivation, behavior and gender dimensions

A number of studies in recent years also provided more insight about the discrepancies in contributor motivation, behavior or gender dimensions in VGI projects, such as OSM. In most cases, an extensive survey was conducted to evaluate the most detailed information about the contributors of the project. Budhathoki (2010) and Budhathoki and Haythornthwaite (2013) showed in a comprehensive study which criteria increase

the contributor motivation in VGI projects. Similar studies were conducted with focus on UGC platforms, such as Wikipedia or other Open Source Software development communities (Coleman et al. 2009), whereas Steinmann et al. (2013a) compared the motivating factors for several online portals, such as OSM, Google Map Maker, Foursquare, Panoramio, Facebook and Wikipedia. Furthermore, the different factors have been classified by the authors in intrinsic and extrinsic or constructive and negative classes. Table 12.2 provides an overview of the classification schemas based on the aforementioned studies.

Table 12.2.: VGI motivating factors (paraphrased from Budhathoki (2010) and Coleman et al. (2009)).

Constructive Side		Negative Side
Intrinsic	Extrinsic	
Altruism	Social reward/relations	Mischief/vandalism
Fun/recreation	Career	Malice or criminal intent
Learning/personal enrichment	Personal reputation	
Unique ethos	Community/project goal	
Self-expression/image	System trust	

The motivation of the different project members was one of the main criteria the researchers investigated in their studies. However, demographic factors, such as age, gender or educational background of the project's participants, were also analyzed in more detail. The results showed that the majority of OSM contributors, i.e. more than 97%, were males (Budhathoki and Haythornthwaite 2013, Stark 2010, Lechner 2011). For other online portals, such as Foursquare or Facebook, the participant data did not show this biased gender distribution; however, the female participation rate dropped substantially when geographic information was introduced, for instance during the geotagging procedure for images or posts on the social networking platforms (Stephens 2013, Steinmann et al. 2013b). The comparison of contributor gender distribution between different mapping platforms showed that women and men contribute to Google Map Maker at equal rates, whereas the number of female contributors significantly dropped when considering contributions to OSM (Stephens 2013). However, these findings differ from the results by Steinmann et al. (2013b), where OSM and Google Map Maker showed equally low female participatory values.

The analysis of the OSM contributor age distribution revealed that the majority of contributors (>60%) are between 20 and 40 years old, and about 20% of the mappers are 40 years or older (Budhathoki and Haythornthwaite 2013, Stephens 2013). The

contributors also provided information about their educational background during the surveys, and the results showed that 63-78% had a college, university or higher education degree (Budhathoki and Haythornthwaite 2013, Stephens 2013, Lechner 2011).

Many research articles stated in the past that VGI is mostly contributed by non-experts (Bishr and Kuhn 2007) or volunteers who are untrained and unqualified (Haklay 2010, Flanagin and Metzger 2008, Goodchild 2008b). Janowicz and Hitzler (2010) summarized that VGI is collected and edited by a heterogeneous online community with different backgrounds and application areas in mind. However, the conducted surveys did not support these statements entirely (Stephens 2013, Budhathoki 2010, Lechner 2011). Almost 50% of the respondents of each survey had degrees or worked in the fields of Geography, Geomatics, Urban Planning or Computer/Information Sciences, highlighting that the OSM community does not necessarily only constitutes of GIS amateurs, as is oftentimes speculated (Budhathoki 2010).

Next to the aforementioned socioeconomic factors, the social interaction between contributors in OSM and their individual contribution patterns played a major role in a number of research articles. Oftentimes, only a few contributors collect the majority of data volume in a predefined area (Mooney and Corcoran 2012a) or entire country (Neis et al. 2012b). Mooney and Corcoran (2012b) and Mooney and Corcoran (2012d) attempted to identify explicit social networks between contributors based on this assumption. The results showed that most collaboration in OSM is purely incidental and that data contributions are mainly done in isolation.

In 2009, a first attempt was made to classify VGI project members into different overlapping contributor categories, such as Neophyte, Interested Amateur, Expert Amateur, Expert Professional and Expert Authority (Coleman et al. 2009). For OSM, the contributors were oftentimes classified or sorted based on their contribution patterns. Mooney and Corcoran (2012a) and Neis and Zipf (2012) grouped the members of the OSM project based on the number of contributions or created objects to analyze the different groups in more detail. Mooney and Corcoran (2012b) classified the OSM contributors of the London (UK) area into four distinct groups, highlighting that the majority of the contributors edited the geometry of objects or their corresponding attributes, but in many cases not both. Steinmann et al. (2013a) utilized a clustering method based on the contribution and feature types of each contributor to identify different patterns in mapping behavior. As a result, contribution profiles, such as "Premium Creator", "Highway Mapper" or "All-Rounder" were created.

12.3.3. Additional developments

Many published research articles in recent years did not intrinsically analyze OSM data quality or contributor patterns, but utilized the available dataset in a number of applications to investigate the fitness for the purposes of the contributions. With the increasing popularity of 3D applications, researchers tested the applicability of OSM for 3D applications or 3D location based services (Schilling et al. 2009). In the following years, the first publications suggested how the OSM schema could be extended to indoor environments (Goetz and Zipf 2011). Other suggested methods on how to transform OSM data to the standardized *City Geography Markup Language* (CityGML) models (Goetz and Zipf 2012a) or for indoor evacuation simulations (Goetz and Zipf 2012b).

In particular, the potential of OSM data for routing or trip-planer applications attracted a high interest in the research community. Neis and Zipf (2008) presented a first approach on how OSM data can be utilized for routing and address- or POI allocations. Luxen and Vetter (2011) improved the OSM routing performance with an open source mobile and server route planning application utilizing a contraction hierarchies method that enabled faster route computations (Geisberger et al. 2008). OSM data was also implemented in robot tasks and autonomous vehicle applications (Hentschel and Wagner 2010), whereas others augmented OSM route network data with *Shuttle Radar Topography Mission* (SRTM) height information to compute the optimal path for electric vehicles (Eisner et al. 2011).

The open approach to data contributions in OSM allows for the development of a plethora of applications, online or printable maps tailored to particular needs, such as hiking, biking, skiing or public transportation. The detailed data requirements for routing applications tailored to disabled people, such as pavement width or surface conditions, can be added to OSM and annotated with a selected number of tags. First studies introduced how OSM and its tagging schema can be utilized for applications tailored to wheelchair users (Holone et al. 2007, Rashid et al. 2010) or visually impaired pedestrians that utilize haptic-feedback (Jacob et al. 2010). The Wheelmap project[1] is a great example for this particular case. Any volunteers can mark locations with wheelchair-friendly environments or accessibility. The results and detailed information of the Wheelmap project is then saved to the OSM database.

A second large potential of OSM and other VGI-related projects lies in their support function on decision making processes during disaster management. Up-to-date

[1] www.wheelmap.org

geodata that includes detailed accessibility information for particular crisis regions can be of crucial importance during the relief response operations of organizations, such as the Red Cross. The data of the OSM project can be utilized in many ways during these events, due to its fast data processing methods and timely map updates. Additionally, the conversion of OSM data to Shapefiles, or other source files for handheld GPS devices, helps to develop tailored LBS applications and run spatial analysis tasks, for instance during natural disaster events. The success of OSM during these events has been proven during a devastating earthquake in Haiti in 2011, a tsunami in Japan in 2011 and after typhoon Yolanda in the Philippines in 2013. Contributors of the OSM project helped to collaboratively collect geodata for the crisis areas. At the latter event, more than 1,500 contributors from 80 countries made more than 3.8 million map changes within 15 days (OpenStreetMap 2013j). This was a significant increase in member activity in comparison to the prior events, where almost 700 (Haiti) and 350 (Japan) contributors collected information for the affected regions. Auer and Zipf (2009) also demonstrated that Open Data and Open Standards can help reduce costs, whereas Goetz et al. (2012) presented a workflow to develop an online map completely based on OSM data.

12.4. Future trends

This extensive review of previous research articles has shown that VGI and, especially, OSM have experienced a strong increase of interest within the research community over the past few years. Although an impressive number of research projects focused on the quality assessment and contributor analysis of VGI, questions remain about different research domains when considering the voluntarily contributed datasets.

Most VGI data quality analyses in the past were conducted using a commercial or governmental reference dataset of high quality. While the general question remains whether these proprietary datasets can really be considered as more accurate than VGI, or if the opposite situation should be considered during the analysis, others focus on intrinsic data evaluations if no reference dataset exists or is not available due to high costs or other factors. In the case of OSM, these intrinsic approaches included the evaluation of the number of edits or the number of contributors in a predefined area (Neis et al. 2012b, Haklay et al. 2010, Neis et al. 2013). However, more research needs to be conducted to evaluate the effectiveness of these studies. If a number of contributors in OSM stopped collecting a certain feature type in a predefined area and starts collecting other information, does this imply that this particular object

type is completely mapped in the area of interest or are there other criteria that can play a role? Additionally, particularly for quality assessment analyses, it would be necessary to have more detailed information about how the data was collected. Did the contributor use a GPS-device, the available areal imagery or is her/his contribution based on a data import? Although the OSM project provides several tags to specify the source or contribution type, the overall usage of those key/value-pairs is only limited within the community.

Similarly, the discussion about trust in VGI is ongoing. Several studies have shown that VGI data can be used as an alternative to commercial or proprietary datasets and in the case of OSM, different companies already switched their mapping products to the freely available dataset. The credibility and trust in VGI plays a major role in these cases, but how can a VGI project, such as OSM, with no strict data specification or quality control, establish some type of trust? None of the previously conducted research projects considered one of the most important components of a VGI project: the individual contributor reputation. How can this reputation of a contributor be computed to provide a better understanding about the quality and trust of the data? What parameters are necessary to assess the quality of the contributions? Some first quantitative parameters, such as the amount of created or edited objects, have been investigated in prior studies to give some first insights. However, for a sophisticated trust estimation of a collaboratively collected dataset, other factors, such as the home location of a contributor, the mapping behavior, especially with regards to the usage of tags that represent a special area of interest of the contributor, or her/his acceptance and reputation within the community can play major roles. Would a contributor rating and reputation system, as discussed by Bishr and Kuhn (2007) and Flanagin and Metzger (2008), be useful to calculate the credibility and trust values as a proxy for VGI quality assessments?

In one of the aforementioned studies, a method was introduced that created contributor profiles such as “Highway Mapper” or “Building Mapper”, based on the added information of the contributor (Steinmann et al. 2013a). A study about Wikipedia contributors revealed, however, a relation between the quality of an article and its authors and concluded that it is more important who contributes to the article and not as much what type of information was added (Stein and Hess 2007). Similar methods that separate contributors by their reputation and experience rather than the type of information they contribute are required for VGI projects.

Neis et al. (2013) proved in their study that for some areas, the majority of OSM data was collected by external contributors, whose home location was more than 1,000 km

away from the area in which the data was contributed. It can be assumed that these data contributions are made through tracing aerial imagery, and prior studies have shown that this mapping behavior can lead to an overrepresentation in the geometry of a feature or to missing feature descriptions, such as street names or other information (Mooney et al. 2010a, Mooney et al. 2010b, Neis et al. 2013). Thus, more detailed analyses are needed to determine whether external or remote members provide data with a better, equal or worse quality when contributing to the project. Additionally, Comber et al. (2013) revealed that contributors oftentimes add information on different scales due to the resolution of the aerial imagery that is available. Similar to the work by Touya and Brando-Escobar (2013), who presented a first approach to calculate the level-of-detail of different OSM features, it needs to be investigated how the scale of the contributed geo-information can potentially be used for VGI quality assessment.

The geographic scope of prior VGI analyses had a strong focus on areas in Europe and, to a lesser extent, the United States. One of the aforementioned studies (Neis et al. 2013) highlighted in a first comparison of selected urban areas of the world that factors, such as population density and income, can have an influence on data contributions and community efforts in the different selected regions. Goodchild (2008a) also pointed out that the digital divide could highly influence the mapping activities in less developed countries. Thus, a strict distinction between the developments of VGI in different geographic world regions and the analysis of potential socio-economic or cultural influential factors could give a better understanding about the individual motivations to contribute to a VGI project for each world region.

Due to the discrepancies in contributor concentration around the world in VGI projects, such as OSM, sometimes, the particular area of interest does not contain the required data or data types that the user would like to implement in a desired project. Hagenauer and Helbich (2012) demonstrated that based on the availability of different feature types in OSM, other missing geographic objects, such as land use information, can be derived. Future research could investigate this process in more detail to see what type of additional geographic features can be derived based on existing objects in the dataset. A different approach that combines and enriches current datasets through data conflation of multiple sources has been the focus of many researchers in recent years. It needs to be noted that licensing conventions, such as the ODbL license in OSM, can be a hindrance in these cases, due to the limitations of the license, whereas other VGI data sources with fewer restrictions, such as geolocated images on Flickr or Panoramio or tweets on Twitter, could help to improve proprietary or governmental datasets (Roick and Heuser 2012).

One of the major concerns that arose in recent years about the OSM project is the lack of detailed and complete address information. Adding this information is a tedious process that does not give contributors the same type of satisfaction as the collection of roads and buildings that are visualized in the projects map. However, how can this issue of missing or incomplete data in VGI projects be solved? Several companies have been supporting the development of a number of tools to display incorrectly mapped information in OSM or that help the contributor to simply collect the required data types. A study has shown that services operating on OSM have a regulative and quality assuring effect (Schmitz et al. 2008). The study, conducted in 2008, showed that the number of topology network errors was reduced after an online route planner based on OSM data was available. Similar approaches have been discussed under the term "gamifaction" to engage new or established contributors of the OSM project to solve errors in the collected dataset in the next few years. A popular example in this domain is the Kort Game, a mobile web-app to repair OSM data, which already showed that OSM contributors and other volunteers are willing to enter missing information, such as the name of a POI, to the dataset.

Due to several data imports, automated data edits and software issues in OSM, it is very important that researchers consider who created, modified or deleted the data in the area of interest during their analyses. Features and objects that were created during a data import, modified by an automated script (bot) or deleted by accident or as an act of vandalism do not represent the general pattern of the dataset and need to be highlighted or excluded during the analysis. Neis and Zipf (2012) stated that before 2011, a software error in one of the OSM editors increased the version number of each object, which falls into the extent of a certain changeset, although it was not changed by the contributor, a major error that especially needs consideration in studies that utilize the OSM full history dump files. Zielstra et al. (2013) showed in their study that the majority of motorized traffic-related data contributions in the U.S. are based on data imports or were changed by automated edits. Unless it is the purpose of the study to identify these patterns, researchers need to be aware of the potential errors that are caused by these procedures.

The detection of the aforementioned vandalism cases in VGI projects has been previously investigated for other UGC-related projects, such as Wikipedia (Potthast et al. 2008). Although only a small number of vandalism cases were detected in OSM in recent years, a study (Neis et al. 2012a) revealed that the number of vandalism cases correlates with the popularity increase of the OSM project. Thus, the need for methods and designated tools that provide secure VGI vandalism detection will spark some of

OSM-related future research. Goodchild and Li (2012) proposed that VGI communities should implement data control methods which could be based on a review system similar to the "edit-reviewer" function in Google Map Maker, in which contributions of newly active members are checked for their eligibility (Elwood et al. 2013). In the case of OSM, the question remains if there are enough volunteers available that are willing to work on manual data validation tasks to approve data edits made by unknown contributors (Neis et al. 2012a). Based on the findings of prior studies, this could be a difficult task, since most contributors restrict their edits and updates to their own collected data (Mooney and Corcoran 2012b, Mooney and Corcoran 2012d).

A number of significant questions about the long-term motivation and the future of VGI platforms and their contributors have been asked by many in recent years. Based on the presented results in this article about the OSM community, we can say that at least every third contributor, of all the active contributors that ever added information, will continue to contribute over several years. Future research could reveal what type of information long-term members contribute. Do they collect new data in different areas or do they start collecting more detailed information, such as trees or sidewalk and surface information in their home area? On the other hand, it would be interesting to know what demotivates contributors and makes them stop contributing data to the project. One possible answer could be that there is "no-more interesting work" left to do (Mooney and Corcoran 2013). Besides the motivation of current contributors, others raise questions on how new volunteers can be attracted to VGI projects (Coleman 2010a). In the past three months (August to October, 2013), OSM increased by 1,000 new members each day, of which 150 actively started contributing. Compared with prior findings, these numbers reveal a decreasing pattern in the number of registered members per day (Neis et al. 2012b). However, the number of newly active members is identical between the years 2011 and 2013.

12.5. Conclusion

UGC and VGI online platforms have developed into a well-known phenomenon on the Internet in recent years. The review of previously conducted research in the realm of VGI in this article has clearly shown that the research community sees a lot of potential in the freely available data sources. OSM, with its exceptionally large community of registered and active contributors, demonstrates that collaboratively collected geographic information by volunteers around the world can lead to an impressive data source for multiple applications. In our study, we provided a comprehensive overview

about the recent research projects in the field of VGI with a strong focus on the OSM project. The research efforts have been separated into two main domains: Data Quality and Contributor Analysis. A detailed discussion of the latest literature for each domain highlights the methodologies and findings for each study. VGI data quality analyses sparked the interest of the research community in the first few years after the platforms attracted more attention, resulting in a large body of studies in this domain. In more recent years, the analysis of contributor behavior, motivation and gender distribution in VGI projects has experienced more attention in the research community.

Many VGI quality analyses have demonstrated in the past that the freely available data can be used for a variety of applications. However, it is still an important and major task to evaluate whether the data is acceptable for each use case. The quality of VGI datasets, as has been proven for OSM, can be heterogeneous when considering different countries or discrepancies between rural and urban areas. There is no reliable estimation if a certain object or other detailed attribute information is included in a VGI dataset, unless the potential user investigates the data in more detail and compares it to ground truth information or a reference dataset of choice. The long term motivation of the volunteers that contribute to VGI projects, which has been questioned in recent years, has been proven, at least for the OSM project in this study.

Based on the most recent developments in VGI research, we were able to discuss potential future trends in research and development, especially for the case of OSM. The intrinsic data assessment approach can be utilized for countries where no reference dataset is available or potential costs are too high to acquire the datasets needed. However, new methods need to be developed that utilize this approach and potentially include multiple VGI datasets. Similarly, conflation methods that utilize several VGI sources or combine VGI with other license conform datasets could be helpful in the near future. To make VGI more respected and eligible for these tasks, however, questions about credibility and trust in VGI datasets, with a focus on the contributor, need to be answered. Thus, the development of new methods that compute a trust factor, contributor reputation or individual contributor data quality is required. Finally, additional surveys are needed to gather more information about the differences in contributor motivation and behavior, especially when considering different continents and cultures.

References

Amelunxen, C. (2010). An Approach to Geocoding Based on Volunteered Spatial Data. In: *Geoinformatik 2010*. (Mar. 17–17, 2010). Kiel, Germany.

Anderson, P. (2007). What is Web 2.0? Ideas, Technologies and Implications for Education. In: *JISC*.

Anthony, D., Smith, S. W., and Williamson, T. (2007). The Quality of Open Source Production: Zealots and Good Samaritans in the Case of Wikipedia. In: *Dartmouth Computer Science Technical Report TR2007-606*. Hanover, NH: Dartmouth College.

Auer, M. and Zipf, A. (2009). How do Free and Open Geodata and Open Standards fit together? From Sceptisim versus high Potential to real Applications. In: *Proceedings of the first Open Source GIS UK Conference*. (June 22, 2009). Nottingham, UK.

Bishr, M. and Janowicz, K. (2010). Can we Trust Information? - The Case of Volunteered Geographic Information.

Bishr, M. and Kuhn, W. (2007). Geospatial Information Bottom-Up: A Matter of Trust and Semantics. In: *The European Information Society - Leading the Way with Geo-information*. Berlin Heidelberg, Germany: Springerpp. 365–387.

Bishr, M. and Mantelas, L. (2008). A Trust and Reputation Model for Filtering and Classification of Knowledge about Urban Growth. *GeoJournal*, 72, 229–37.

Bono, F. and Gutiérrez, E. (2011). A network-based analysis of the impact of structural damage on urban accessibility following a disaster: the case of the seismically damaged Port Au Prince and Carrefour urban road networks. *Journal of Transport Geography*, 19, 1443–1455.

Brando, C. and Bucher, B. (2010). Quality in User-Generated Spatial Content: A Matter of Specifications. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

Brassel, K., Bucher, F., Stephan, E., and Vckovski, A. (1995). Completeness. In: *Elements of Spatial Data Quality*. Oxford, UK: Elsevier81–108.

Budhathoki, N. (2010). Participants' Motivations to Contribute to Geographic Information in an Online Community. Ph.D. Dissertation. University of Illinois, Urbana-Champaign, Urbana, IL, USA.

Budhathoki, N. and Haythornthwaite, C. (2013). Motivation for Open Collaboration: Crowd and Community Models and The Case of OpenStreetMap. *American Behavioral Scientist*, 57, 548–575.

Canavosio-Zuzelski, R., Agouris, P., and Doucette, P. (2013). A Photogrammetric Approach for Assessing Positional Accuracy of OpenStreetMap Roads. *ISPRS International Journal of Geo-Information*, 2, 276–301.

Ciepluch, B., Jacob, R., Mooney, P., and Winstanley, A. (2010). Comparison of the Accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps. In: *Proceedings of the Ninth International Symposium on Spatial Accuracy Assessment in Natural Resuorces and Enviromental Sciences*. (July 20–23, 2010). Leicester, UK.

Coleman, D. (2010a). The Potential and Early Limitations of Volunteered Geographic Information. *Geomatica*, 64, 27–39.

Coleman, D. (2010b). Volunteered Geographic Information in Spatial Data Infrastructure: An Early Look At Opportunities And Constraints. In: *Spatially Enabling Society: Research, Emerging Trends and Critical Assessment*. Leuven, Belgium: Leuven University Press.

Coleman, D., Georgiadou, Y., and Labonte, Y. (2009). Volunteered Geographic Information: The Nature and Motivation of Producers. *International Journal of Spatial Data Infrastructures Research*, 4, 332–358.

Comber, A., See, L., Fritz, S., Velde, M. van der, Perger, C., and Foody, G. (2013). Using Control Data to Determine the Reliability of Volunteered Geographic Information About Land Cover. *International Journal of Applied Earth Observation and Geoinformation*, 23, 37–48.

Devogele, T., Parent, C., and Spaccapietra, S. (1998). On Spatial Database Integration. *International Journal of Geographical Information Science*, 12, 335–352.

Eisner, J., Funke, S., and Storandt, S. (2011). Optimal Route Planning for Electric Vehicles in Large Networks. In: *Proceedings of the 25th Conference on Artificial Intelligence*. (Aug. 7, 2011–Aug. 11, 2010). San Francisco, CA, USA.

Elwood, S. (2006). Critical Issues in Participatory GIS: Deconstructions, Reconstructions, and new Research Directions. *Transaction in GIS*, 10, 693–708.

Elwood, S. (2008). Volunteered Geographic Information: Future Research Directions Motivated by Critical, Participatory, and Feminist GIS. *GeoJournal*, 72, 173–183.

Elwood, S. (2010). Geographic Information Science: Emerging Research on the Societal Implications of the Geospatial Web. *Progress in Human Geography*, 34, 349–357.

Elwood, S., Goodchild, M.F., and Sui, D. (2013). Prospects for VGI Research and the Emerging Fourth Paradigm. In: *Crowdsourcing Geographic Knowledge. Volunteered Geographic Information (VGI) in Theory and Practice*. Berlin/Heidelberg, Germany: Springer.

Elwood, S., Goodchild, M., and Sui, D. (2012). Researching Volunteered Geographic Information (VGI): Spatial Data, Geographic Research, and new Social Practice. *Annals of the Association of American Geographers*, 102, 571–590.

Fairbairn, D. and Al-Bakri, M. (2013). Using Geometric Properties to Evaluate Possible Integration of Authoritative and Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, 2, 349–370.

Flanagin, A. J. and Metzger, M. J. (2008). The Credibility of Volunteered Geographic Information. *GeoJournal*, 72, 137–148.

Flickr (2013). *Around the world and back again*. URL: `http://blog.flickr.net/en/2008/08/12/around-the-world-and-back-again` (visited on 09/23/2013).

Foursquare (2013). *foursquare is joining the OpenStreetMap movement! Say hi to pretty new maps!* URL: `hhttp://blog.foursquare.com/2012/02/29/foursquare-is-joining-the-openstreetmap-movement-say-hi-to-pretty-new-maps` (visited on 09/23/2013).

Geisberger, R., Sanders, P., Schultes, D., and Delling, D. (2008). Contraction Hierarchies: Faster and Simpler Hierarchical Routing in Road Networks. In: *Proceedings of the 7th Workshop on Experimental Algorithms, Volume 5038 of Lecture Notes in Computer Science*. Springer.

Girres, J.F. and Touya, G. (2010). Quality Assessment of the French OpenStreetMap Dataset. *Transaction in GIS*, 14, 435–459.

Goetz, M. (2012). Using Crowd-sourced Indoor Geodata for the Creation of a Three-dimensional Indoor Routing Web Application. *Future Internet*, 4, 575–591.

Goetz, M., Lauer, J., and Auer, A. (2012). An Algorithm Based Methodology for the Creation of a Regularly Updated Global Online Map Derived From Volunteered Geographic Information. In: *Proceedings of the fourth International Conference on Advanced Geographic Information Systems, Applications and Services*. (Jan. 30–Feb. 4, 2012). Valencia, Spain.

Goetz, M. and Zipf, A. (2011). Extending OpenStreetMap to Indoor Environments: Bringing Volunteered Geographic Information to the Next Level. In: *Urban and Regional Data Management: Udms Annual 2011*. Delft, The Netherlands.

Goetz, M. and Zipf, A. (2012a). Towards Defining a Framework for the Automatic Derivation of 3D CityGML Models from Volunteered Geographic Information. *International Journal of 3-D Information Modeling*, 1, 1–16.

Goetz, M. and Zipf, A. (2012b). Using Crowdsourced Indoor Geodata for Agent-Based Indoor Evacuation Simulations. *ISPRS International Journal of Geo-Information*, 1, 186–208.

Goodchild, M.F. (2007). Citizens as Sensors: The World of Volunteered Geography. *GeoJournal*, 69, 211–221.

Goodchild, M.F. (2008a). Assertion and Authority: The Science of User-generated Geographic Content. In: *Proceedings of the Colloquium for Andrew U. Frank's 60th Birthday, Department of Geoinformation and Cartography.* (June 30–July 1, 2008). Vienna, Austria.

Goodchild, M.F. (2008b). Spatial Accuracy 2.0. In: *Proceedings of the 8th International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences.* (June 25–27, 2008). Shanghai, China.

Goodchild, M.F. and Glennon, J.A. (2010). Crowdsourcing Geographic Information for Disaster Response: A Research Frontier. *Internationl Journal of Digital Earth*, 3, 231–241.

Goodchild, M.F. and Hunter, G.J. (1997). A Simple Positional Accuracy Measure for Linear Features. *International Journal of Geographical Information Science*, 11, 299–306.

Goodchild, M.F. and Li, L. (2012). Assuring the Quality of Volunteered Geographic Information. *Spatial Statistics*, 1, 110–120.

Hagenauer, J. and Helbich, M. (2012). Mining Urban Land Use Patterns from Volunteered Geographic Information by Means of Genetic Algorithms and Artificial Neural Networks. *International Journal of Geographical Information Science*, 26, 963–982.

Haklay, M. (2010). How good is OpenStreetMap Information: A Comparative Study of OpenStreetMap and Ordnance Survey Datasets for London and the Rest of England. *Environment and Planning B*, 37, 682–703.

Haklay, M., Basiouka, S., Antoniou, V., and Ather, A. (2010). How many Volunteers does it take to Map an Area well? The Validity of Linus' Law to Volunteered Geographic Information. *The Cartographic Journal*, 47, 315–322.

Haklay, M. and Ellul, C. (2011). *Completeness in Volunteered Geographical Information - The Evolution of OpenStreetMap Coverage in England (2008–2009).* URL: `http://povesham.wordpress.com/2010/08/13/completeness-in-volunteered-geographical-information-%E2%80%93-the-evolution-of-openstreetmap-coverage-2008-2009/`.

Haklay, M., Singleton, A., and Parker, C. (2008). Web Mapping 2.0: The Neogeography of the GeoWeb. *Geography Compass*, 2, 2011–2039.

Haklay, M. and Weber, P. (2008). OpenStreetMap: User-generated street map. *IEEE Pervasive Computing*, 7, 12–18.

Heipke, C. (2010). Crowdsourcing Geospatial Data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 65, 550–557.

Helbich, M., Amelunxen, C., Neis, P., and Zipf, A. (2012). Comparative Spatial Analysis of Positional Accuracy of OpenStreetMap and Proprietary Geodata. In: *Proceedings of GI Forum 2012: Geovisualization, Society and Learning.* (July 4–6, 2012). Salzburg, Austria.

Hentschel, M. and Wagner, B. (2010). Autonomous Robot Navigation Based on OpenStreetMap Geodata. In: *Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems.* (Sept. 19–22, 2010). Madeira Island, Portugal.

Hochmair, H.H. and Zielstra, D. (2013). Development and Completeness of Points Of Interest in Free and Proprietary Data Sets: A Florida Case Study. In: *Proceedings of GI Forum 2013: Creating the GISociety.* (July 3–5, 2013). Salzburg, Austria.

Hochmair, H.H., Zielstra, D., and Neis, P. (2013). Assessing the Completeness of Bicycle Trails and Designated Lane Features in OpenStreetMap for the United States and Europe. In: *Proceedings of the the Transportation Research Board - 92nd Annual Meeting.* (Jan. 13–17, 2013). Washington, DC, USA.

Holone, H., Misund, G., and Holmstedt, H. (2007). Users Are Doing It For Themselves: Pedestrian Navigation with User Generated Content. In: *Proceedings of the 2007 International Conference on Next Generation Mobile Applications, Services and Technologies.* (Sept. 12–14, 2007). Cardiff, Wales, UK.

Horita, F. and De Albuquerque, J.P. (2013). An Approach to Support Decision-Making in Disaster Management based on Volunteer Geographic Information (VGI) and Spatial Decision Support Systems (SDSS). In: *Proceedings of the 10th International Conference on Information Systems for Crisis Response and Management.* (May 12–15, 2013). Baden-Baden, Germany.

Hristova, D., Mashhadi, A., Quattrone, G., and Capra, L. (2012). Mapping Community Engagement with Urban Crowd-Sourcing. In: *Proceedings of When the City Meets the Citizen Workshop.* (June 4, 2012). Dublin, Ireland.

Hudson-Smith, A., Batty, M., Crooks, A., and Milton, R. (2009). Mapping for the Masses: Accessing Web 2.0 through Crowdsourcing. *Social Science Computer Review*, 27(4), 524–538.

Hunter, G.J. (1999). New tools for handling spatial data quality: moving from academic concepts to practical reality. *URISA Journal*, 11, 25–34.

Jackson, S.P., Mullen, W., Agouris, P., Crooks, A., Croitoru, A., and Stefanidis, A. (2013). Assessing Completeness and Spatial Error of Features in Volunteered Geographic Information. *ISPRS International Journal of Geo-Information*, 2, 507–530.

Jacob, R., Mooney, P., Corcoran, P., and Winstanley, A.C. (2010). Haptic-gis: exploring the possibilities. In: *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. (Nov. 5, 2010). San Jose, CA, USA.

Janowicz, K. and Hitzler, P. (2010). The Digital Earth as Knowledge Engine. *Semantic Web*, 3, 213–221.

Javanmardi, S., Ganjisaffar, Y., Lopes, C., and Baldi, P. (2009). User Contribution and Trust in Wikipedia. In: *Proceedings of the 5th International Conference on Collaborative Computing: Networking, Applications and Worksharing*. (Nov. 11–14, 2009). Washington, DC, USA.

Jo, H.H., Karsai, M., Kertész, J., and Kaski, K. (2012). Circadian pattern and burstiness in mobile phone communication. *New Journal of Physics*. DOI: 10.1088/1367-2630/14/1/013055.

Kessler, C. and Groot, R.T.A. de (2013). Trust as a Proxy Measure for the Quality of Volunteered Geographic Information in the Case of OpenStreetMap.

Kessler, C., Trame, J., and Kauppinen, T. (2011). Tracking Editing Processes in Volunteered Geographic Information: The Case of OpenStreetMap. In: *Proceedings of the COSIT'11 Workshop: Identifying Objects, Processes and Events in Spatio-Temporally Distributed Data*. (Sept. 12–16, 2011). Belfast, Maine, USA.

Koukoletsos, T., Haklay, M., and Ellul, C. (2012). Assessing Data Completeness of VGI through an Automated Matching Procedure for Linear Data. *Transaction in GIS*, 16, 477–498.

Kuhn, W. (2007). National Center for Geographic Information and Analysis and Vespucci Specialist Meeting on Volunteered Geographic Information. In: *Volunteered Geographic Information and Giscience*. Santa Barbara, CA, USA.

Lechner, M. (2011). Nutzungspotentiale crowdsource-erhobener Geodaten auf verschiedenen Skalen. Ph.D. Dissertation. University Freiburg, Freiburg, Germany.

Ludwig, I., Voss, A., and Krause-Traudes, M. (2011). A Comparison of the Street Networks of Navteq and OSM in Germany. *Advancing Geoinformation Science for a Changing World*, 1, 65–84.

Luxen, D. and Vetter, C. (2011). Real-Time Routing with OpenStreetMap data. In: *Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. (Nov. 1–4, 2011). Chicago, Illinois, USA.

Manfré, L.A., Hirata, E., Silva, J.B., Shinohara, E.J., Giannotti, M.A., Larocca, A.P.C., and Quintanilha, J.A. (2012). An Analysis of Geospatial Technologies for Risk and

Natural Disaster Management. *ISPRS International Journal of Geo-Information*, 1, 166–185.

Mashhadi, A. J., Quattrone, G., Capra, L., and Mooney, P. (2012). On the Accuracy of Urban Crowd-sourcing for Maintaining Large-scale Geospatial Databases. In: *Proceedings of the Eighth Annual International Symposium on Wikis and Open Collaboration*. (Aug. 27–29, 2012). Linz, Austria.

Mondzech, J. and Sester, M. (2011). Quality Analysis of OpenStreetMap Data based on Application need. *Cartographica*, 46, 115–125.

Mooney, P. and Corcoran, P. (2012a). Characteristics of heavily edited Objects in OpenStreetMap. *Future Internet*, 4, 285–305.

Mooney, P. and Corcoran, P. (2012b). How Social is OpenStreetMap? In: *Proceedings of the 15th AGILE International Conference on Geographic Information Science*. (Apr. 24, 2010–Apr. 27, 2012). Avignon, France.

Mooney, P. and Corcoran, P. (2012c). The Annotation Process in OpenStreetMap. *Transaction in GIS*, 16, 561– 579.

Mooney, P. and Corcoran, P. (2012d). Who are the Contributors to OpenStreetMap and what do they do? In: *Proceedings of 20th Annual GIS Research UK*. (Apr. 11–13, 2012). Lancaster, UK.

Mooney, P. and Corcoran, P. (2013). Has OpenStreetMap a role in Digital Earth Applications? *International Journal of Digital Earth*. DOI: 10.1080/17538947.2013.781688.

Mooney, P., Corcoran, P., and Ciepluch, B. (2013). The Potential for using Volunteered Geographic Information in Pervasive Health Computing Applications. *Journal of Ambient Intelligence and Humanized Computing*, 4(6), 731–745.

Mooney, P., Corcoran, P., and Winstanley, A. (2010a). A Study of Data Representation of Natural Features in OpenStreetMap. In: *Proceedings of the 6th GIScience International Conference on Geographic Information Science*. (Sept. 14–17, 2010). Zurich, Switzerland.

Mooney, P., Corcoran, P., and Winstanley, A. (2010b). Towards Quality Metrics for OpenStreetMap. In: *Proceedings of the 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. (Nov. 2–5, 2010). San Jose, CA, USA.

Mülligann, C., Janowicz, K., Ye, M., and Lee, W. (2011). Analyzing the Spatial-Semantic Interaction of Points of Interest in Volunteered Geographic Information. In: *Spatial Information Theory - Lecture Notes in Computer Science Volume 6899*. (Sept. 12–16, 2011). Belfast, ME, USA.

Neis, P., Goetz, M., and Zipf, A. (2012a). Towards Automatic Vandalism Detection in OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 315–332.

Neis, P., Singler, P., and Zipf, A. (2010). Collaborative Mapping and Emergency Routing for Disaster Logistics – Case Studies from the Haiti Earthquake and the UN Portal for Afrika. In: *Proceedings of the Geoinformatics Forum*. (July 6–9, 2010). Salzburg, Austria.

Neis, P., Zielstra, D., and Zipf, A. (2012b). The Street Network Evolution of Crowdsourced Maps: OpenStreetMap in Germany 2007–2011. *Future Internet*, 4, 1–21.

Neis, P., Zielstra, D., and Zipf, A. (2013). Comparison of Volunteered Geographic Information Data Contributions and Community Development for Selected World Regions. *Future Internet*, 5, 282–300.

Neis, P. and Zipf, A. (2008). OpenRouteService.org is Three Times "Open": Combining OpenSource, OpenLS and OpenStreetMaps. In: *Proceedings of the GIS Research UK 16th Annual conference GISRUK 2008*. (Apr. 2–4, 2008). Manchester, UK.

Neis, P. and Zipf, A. (2012). Analyzing the Contributor Activity of a Volunteered Geographic Information Project - The Case of OpenStreetMap. *ISPRS International Journal of Geo-Information*, 1, 146–165.

Nemoto, K., Gloor, P., and Laubacher, R. (2011). Social Capital Increases Efficiency of Collaboration Among Wikipedia Editors. In: *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*. (June 6–9, 2011). Eindhoven, The Netherland.

Nielsen, J. (2006). *Participation Inequality: Encouraging More Users to Contribute*. URL: `http://www.nngroup.com/articles/participation-inequality/` (visited on 09/23/2013).

OpenStreetMap (2013a). *Apple Maps | OpenStreetMap Blog*. URL: `http://blog.osm.org/2012/10/02/apple-maps/` (visited on 09/23/2013).

OpenStreetMap (2013b). *Bing – OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Bing` (visited on 09/20/2013).

OpenStreetMap (2013c). *Editing – OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Editing` (visited on 09/21/2013).

OpenStreetMap (2013d). *Foundation*. URL: `http://wiki.osmfoundation.org/wiki/Main_Page` (visited on 09/24/2013).

OpenStreetMap (2013e). *Humanitarian OSM Team – OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/Humanitarian_OSM_Team` (visited on 09/15/2013).

OpenStreetMap (2013f). *IRC – OpenStreetMap Wiki*. URL: `http://wiki.osm.org/wiki/IRC` (visited on 11/24/2013).

OpenStreetMap (2013g). *Mailing lists – OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Mailing_lists (visited on 11/24/2013).

OpenStreetMap (2013h). *Map Features – OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Map_Features (visited on 11/24/2013).

OpenStreetMap (2013i). *Planet OSM – Full History*. URL: http://planet.osm.org/planet/full-history/ (visited on 11/24/2013).

OpenStreetMap (2013j). *Typhoon Haiyan – OpenStreetMap Wiki.:Statistics*. URL: http://wiki.osm.org/wiki/Typhoon_Haiyan (visited on 11/24/2013).

OpenStreetMap (2013k). *Yahoo! Aerial Imagery – OpenStreetMap Wiki*. URL: http://wiki.osm.org/wiki/Yahoo!_Aerial_Imagery (visited on 09/20/2013).

O'Reilly, T. (2005). *What Is Web 2.0: Design Patterns and Business Models for the Next Generation of Software*. URL: http://oreilly.com/web2/archive/what-is-web-20.html (visited on 09/21/2013).

OSMF (2013). *Working Groups - OSM Foundation*. URL: http://www.osmfoundation.org/wiki/Working_Groups (visited on 09/24/2013).

OSMstats (2013). *Statistics of the free wiki world map*. URL: http://osmstats.altogetherlost.com (visited on 10/05/2013).

Ostermann, F.O. and Spinsanti, L. (2011). A Conceptual Workflow For Automatically Assessing The Quality Of Volunteered Geographic Information For Crisis Management. In: *Proceedings of the 14th AGILE International Conference on Geographic Information Science*. (Apr. 18–21, 2011). Utrecht, The Netherlands.

Poser, K. and Dransch, D. (2010). Volunteered Geographic Information for Disaster Management with Application to Rapid Flood Damage Estimation. *Geomatica*, 64, 89–98.

Potthast, M., Stein, B., and Gerling, R. (2008). Automatic Vandalism Detection in Wikipedia. In: *Proceedings of the IR Research, 30th European Conference on Advances in Information Retrieval*. (Mar. 30–Apr. 3, 2008). Glasgow, Scotland.

Pourabdollah, A., Morley, J., Feldman, S., and Jackson, M. (2013). Towards an Authoritative OpenStreetMap: Conflating OSM and OS OpenData National Maps' Road Network. *ISPRS International Journal of Geo-Information*, 2, 704–728.

Pultar, E., Raubal, M., Cova, T.J., and Goodchild, M.F. (2009). Dynamic GIS Case Studies: Wildfire Evacuation and Volunteered Geographic Information. *Transaction in GIS*, 13, 85–104.

Ramm, F. and Stark, H.J. (2008). Crowdsourcing Geodata. *Géomatique Suisse*, 6, 315–318.

Ramm, F., Topf, J., and Chilton, S. (2010). Cambridge, UK: UIT.

Rashid, O., Dunbar, A., Fisher, S., and Rutherford, J. (2010). Users Helping Users - User generated content to assist wheelchair users in an urban environment. In: *Proceedings of the 9th International Conference on Mobile Business and 2010 Ninth Global Mobility Roundtable*. (June 13–15, 2010). Athens, Greece.

Raymond, E.S. (1999). The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary. In: Beijing, China: O'Reilly.

Roick, O. and Heuser, S. (2012). Location Based Social Networks – Definition, Current State of the Art and Research Agenda. *Transaction in GIS*, 17, 763–784.

Rouse, L.J., Bergeron, S.J., and Harris, T.M. (2007). Participating in the Geospatial Web: Collaborative Mapping, Social Networks and Participatory GIS. In: *The Geospatial Web - Advanced Information and Knowledge Processing*. London, UK: Springer.

Scheider, S. and Possin, J. (2012). Affordance-based Individuation of Junctions in Open Street Map. *Journal of Spatial Information Sciences*, 4, 31–56.

Schilling, A., Over, M., Neubauer, S., Neis, P., Walenciak, G., and Zipf, A. (2009). Interoperable Location Based Services for 3D Cities on the Web Using User Generated Content from OpenStreetMap. In: *Proceedings of the 27th Urban Data Management Symposium*. (June 24–26, 2009). Ljubljana, Slovenia.

Schmitz, S., Neis, P., and Zipf, A. (2008). New Applications Based on Collaborative Geodata - the Case of Routing. In: *Proceedings of XXVIII INCA International Congress on Collaborative Mapping and Space Technology*. (Nov. 4–6, 2008). Gandhinagar, Gujarat, India.

Sieber, R. (2006). Public Participation Geographic Information Systems: A Literature Review and Framework. *Annals of the American Association of Geography*, 96, 491–507.

Song, W. and Sun, G. (2010). The Role of Mobile Volunteered Geographic Information in Urban Management. In: *Proceedings of 18th International Conference on Geoinformatics*. (June 18–20, 2010). Beijing, China.

Stark, H.J. (2010). Umfrage zur Motivation von Freiwilligen im Engagement in Open Geo-Data Projekten. In: *Proceedings of FOSSGIS Anwenderkonferenz für Freie und Open Source Software für Geoinformationssysteme*. (Mar. 2–5, 2010). Osnabrück, Germany.

Stein, K. and Hess, C. (2007). Does it Matter Who Contributes: A Study on Featured Articles in the German Wikipedia. In: *Proceedings of the eighteenth Conference on Hypertext and Hypermedia*. (Sept. 10–12, 2007). Manchester, UK.

Steinmann, R., Gröchenig, S., Rehrl, K., and Brunauer, R. (2013a). Contribution Profiles of Voluntary mappers in OpenStreetMap. In: *Online Proceedings of the International Workshop on Action and Interaction in Volunteered Geographic Information (ACTIVITY) In: 16th AGILE Conference on Geographic Information Science*. (May 14–17, 2013). Leuven, Belgium.

Steinmann, R., Häusler, E., Klettner, S., Schmidt, M., and Lin, Y. (2013b). Gender Dimensions in UGC and VGI - A Desk-based Study. In: *Proceedings of Angewandte Geoinformatik 2013*. (July 3–5, 2013). Salzburg, Asutria.

Stephens, M. (2013). Gender and the GeoWeb: Divisions in the Production of User-generated Cartographic Information. *GeoJournal*, 1–16.

Sui, D. (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.

Tapscott, D. (1997). The Digital Economy: Promise and Peril. In: *The Age of Networked Intelligence*. New York, USA: McGraw Hill.

Touya, G. and Brando-Escobar, C. (2013). Detecting Level-of-Detail Inconsistencies in Volunteered Geographic Information Data Sets. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 48, 134–143.

Ushahidi (2013). *Ushahidi*. URL: `http://ushahidi.com` (visited on 09/15/2013).

Van Oort, P.A.J. (2006). Spatial Data Quality: From Description to Application. Ph.D. Thesis. Wageningen University.

Walter, V. and Fritsch, D. (1999). Matching Spatial Data Sets: A Statistical Approach. *International Journal of Geographical Information Science*, 13, 445–473.

Wikipedia (2013). *Wikipedia:Statistics*. URL: `http://en.wikipedia.org/wiki/Wikipedia:Statistics` (visited on 10/05/2013).

Wilkinson, D. and Huberman, B. (2007). Assessing the Value of Cooperation in Wikipedia. *First Monday*, 12.

Wunsch-Vincent, S. and Vickery, G. (2007). Participative Web: User-Created Content: Web 2.0, Wikis and Social Networking. In: *Organisation for Economic Co-operation and Development*. Paris, France.

Yasseri, T., Sumi, R., and Kertész, J. (2012). Circadian Patterns of Wikipedia Editorial Activity: A Demographic Analysis. *PLoS ONE*. DOI: `10.1371/journal.pone.0030091`.

Zielstra, D. and Hochmair, H.H. (2011a). A Comparative Study of Pedestrian Accessibility to Transit Stations using Free and Proprietary Network Data. *Transportation Research Record: Journal of the Transportation Research Board*, 2217, 145–152.

Zielstra, D. and Hochmair, H.H. (2011b). Digital Street Data: Free versus Proprietary. *GIM International*, 25, 29–33.

Zielstra, D. and Hochmair, H.H. (2012). Comparison of Shortest Path Lengths for Pedestrian Routing in Street Networks Using Free and Proprietary Data. In: *Proceedings of the Transportation Research Board - 91st Annual Meeting*. (Jan. 22–26, 2012). Washington, DC, USA.

Zielstra, D., Hochmair, H.H., and Neis, P. (2013). Assessing the Effect of Data Imports on the Completeness of OpenStreetMap – A United States Case Study. *Transaction in GIS*, 17, 315–334.

Zielstra, D. and Zipf, A. (2010). A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany. In: *Proceedings of the 13th AGILE International Conference on Geographic Information Science*. (May 10–14, 2010). Guimarães, Portugal.

Eidesstattliche Versicherung gemäß § 8 der Promotionsordnung der Naturwissenschaftlich-Mathematischen Gesamtfakultät der Universität Heidelberg

1. Bei der eingereichten Dissertation zu dem Thema *Analysis of User-generated Geodata Quality for the Implementation of Disabled People Friendly Route Planning* handelt es sich um meine eigenständig erbrachte Leistung.

2. Ich habe nur die angegebenen Quellen und Hilfsmittel benutzt und mich keiner unzulässigen Hilfe Dritter bedient. Insbesondere habe ich wörtlich oder sinngemäß aus anderen Werken übernommene Inhalte als solche kenntlich gemacht.

3. Die Arbeit habe ich bislang nicht an einer Hochschule des In- oder Auslands als Bestandteil einer Prüfungs- oder Qualifikationsleistung vorgelegt.

4. Die Richtigkeit der vorstehenden Erklärungen bestätige ich.

5. Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt.

Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Hünstetten Wallbach, 22. Februar 2014

Pascal Neis