A Comparative Study of Proprietary Geodata and Volunteered Geographic Information for Germany

Dennis Zielstra[1], Alexander Zipf[2]

[1]University of Bonn (Germany), Department of Geography, Cartography Research Group

[2]University of Heidelberg (Germany), Department of Geography, Chair of Geoinformatics

INTRODUCTION

In connection with the Web 2.0 movement of the Internet (O'Reilly, 2005) and the progressive development of tools and applications for the collection and provision of spatial information (Turner, 2006), the quality and quantity of so-called Volunteered Geographic Information (VGI) (Goodchild, 2007) underwent a fast-paced worldwide development. Some even speak of a "Wikification of GIS" (Sui, 2008). This spatial data, mostly collected by volunteers, is freely available for the Internet user and can (under certain licensing conditions) be applied to numerous GIS projects and applications. Through advanced data donations, but also by a variety of other non-proprietary data sources, some of these free data providers are able to offer a vast variety of different information.

This development in recent years stands in strong contrast to the very expensive commercial spatial data provided by a few companies. Much of this proprietary data is widely used today, for example, in car navigation devices or cell phones. The strong demand for freely available spatial data, though, has boosted the number of VGI available on the Internet. They can be found in very simple forms such as in Wikipedia entries that provide some spatial information like lat-long coordinates (geotag), or in so-called mashups in Google Earth or Google Maps, which combine different information sources. One of the most complex and promising projects in recent years, however, is OpenStreetMap[1] (OSM).

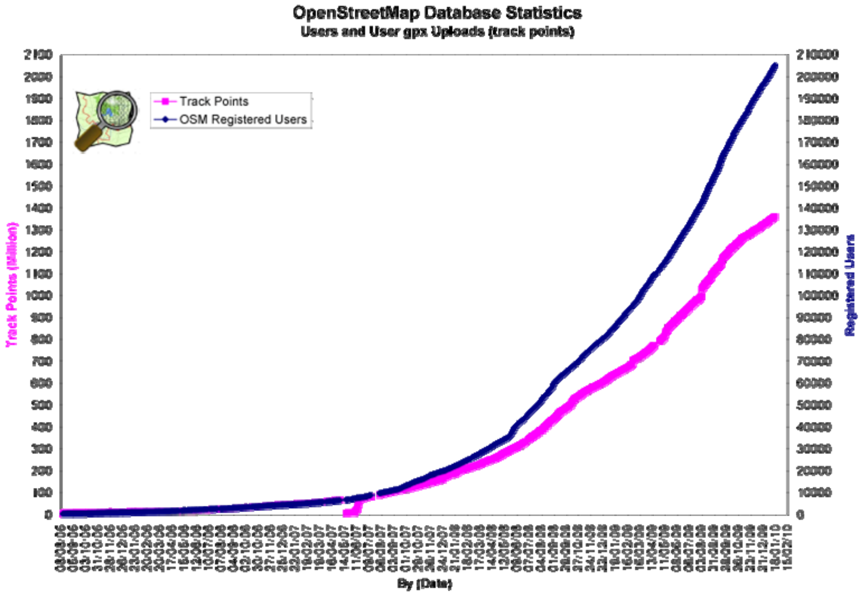

Figure 1: User- and Data development of OpenStreetMap (2005-2010)[2]

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

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

Initiated in 2004 at University College London, by Steve Coast, OpenStreetMap gives all Internet users the opportunity to download spatial data without any costs or fees and to use it for their own projects. The goal of the OpenStreetMap community is to create a map of the world that will contain as much detailed information as possible, and this information is being collected by volunteers. As Figure 1 illustrates, both the membership numbers and the amount of data since the beginning of the project rose rapidly in an impressive manner. Since March 2009, the number of 100,000 members has been exceeded; in January 2010 it has again doubled to over 200,000 members. Since 2004, OpenStreetMap has collected, particularly in Europe, a large amount of geodata, with the greatest gains coming within the last two years.

But, as in many other projects related to the Web2.0 movement, including Wikipedia and others, questions are being raised about the accuracy and correctness of the information provided. VGI, including those of the OpenStreetMap project, is no exception to this concern, and it raises numerous doubts about their quality and reliability (Goodchild, 2007). Despite the positive aspects of the project, there are still concerns regarding free data, as compared to data provided by professional manufacturers such as TeleAtlas and Navteq (Flanagin & Metzger, 2008).

The goal of this paper is to make a contribution to this discussion and to find some answers to widely asked questions regarding comparisons of OpenStreetMap and TeleAtlas data. Based on initial studies from England (Haklay, 2008; Ather, 2009), an examination and analysis of German datasets is presented (Zielstra, 2009, Zielsta & Zipf 2009). Some results from this research, as well as follow-up investigations are described and discussed in this paper. Further aspects of OSM data quality and usability are investigated for example in (Schmitz et al. 2008, Auer & Zipf 2009, Neis et al. 2010. or Amelunxen 2010).

AVAILABLE DATASETS

There are different ways to obtain the data from the OpenStreetMap.org website. One way is to define a desired area, and then store the information contained in it (roads, etc.) to an XML file. Companies like "Geofabrik"[3] and „Cloudmade“[4] offer OpenStreetMap data in shapefile, XML and several other formats as a free download, along with many tools and information about OpenStreetMap. The data is already divided into hierarchical regions. Since the OpenStreetMap data, as shown in Figure 1, has experienced an almost exponential growth, the data provided by the Geofabrik is updated on a daily basis and the Cloudmade data on a weekly basis. As a counterpart to the freely available geodata from OpenStreetMap, a proprietary dataset from the commercial provider, TeleAtlas, was used. To be as accurate as possible, the TeleAtlas MultiNet data package has been used, which has widespread availability in different versions and is used by many applications. The cost for the data depends on the license and data size (e.g., All of Europe or only one Country), intended use, output medium, and other factors. The TeleAtlas MultiNet data is available in shapefile (.Shp) format and is being updated on a regular basis with quarterly releases.

It needs to be noted explicitly, that the results of the following analyzes only show a relative comparison between two available datasets. It cannot show the real completeness or absolute quality of OSM in respect to the real world (ground truth). But as the Tele Atlas Multinet dataset is a very successful commercial dataset it can be used as a relative reference for comparison in particular with respect to navigation tasks. Both datasets (and others such as Navteq etc.) offer of course a wider range of data types that are being investigated in further research work in the research group.

[3] http://www.geofabrik.de
[4] http://cloudmade.com

METHODS

One of the major goals of the data analysis was to include information on the completeness of the OpenStreetMap data in comparison to the TeleAtlas MultiNet dataset. The completeness of a road network can be determined by calculating the total length of the roads of one of the dataset providers within a predefined area and then comparing it with the total length of roads of the other provider within the same area. If there is a difference in the overall lengths, it indicates that one of the datasets is more complete than the other (see also Haklay 2008). This is, of course, only a relative measure, as we do not have any hints about the actual length of the entire street network in the real world.

The first investigations that were carried out involved a set of calculations using different datasets in which the differences of the total data length for the entire federal territory of Germany were defined.

After receiving the general information about the differences in the datasets, a closer examination of some selected areas (medium-sized towns and large cities) took place. First, the five biggest German cities (Berlin, Hamburg, Munich, Cologne, and Frankfurt) were analyzed. Then five medium-sized cities, selected by specific criteria such as location and population, were analyzed. The differences between OpenStreetMap and TeleAtlas in each city were calculated for three OSM datasets (April, July and December) and visualized in a chart with relative values.

Once the differences in the total lengths for all of Germany and the various cities were calculated, a different method, based on research by Haklay (2008) in England, was used to analyze and visualize the regional differences between the two datasets. This method includes the calculation of total lengths and differences by square km (OpenStreetMap minus TeleAtlas = difference in each grid area). To reinforce the significance of the results, however, the differences were calculated not only in absolute terms, as Haklay computed and displayed, but also in relative values for each grid space included, as there is wide variation in total lengths of street network data between rural and more densely populated areas, even in Germany.

Further investigations regarding this issue were processed by making different comparisons between the cities used in the analysis before. From the center of the cities' circular buffers with different distances (large cities: 0-10 km, 10-25 km, 25-50 km; medium-sized cities: 0-5 km, 5-15 km, 15-25 km) were created. Thereafter, the difference between OpenStreetMap and TeleAtlas within each buffer was calculated. This allowed the increases in the differences of the data toward the rural areas to be accurately quantified.

RESULTS

Figures 2 through 4 show that, in total, the length of the streets available in OSM is still smaller than that of the street-length data available in TeleAtlas MultiNet. But the length growth rate of the entire street network in OSM is tremendous, as within only 8 months the difference between OSM and TeleAtlas was reduced from 29% to 7% (see Figure 2).

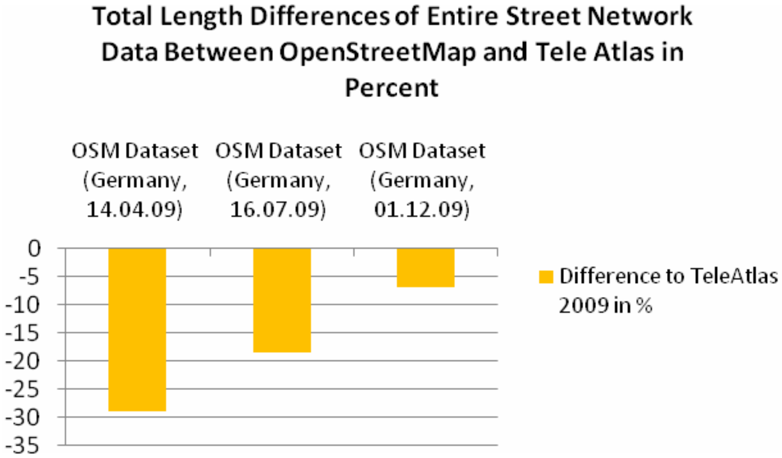

Figure 2: Comparison of the used datasets with respect to the entire street network

The biggest differences between the two datasets can be found in the car navigation related data (see Figure 3). Since this is the field that Tele Atlas specializes in, this result is not very surprising but gives a good example on how the different datasets specialize in different fields of data. Also the results of the comparison of the pedestrian navigation related data (see Figure 4) provide further evidence for the assumption that OSM specializes on smaller streets and alleys. Of course, however, the resulting pressing question is how this differs from region to region (e.g., rural to urban areas) between different object types. Therefore, more detailed investigations were carried out and still are continuously being conducted.

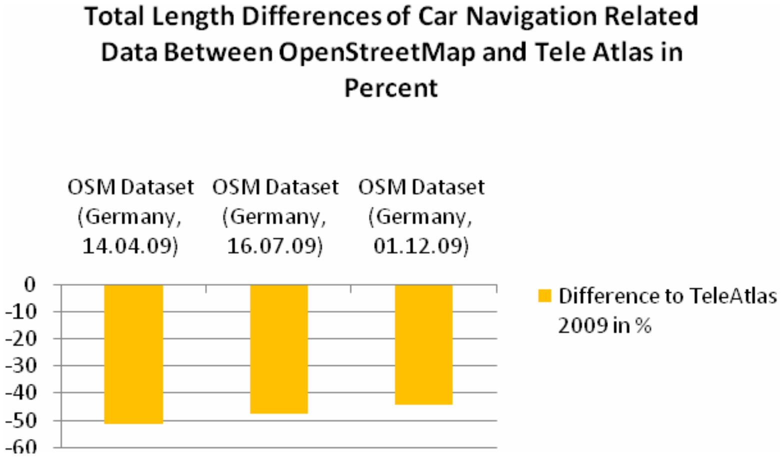

Figure 3: Comparison of the datasets used with respect to car navigation

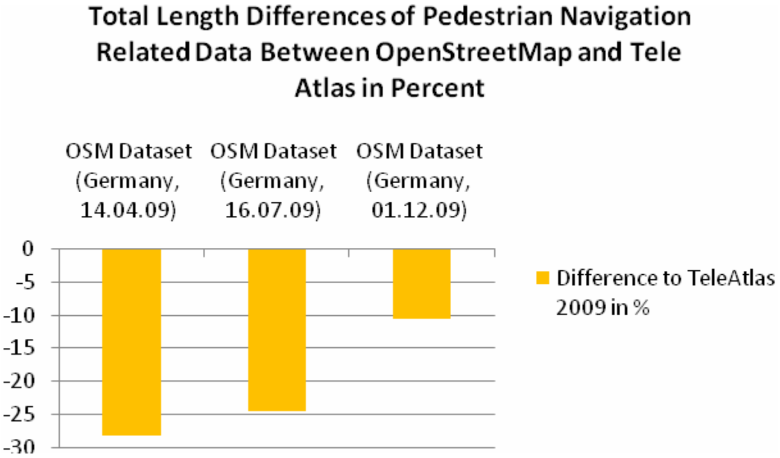

Figure 4: Comparison of the datasets used with respect to pedestrian navigation

Once again, the basic comparison calculations of the different databases shown in Figures 2 through 4 represent and prove the impressive and rapid development of the OpenStreetMap Project. An increase of roughly 20% in just three months is an enormous amount of data. The calculations for the large cities show that of all the cities studied, the OpenStreetMap community had collected more data than TeleAtlas (see Figure 5) in the summer of 2009. This means that, especially in larger cities in Germany, OpenStreetMap covers many small trails and pathways for pedestrians, but also smaller side streets that are used by cars.

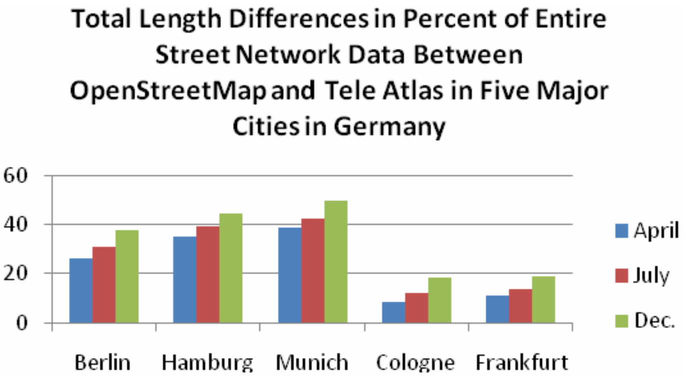

Figure 5: Comparison of the datasets used in five major towns with respect to the entire street network
(OpenStreetMap April, July, and Dec. 2009, TeleAtlas 2009)

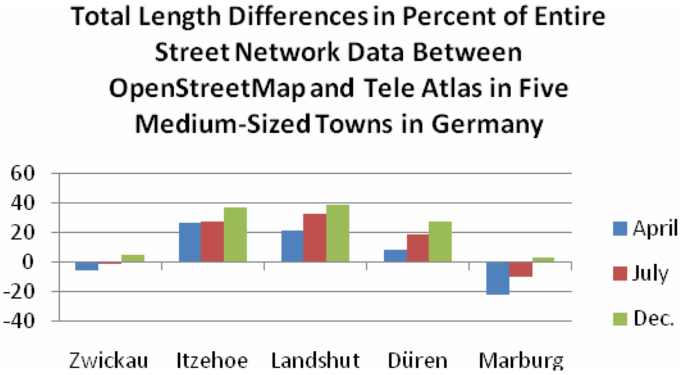

Figure 6: Comparison of the datasets used in five medium-sized towns with respect to the entire street network
(OpenStreetMap April, July, and Dec. 2009, TeleAtlas 2009)

Even in medium-sized towns with fewer active OpenStreetMap members, the results of the calculations showed more collected geodata than that provided by TeleAtlas in December 2009 (see Figure 6). Larger differences can be found in the car-routing-related data, which is where TeleAtlas's proficiency is revealed. Two out of five major cities are still showing deficits in OSM regarding this data but the differences are decreasing with time (see Figure 7).

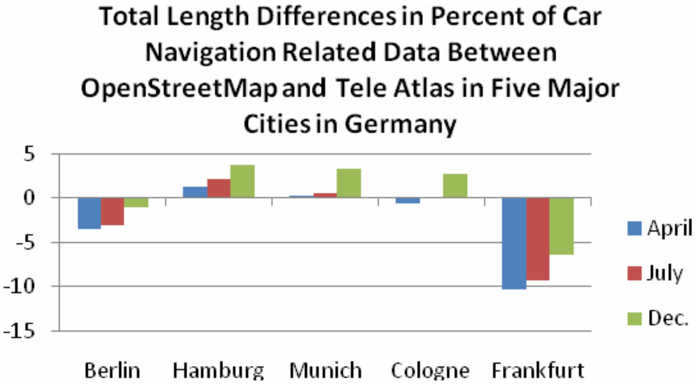

Figure 7: Comparison of the datasets used in five major cities with respect to car navigation
(OpenStreetMap April, July, and Dec. 2009, TeleAtlas 2009)

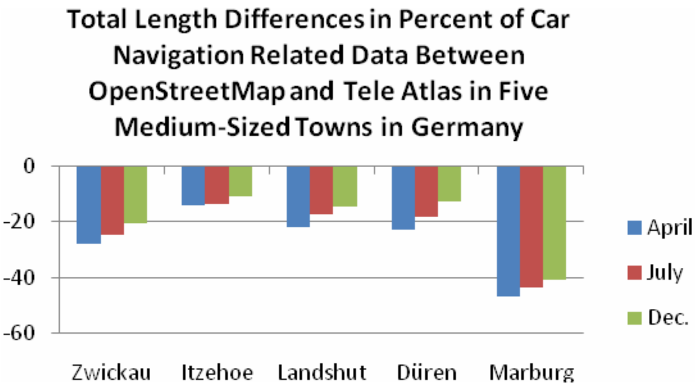

Figure 8: Comparison of the datasets used in five medium-sized towns with respect to car navigation
(OpenStreetMap April, July, and Dec. 2009, TeleAtlas 2009)

Every medium-sized town analyzed during the research showed large deficits in the car routing related data (see Figure 8). The comparison of data related to pedestrian navigation, however, showed again the focus of OpenStreetMap on little paths and ways that are either not covered at all, or only in limited amounts, by TeleAtlas (see Figures 9 and 10).

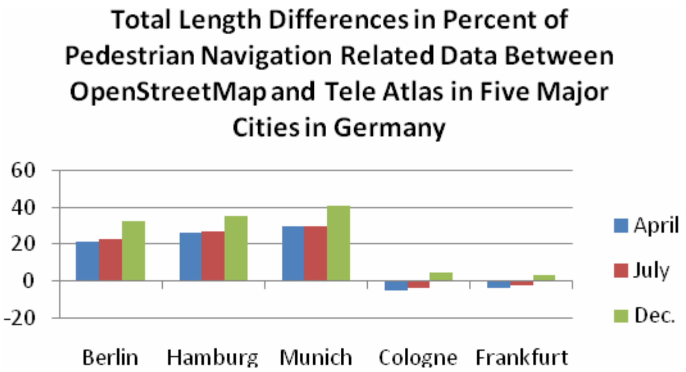

Figure 9: Comparison of the datasets used in five major cities with respect to pedestrian navigation
(OpenStreetMap April, July, and Dec. 2009, TeleAtlas 2009)

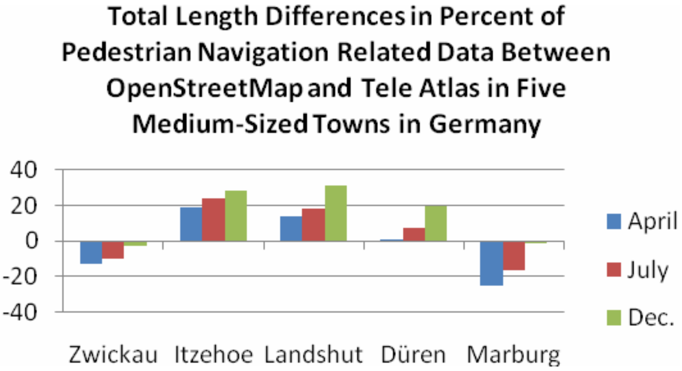

Figure 10: Comparison of the datasets used in five medium-sized towns with respect to pedestrian navigation
(OpenStreetMap April, July, and Dec. 2009, TeleAtlas 2009)

The newly calculated maps (Figure 11 and Figure 12) show a clear decrease in the OpenStreetMap dataset from the metropolitan areas and city centers to the surrounding rural areas. But it also shows that within significant parts of Germany (the more densely populated ones), OSM data now has a larger length of street networks - which means it offers more data in these specific areas than does the commercial provider. This, of course, does not yet say anything about the quality of the data at the attribute level or the geometric precision or the homogeneity according to different object types; however, it does provide a first impression about the potential of VGI and OpenStreetMap.

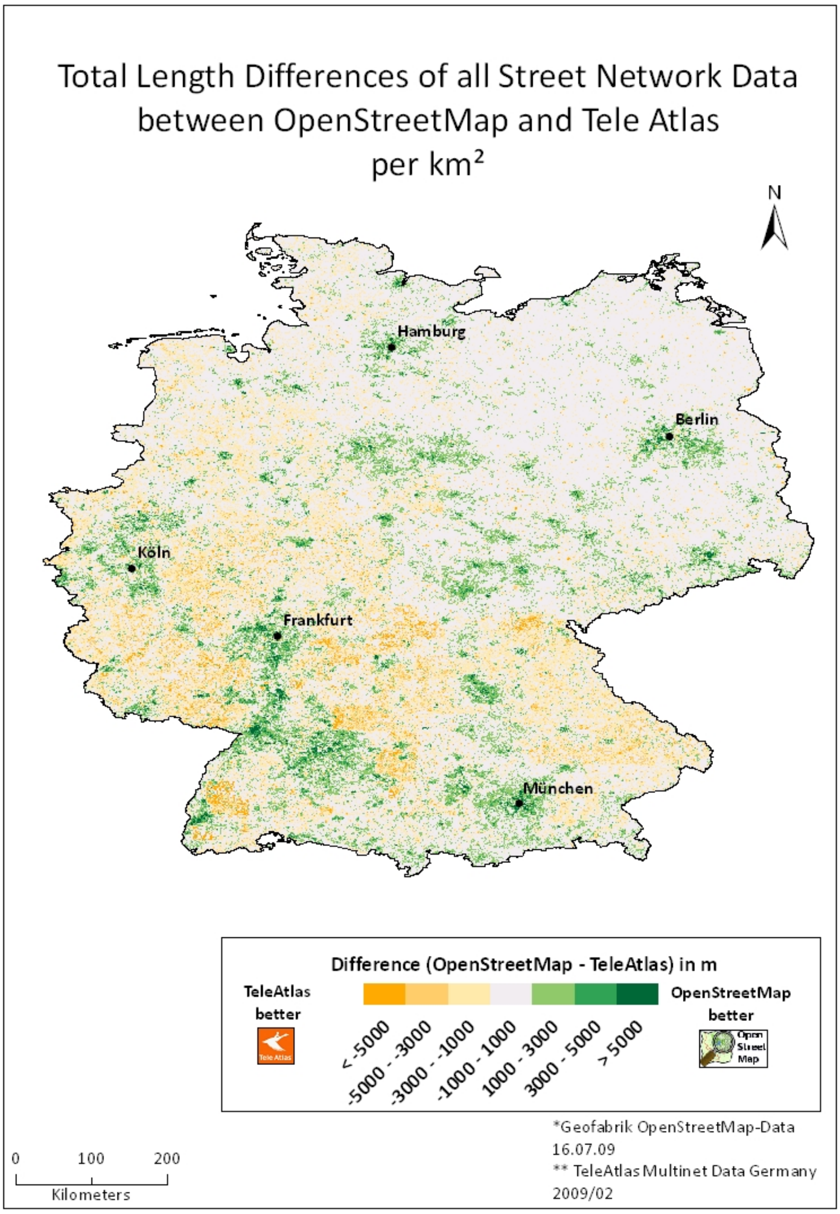
Total Length Differences of all Street Network Data between OpenStreetMap and Tele Atlas per km²
N
Hamburg
Berlin
Köln
Frankfurt
München
Difference (OpenStreetMap - TeleAtlas) in m
TeleAtlas better
OpenStreetMap better
< -5000
-5000 - -3000
-3000 - -1000
-1000 - 1000
1000 - 3000
3000 - 5000
> 5000
0
100
200
Kilometers
*Geofabrik OpenStreetMap-Data 16.07.09
** TeleAtlas Multinet Data Germany 2009/02

Figure 11: Map showing the results of the total length difference calculations in absolute values.

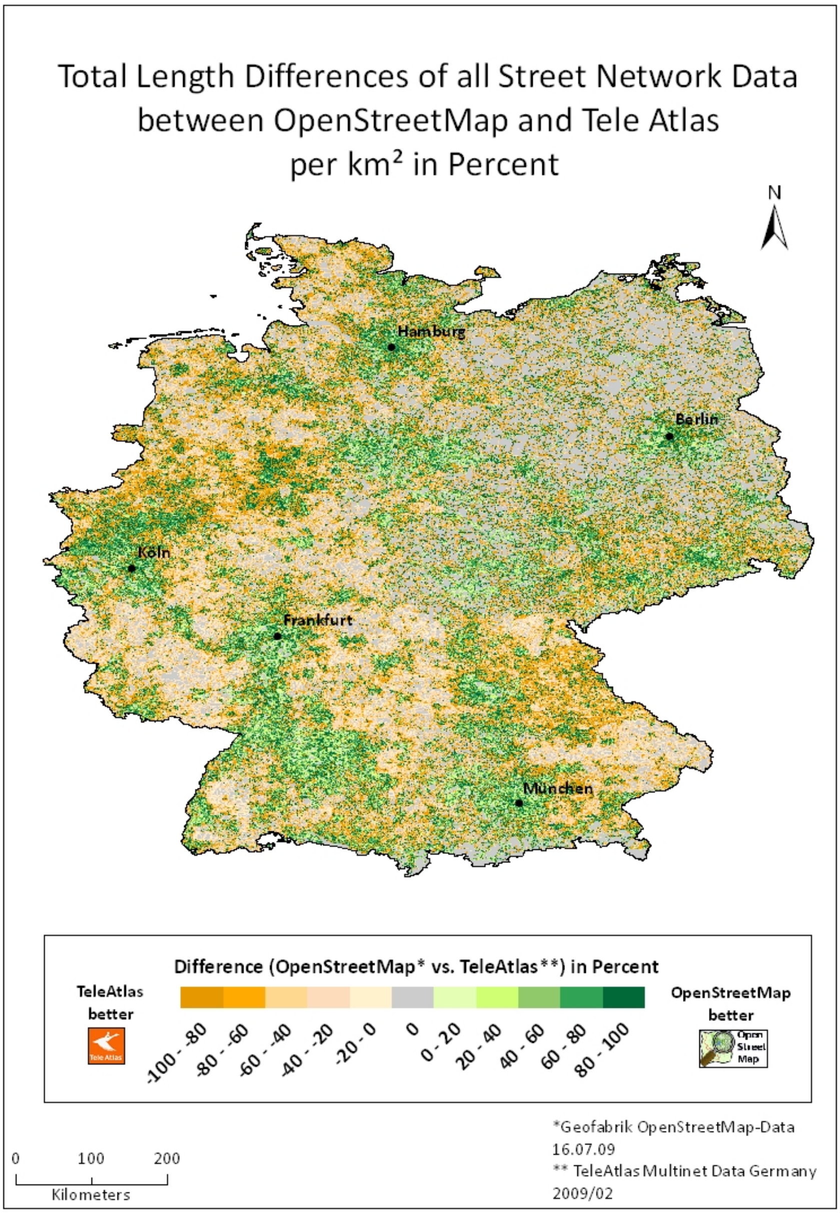
Total Length Differences of all Street Network Data
between OpenStreetMap and Tele Atlas
per km² in Percent
N
Hamburg
Berlin
Köln
Frankfurt
München
Difference (OpenStreetMap* vs. TeleAtlas**) in Percent
TeleAtlas better
-100 - -80
-80 - -60
-60 - -40
-40 - -20
-20 - 0
0
0 - 20
20 - 40
40 - 60
60 - 80
80 - 100
OpenStreetMap better
0
100
200
Kilometers
*Geofabrik OpenStreetMap-Data 16.07.09
** TeleAtlas Multinet Data Germany 2009/02

Figure 12: Map showing the results of the total length difference calculations in relative values

OSM also offers data relevant for pedestrians and bikers while Tele Atlas focuses more on car navigation. This can be seen when comparing the relative results from the analysis between the streets (classes) relevant for cars or for pedestrian navigation (see Figures 13 and 14).

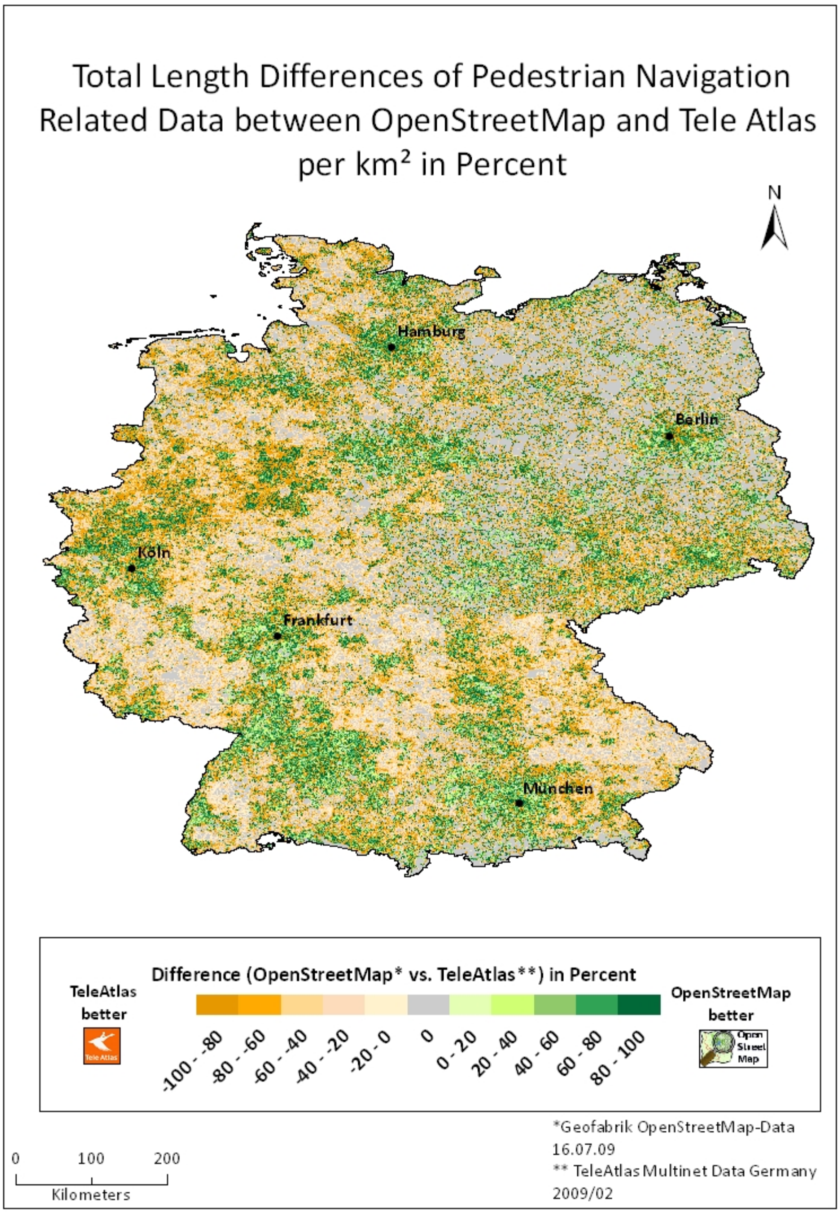
Total Length Differences of Pedestrian Navigation Related Data between OpenStreetMap and Tele Atlas per km² in Percent
N
Hamburg
Berlin
Köln
Frankfurt
München
TeleAtlas better
Difference (OpenStreetMap* vs. TeleAtlas**) in Percent
-100 - -80
-80 - -60
-60 - -40
-40 - -20
-20 - 0
0
0 - 20
20 - 40
40 - 60
60 - 80
80 - 100
OpenStreetMap better
0
100
200
Kilometers
*Geofabrik OpenStreetMap-Data 16.07.09
** TeleAtlas Multinet Data Germany 2009/02

Figure 13: Map showing relative analysis results for pedestrian routing network

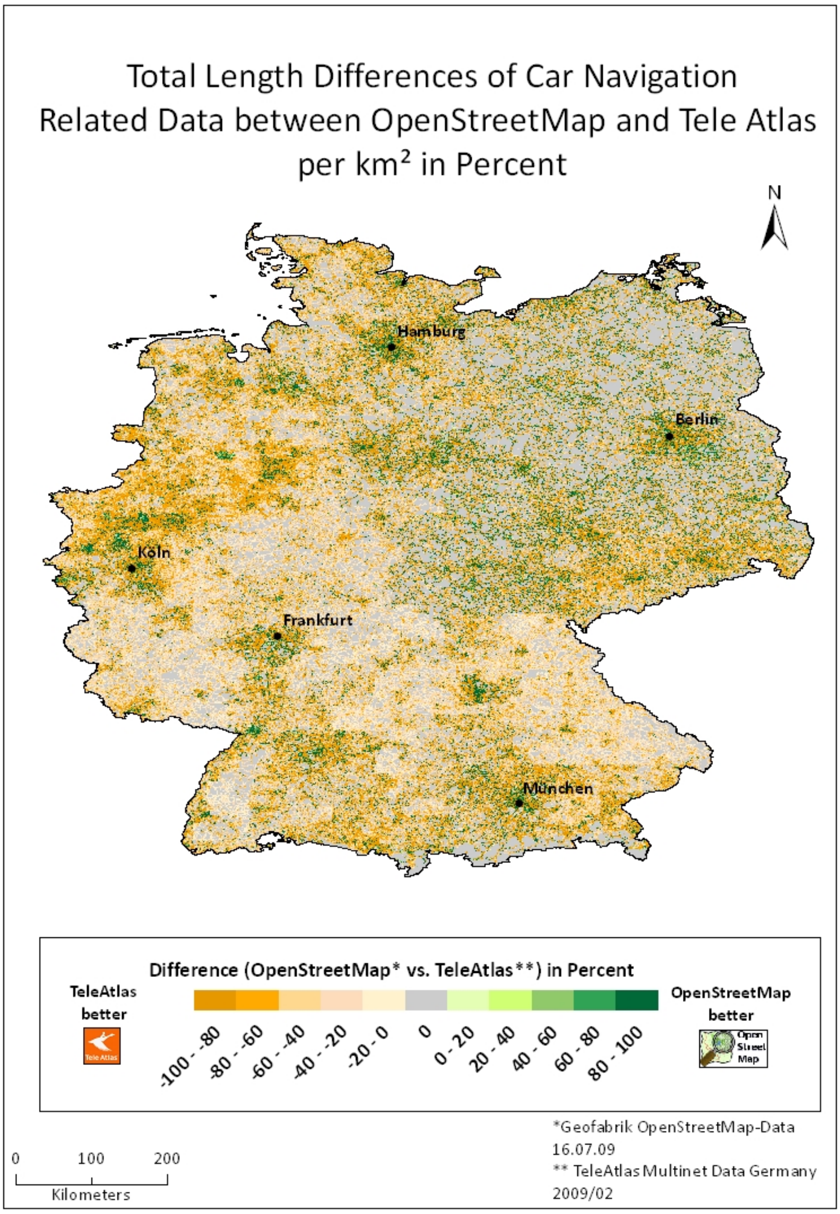
Total Length Differences of Car Navigation Related Data between OpenStreetMap and Tele Atlas per km² in Percent
N
Hamburg
Berlin
Köln
Frankfurt
München
Difference (OpenStreetMap* vs. TeleAtlas**) in Percent
TeleAtlas better
OpenStreetMap better
-100 - -80
-80 - -60
-60 - -40
-40 - -20
-20 - 0
0
0 - 20
20 - 40
40 - 60
60 - 80
80 - 100
*Geofabrik OpenStreetMap-Data 16.07.09
** TeleAtlas Multinet Data Germany 2009/02
0
100
200
Kilometers

Figure 14: Map showing relative analysis results for routing network only relevant to cars

The results of the circular buffer method clearly show a large decrease in the completeness of the data as the distance from the center of the cities increases (see Figure 15). Nearly all the major cities tested show this same pattern and have differences of up to 23%. The mid-sized cities showed even greater problems, in spite of the smaller buffer sizes, as very large differences of up to 57% were calculated (see Figure 16).

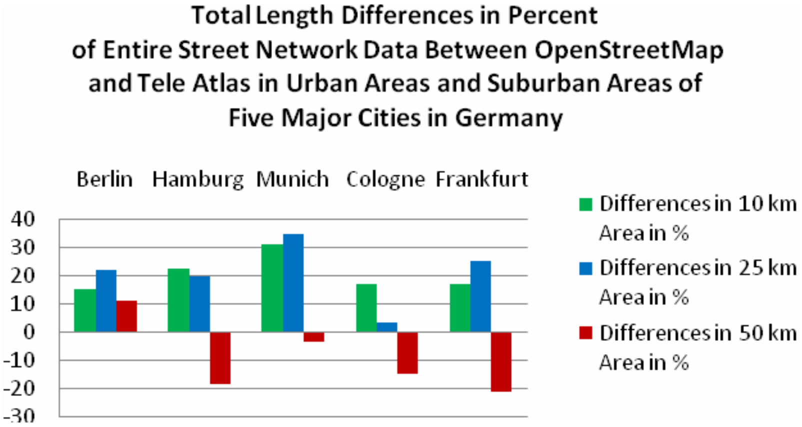

Figure 15: Comparison of the used datasets in five major cities with respect to the entire street network (OpenStreetMap 16.07.2009, Tele Atlas 2009)

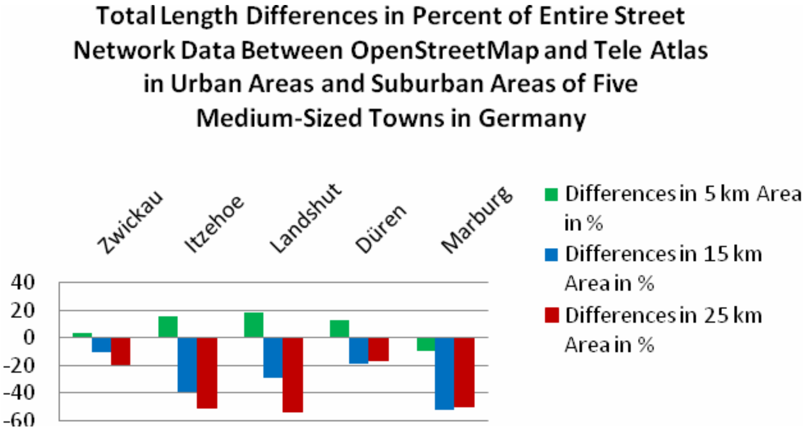

Figure 16: Comparison of the used datasets in five medium-sized towns with respect to the entire street network (OpenStreetMap 16.07.2009, Tele Atlas 2009)

CONCLUSION AND FUTURE WORK

The results of this analysis have shown that the Web 2.0 movement in the field of freely available spatial data has been very active during the last couple of years. While the first VGI used to offer very basic information, such as geotags on Wikipedia articles, over time, projects like OpenStreetMap excelled due to its many members and could thus offer a complex diversity of spatial data. Thus the initial questions regarding the completeness of the German OpenStreetMap data in comparison to proprietary TeleAtlas MultiNet data were raised and answered.

As the results of this paper have shown, there is still a very strong heterogeneity of the OpenStreetMap data in terms of their completeness. In all cities studied, the diversity of the freely available data is significantly higher between inner-city and rural areas, which can be explained by the presence of more active members on the project in the larger cities. Further tests showed that the completeness of the data is marked by strong differences between the large and medium-sized cities. By using a circular buffer method, a significant decrease in the data was observed as the distance from the city center increased. Further, the calculated difference maps, both in absolute and relative terms, were able to visualize the different concentrations of the data. Both datasets (and others such as Navteq etc.) offer, of course, a wider range of data types that are being investigated in further research work.

As a result of the entire analysis, it can be noted that the VGI of the OpenStreetMap project can certainly offer a large amount of data. The theory developed by Goodchild about "Citizens as Sensors" (Goodchild, 2007) is well reflected in this project and demonstrates the potential that lies within OpenStreetMap if its current membership continues to stay active and new members can be gained. However, it is also clear that the freely available data provided is not yet a sufficient replacement for the proprietary TeleAtlas data for all types of applications - in particular, if a more consistent coverage in rural areas is needed. The usability of a dataset will always depend upon the usage needed and its characteristics with respect to completeness, accuracy, homogeneity, and other factors. If coverage is needed only in the densely populated urban areas of Germany (e.g., by regional traffic providers or logistics companies), OpenStreetMap may already be an interesting - and very cost-efficient - alternative to commercial datasets. But again this depends on the actual application.

Of course, the professional data is not without faults, which can be read in many forums on the Internet, but the coverage of OpenStreetMap data in rural areas is too small to be a sophisticated alternative for any application. In larger cities, however, the data diversity is so rich that already projects that are based on proprietary data are being replaced with OpenStreetMap data. It remains to be seen whether the very good collection of data from OpenStreetMap in the major cities can also be transferred to the countryside to obtain even better results.

REFERENCES

Ather, A., 2009 A Quality Analysis of OpenStreetMap Data, M.Eng. Dissertation, Department of Civil, Environmental & Geomatic Engineering, University College London.

Amelunxen, C. 2010 An Approach to gecoding based on volunteered Spatial Data. Geoinformatik 2010. Die Welt im Netz. Kiel.

Auer, M. and Zipf, A. 2009 How do Free and Open Geodata and Open Standards fit together? From Sceptisism versus high Potential to real Applications. The First Open Source GIS UK Conference. University of Nottingham. UK.

Flanagin, A. J. und Metzger, M., 2008 The credibility of volunteered geographic information GeoJournal, 72(3), 137-148.

Goodchild, M. F., 2007 Citizens as sensors: the world of volunteered geography. GeoJournal, (69), 211-221.

Haklay, M., 2008 How good is OpenStreetMap information? A comparative study of OpenStreetMap and Ordnance Survey datasets for London and the rest of England, Under review in Environment & Planning B: Planning and Design.

Neis, P., Zielstra, D., Zipf, A., Struck , A. 2010: Empirische Untersuchungen zur Datenqualität von OpenStreetMap - Erfahrungen aus zwei Jahren Betrieb mehrerer OSM-Online-Dienste. AGIT 2010. Symposium für Angewandte Geoinformatik. Salzburg. Austria.

O'Reilly, T., 2005 What is web 2.0: Design patterns and business models for the next generation of software. O'Reilly Media.

Schmitz,S., Neis, P. and A. Zipf, 2008: New Applications based on Collaborative Geodata - the Case of Routing. XXVIII INCA International Congress on Collaborative Mapping and SpaceTechnology, Gandhinagar, Gujarat, India.

Sui, D. Z., 2008 The wikification of GIS and its consequences: Or Angelina Jolie's new tattoo and the future of GIS. Computers, Environment and Urban Systems, 32(1), p 1-5.

Turner, J. A., 2006 Introduction to Neogeography. Short Cuts, O'Reilly Media.

Zielstra, D., 2009 Datenqualität und Anwendbarkeit von Volunteered Geographic Information. Vergleich von proprietären und frei verfügbaren Geodaten. Diploma Thesis, Department of Geography, Cartography Research Group, University of Bonn.

Zielstra, D. & Zipf, A., 2009: Datenqualität von OpenStreetMap - Erste Ergebnisse empirischer Untersuchungen. AGIT 2009. Symposium für Angewandte Geoinformatik. Salzburg. Austria.