Building Generic Quality Indicators for OpenStreetMap

Błażej Ciepłuch, Peter Mooney, and Adam C. Winstanley

Department of Computer Science, National University of Ireland Maynooth, Co. Kildare, Ireland
Tel. (+353 1 708 3847) Fax (+353 1 708 3848)
bciepluch@cs.nuim.ie (corresponding author)

ABSTRACT: OpenStreetMap (OSM) is a very good example of the Volunteered Geographical Information (VGI) paradigm. After a literature review and technology overview of quality assessment in OpenStreetMap we feel that some new methods of quality assessment in VGI such as OSM are required. Currently OSM quality assessment is heavily reliant on "bug checking" and rely heavily upon human interaction to complete. In this paper we provide some alternative methods for quality assessment in OSM. Included in this discussion will be suggestions on how these methods can be implemented, categorization of the difficulty in implementation, and issues of automated quality assessment.

KEYWORDS: OpenStreetmap , Data Quality , VGI quality

1. Collecting Data for OpenStreetMap

Volunteered Geographical Information (VGI) is now an important topic in GIS research. OpenStreetMap (OSM) is a very popular example of VGI. One of the problems affecting the uptake of OSM for mainstream GIS is concerns about the quality of the spatial data stored in the OSM database. The OSM community has developed by consensus a default model of metadata tags for geographical objects (points, lines, polygons, etc). OSM editing software packages implement this model and check for compliance. However contributors to OSM are not obliged to use these tags. They can still add their own metadata. However for more effective data contributions the use of the OSM tag model will assist in creation of: consistent styles in maps generated by tile generators, more effective us in routing applications, and potentially assist in minimizing problems associated with multilingual place and feature names. Another problem affecting the uptake of OSM for mainstream GIS is the variation in user skill and how OSM contributors go about the task of collecting spatial data for the OSM project. As Goodchild (2007) remarks that every person using GPS-enabled mobile devices can be considered as a sensor and he concludes that are "six billions sensors down there" (on earth). While the number of contributors to OSM is relatively small (in the order of tens of thousands) there is variation amongst these contributors. The quality of spatial data collected by a contributor appears to be determined by three factors. Firstly there is the issue of the GIS training and experience of the contributor; how much has this user contributed to the project in the past; and how do they apply the default metadata model to their contributed data, apply their own open metadata model, etc. In this paper we discuss the issues arising from the development of generic quality indicators for OSM data.

2. Quality in OSM

Contributors to OSM can impose their own personal "style" on different objects and features. This is particularly evident where contributors only map certain features (roads, or parks, or bicycle facilities, etc). In some cases problems with tagging and accuracy are carried over consistently in all of a user's contributions. While contributors can trace over aerial imagery (Bing and Yahoo!) using OSM editor software many contributing users collect data manually by field survey. The accuracy of a contributor's data logger may vary considerably from other contributors. This can be particularly problematic in areas where many OSM contributors are operating and all are using different GPS devices. These ground-truth accuracy problems are compounded when contributors using OSM editor

software contribute new features (by GPS trace upload or tracing aerial imagery) in relation to nearby features created by other contributors (where accuracy may be a problem).

Manual quality checking of contributed data in OSM is possible at the post-contribution stage. This manual checking is usually based on other users looking at maps and "hunting for errors". Several tools provide assistance in this manual checking – for example OpenStreetBugs http://openstreetbugs.schokokeks.org/ where we can report potentially erroneous features on the map. It is hoped that at some later stage these problems will be solved by other contributors. As Hudson-Smith et al. (2009) remarks "the law of large numbers dominates in this instance" and most obvious bugs should be corrected quickly. There are other more advanced tools which are semi -automated. One such tool is the turn restriction analyzer http://osm.virtuelle-loipe.de/restrictions/ . The turn restriction analyzer presents all turn restrictions on a map and OSM contributors can then decide which parts of the road networks needs improvement and more data and information for applications such as routing. The turn restriction analyzer can also be used to highlight roads without any name or identification tags. The Tag Watch http://tagwatch.stoecker.eu/Europe/En/ software tool can provide simple statistical output based on the tags and metadata for all features in OSM. It reports on; the most popular tags; number of data points, number of undefined custom tags used etc. This tool can really only be considered as an assistance tool as it is not oriented directly on error identification. There are a number of other similar tools for OSM data. In Kounadi (2009) the author provides a complete list of all these tools together with a description of key uses of each of the tools.

3. A suite of quality indicators for OSM

In comparison to the tools described above and are used by OSM community we are not specifically focused on "error hunting". Our objective is to find some good OSM quality measurement indicators which can give us, potentially quantitative, answers to common queries such as. "Is OSM in particular area good or bad?". When humans look at a map image rendered from OSM data they usually judge for themselves very quickly if the OSM data in the map is "good or bad". These judgements can be based on a number of factors: number of features, spatial distribution of features and data points, map labelling information, etc. OSM is a very unique source of spatial data. Therefore traditional approaches to quality assessment may be irrelevant or need to be modified for the OSM context. The question for the remainder of the paper is how to implement strong and robust quality indicators for OSM data. In the next section we show examples of some sample indicators while other more complicated analysis is described and is part of ongoing work.

Easily implemented quality indicators for OSM

If one has access to a ground-truth dataset then a comparison can be made with OSM for a particular region. A popular quality indicator described in the literature on OSM quality is to compare the lengths of particular features. We have performed this check for Ireland as we had access to an Ordnance Survey Ireland (OSI) 1:5000 data set. Both OSM for Ireland and the OSI dataset were stored in a PostGIS database. A script then automatically measured the lengths of line features (roads, paths, trails, railway, river, etc) for 5KM grid squares. Our script provided output in tabular format which then are easily displayed within a GIS. The grid-squares method for ground-truth comparison of OSM data was first described and evaluated by Haklay (2010) who compared UK OSM against Ordnance Survey GB data. Zielstra et al. (2010) also used this method to compare OSM data set for Germany against TeleAtlas for Germany. However one of the problems with this method is that in many cases we do not have access to a suitable ground-truth datasets for comparison. The reasons for this can include the high cost or licensing terms of proprietary spatial data or that the spatial data is available but at an unsuitable geographical scale. In these cases ground-truth comparison is not possible. In the absence of availability of a ground truth dataset a simple alternative solution is to analyse the density of data points within the grid squares. In OSM every node or point is stored explicitly within the data model. The points for every OSM line, polyline, or polygon feature are available for analysis. While this is a simplification of more complex spatial analysis problems it

provides us with a simple overview of OSM activity for a particular region.

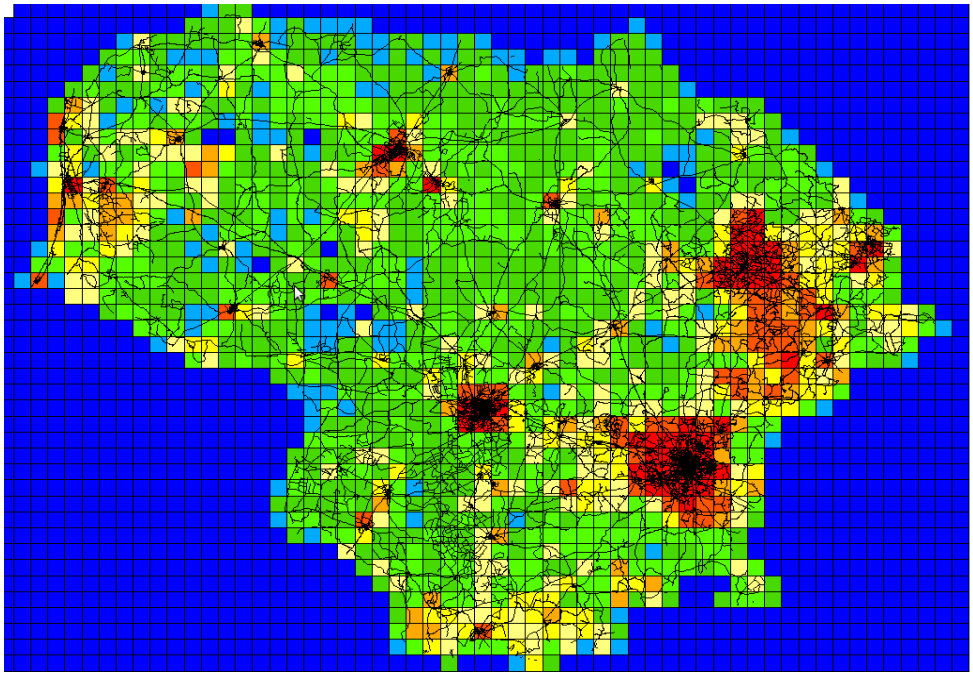

Figure 1. Density of points in OSM Lithuania using 5KM grid squares

Figure 1 shows the density of points for OSM Lithuania. OSM mapping activity is concentrated around areas of high population. Immediately we can see blue grid squares (empty space) in the map. This point density analysis for grid-squares can be configured easily to display other characteristics: density of certain types of features, Points of Interest (POI), frequency of occurrence of specific tags, time since last contribution to each grid-square,. etc. Such information can be valuable in assessment of the quality of the OSM data by verifying where the most intensive OSM contributor activity is occurring. Not all geographical features are traced directly by contributors from available aerial imagery. Some features such as pubs, restaurants, street furniture, building names, signposts, etc require either an OSM contributor to physically visit those locations and record their location or use local knowledge to place these features correctly in the OSM database using an OSM editor software tool.

Advanced quality indicators for OSM

In OSM any contributor can edit any feature in the OSM database and they can collect spatial data about any real world geographic feature and upload this information to the global OSM database. What about the contribution profile or history of a particular OSM contributor? What type of features does this contributor usually edit or collect data on? Do they focus on POI type data collection, natural features, roads, public amenities, etc? For new contributors to OSM these type of questions could be difficult to answer as they may not have contributed much spatial data up to this time.

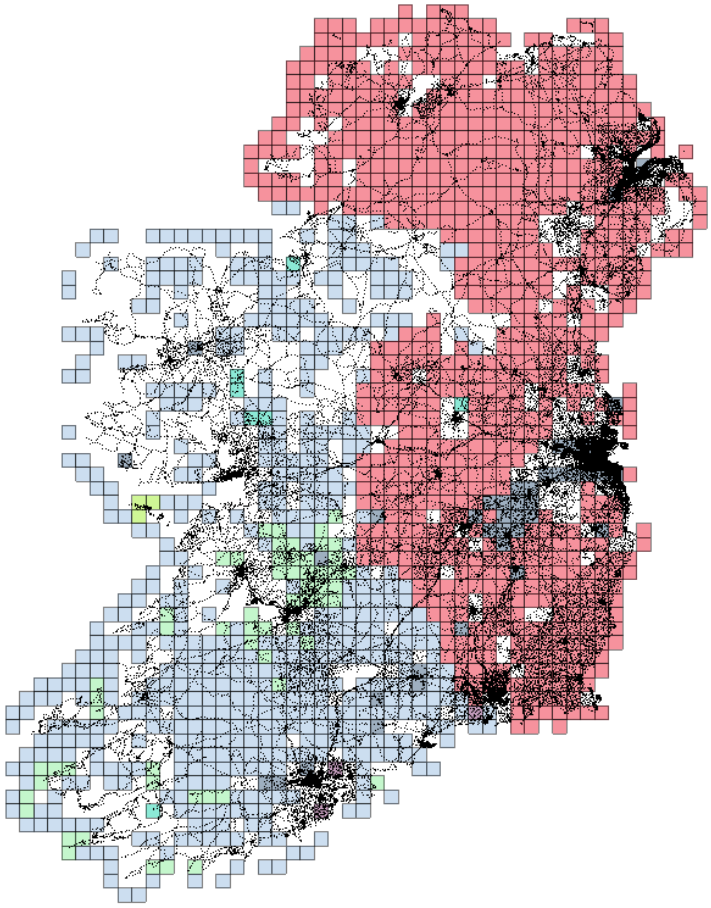

Figure 2. Users (colour coded) who have performed most editing to natural features in OSM Ireland

In figure 2 an example is shown for Ireland (5KM grid squares) which shows the most dominant contributor in each of grid square. Figure 2 shows a unique colour for the user who has edited the most polygons representing natural features in each 5KM gird square. It almost appears that two users have divided the country up into two sections. Given the geographical scale to which these two users have contributed it is probably fair to say that much of their contributions have come through traceing aerial imagery of natural features. In OSM mapping road networks could feasibly be performed at a national level by contributors driving these routes and collecting the data. After the roads have been collected and contributed then the next strata of features are added such as public amenities along the sides of the roads or highways. We are developing software to analyse the order in which features are contributed by analysing the history file for each geographical feature. We believe that the OSM maps grow “from the road outwards”. After amenities such as pubs, restaurants, service stations, buildings etc are added more localised features are added (postboxes, bins, traffic signals, signs, speed bumps, etc). After this strata of features natural features such as parks, fields, grassland, car-parks, etc are added. For this type of analysis it will be necessary to process OSM-XML data. In Figure 3 an extract from the OSM-XML data for Portsmouth UK is shown. Useful information such as contributor ID, timestamp of edit, version number, and tags are provided. Using the OSM API the history file for any object can be accessed. We are also investigating an hypothesis that physical accessibility is highly correlated with data collection for OSM. We will report the results of how far OSM features are from roads – as roads provide access to other geographical features for contributors collecting data by bicycle or on foot. The logical answer is that most features will be close to roads with white spaces appearing in the OSM maps for very rural or difficult to access regions. Most contributors who uploads to OSM usually do so with a GPX file. It is possible to download this files using the OSM API for a particular area. Combining the GPX traces for several contributors in the

same region may provide usual information on how contributors went about mapping that region.

```
<way id="32394317" user="Chris Parker" uid="51722" visible="true"
 version="4" changeset="5432895" timestamp="2010-08-08T13:06:05Z">
 <nd ref="364433886"/>.......
 <nd ref="364433886"/>
 <tag k="amenity" v="university"/>
 <tag k="building" v="yes"/>
 <tag k="name" v="St Michaels Building"/>
 <tag k="source" v="os_opendata_streetview"/>
</way>
<way id="85440" user="IknowJoseph" uid="30587" visible="true" version="5"
changeset="4941609" timestamp="2010-06-09T00:18:04Z">
 <nd ref="163557"/>...<nd ref="194392"/>
 <tag k="highway" v="primary"/>
 <tag k="maxspeed" v="30mph"/>
 <tag k="name" v="Anglesea Road"/>
 <tag k="oneway" v="yes"/>
 <tag k="ref" v="A3"/>
</way>
```

Figure 1: Example of an OSM XML file for two objects in Portsmouth city

4. Conclusions and Final Remarks

In this paper we have described our work in developing a suite of quality indicators for OSM which could be applied consistently to OSM for different regions/databases. VGI, such as OSM, is a unique spatial data resource. Consequently quality indicators must examine a wide range of characteristics. Some of the quality assessments mentioned above are easily implemented and are fully automated with results returned quickly. Other quality investigations such as user profiling will take considerably longer. In future work we will develop a formal categorization of these indicators from which prospective users of OSM can choose an appropriate subset corresponding to the requirements of their application.

5. References

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

Ciepłuch, B.; Jacob, R.; Mooney, P. & Winstanley, A. (2010), 'Comparison of the accuracy of OpenStreetMap for Ireland with Google Maps and Bing Maps"-', Accuracy 2010 Conference.

Hudson-Smith, A.; Batty, M. & Crooks, A. (2009), 'Mapping for the Masses Accessing Web 2.0 Through Crowdsourcing', *Social Science Computer Review* **27**(4), 524-538.

Kounadi, O. (2009), 'Assessing the quality of OpenStreetMap data', Master's thesis, University College of London Department of Civil, Environmental And Geomatic Engineering.

Haklay, M. (2010), 'How good is Volunteered Geographical Information? Acomparative study of OpenStreetMap and Ordnance Surveydatasets', *Environment and Planning B: Planning and Design*.

Zielstra, D. & Zipf, A. (2010), 'A Comparative Study of Proprietary Geodata andVolunteered Geographic Information for Germany', *13th AGILE International Conference on Geographic Information Science 2010*.

6. Biography

Błażej Ciepłuch obtained a B.Sc Degree in Electronics at PWSZ Piła, Poland in 2004 and obtained a Masters Degree in Electronic Automation in Poznań technical university in 2006. In November 2008 he commenced his PhD in Computer Science at NUI Maynooth under the supervision of Dr. Peter Mooney and Dr. Adam C. Winstanley. His research interest include visualization of spatial data, access to environmental information, and quality assessment of Volunteered Geographic Information.