A step towards the improvement of spatial data quality of Web 2.0 geo-applications: the case of OpenStreetMap

Vyron Antoniou[1], Mordechai (Muki) Haklay[2]
Jeremy Morley[3]

[1,2]University College London, Gower Street, London, WC1E 6BT, UK
+44 20 7679 2745
{v.antoniou, m.haklay}@ucl.ac.uk
[3]University of Nottingham, University Park, Nottingham, NG7 2RD, UK
+44 (0) 115 84 68411
Jeremy.Morley@nottingham.ac.uk

KEYWORDS: VGI, OSM, Spatial Data Quality, Quality Assurance, Vector Data

1. Introduction

OpenStreeMap (OSM) is a Web 2.0 system that allow users to create and freely use spatial data. The success of OSM initiative has drawn the attention of scholars and researchers that started to examine issues like the credibility, quality and fitness for purpose of such data (Goodchild 2007, Sui 2008, Flanagin and Metzer 2008, Haklay 2008). In this study we also look into data quality issues (mostly focusing on the attribution of the entities), but from a different perspective. We analyze and formalise the knowledge submitted to the wiki pages by the contributors of OSM regarding the map creation process. This set of rules functions as a user guide for the creation of spatial data for OSM and it can be considered as the 'specification' of the OSM geodata product. After explaining the nature of these specifications, they are used to evaluate the quality of the data created.

2. The OSM Rules

There is a widespread manifestation in the OSM wiki pages that OSM community does not want to impose any rules on its participants. On the contrary, the wiki pages claim that participants can freely use any lawful method and practice to create spatial content and also that they are free to assign any kind and type of attributes (using tags) to real world features (OSM 2009):

"OpenStreetMap does not have any content restrictions on tags that can be assigned to Nodes, Ways or Areas. You can use any tags you like".

In practice though, OSM users have created numerous wiki pages that are full of instructions regarding procedures to describe geographical objects (OSM 2009):

"However, there is benefit in agreeing on a recommended set of features and corresponding tags in order to create, interpret and display a common basemap".

These instructions are not presented as hard and fast rules but rather as lessons from other contributors' experiences or as best practice proposals. Nonetheless, this wiki-made user guide has evolved into a quite complicated and some times hard to follow technical document.

It is interesting to note that the road map to create or change such a rule is totally democratic. In brief, users can start a proposal procedure whenever they feel that a mapping feature should be added or changed. This procedure includes a discussion and a voting step which determines whether the proposal will be rejected or accepted and consequently implemented. The active and approved map

features are documented with proper instructions and both written and visual examples. This is a continuous process; entities from the map features list can be replaced with new ones and the old entities become deprecated.

2. Quality assurance and OSM

This open and democratic process perhaps is one of the key factors for the popularity of the OSM endeavour. However, such freedom created a lot of inconsistencies and therefore there was a need for some form of quality assurance mechanism that would enable users to correct inaccuracies. Indeed, today there is a variety of options for an OSM contributor to achieve that: from assigning a simple "fixme" tag to a feature to indicate that it requires updating to using one of the dedicated applications for identifying errors in the database (the list of those applications is available at http://wiki.openstreetmap.org/wiki/Quality_Assurance).

These early attempts for identifying and correcting mistakes in the OSM dataset, while they present interesting paradigms of a self-correcting mechanism for a crowd–sourced, Web 2.0 application, are still incomplete and patchy. The 'Keep Right' application that was created by the OSM community and monitors the violation of some OSM rules provides an example for this (Figure 1). The application evaluates the data conformance against a set of pre-defined rules and presents to the users the positions of possible mistakes.

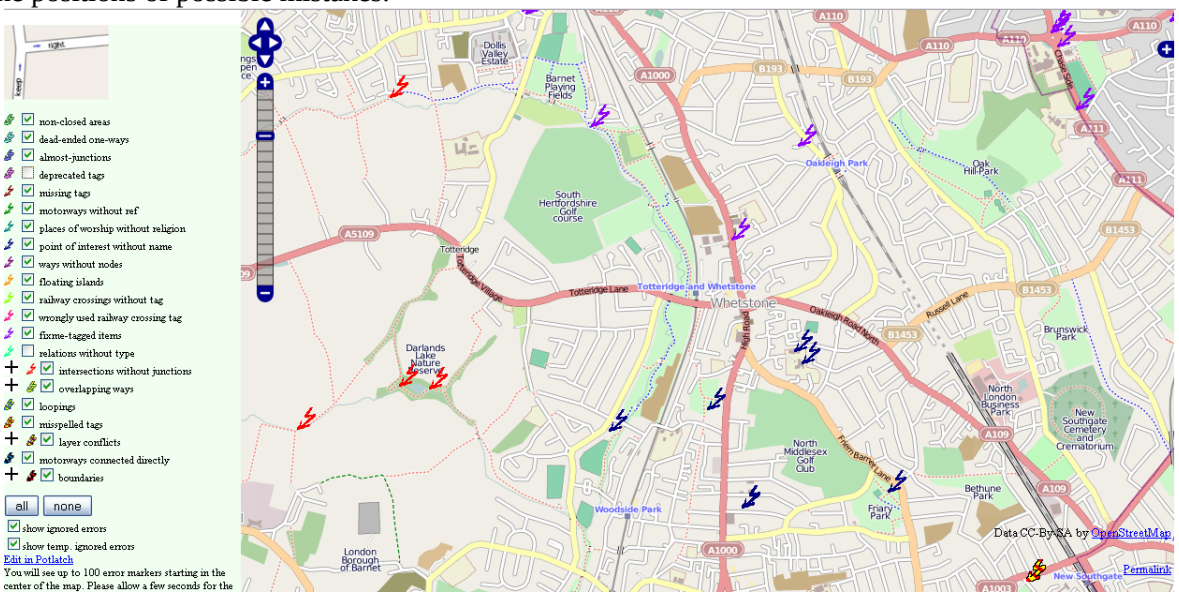
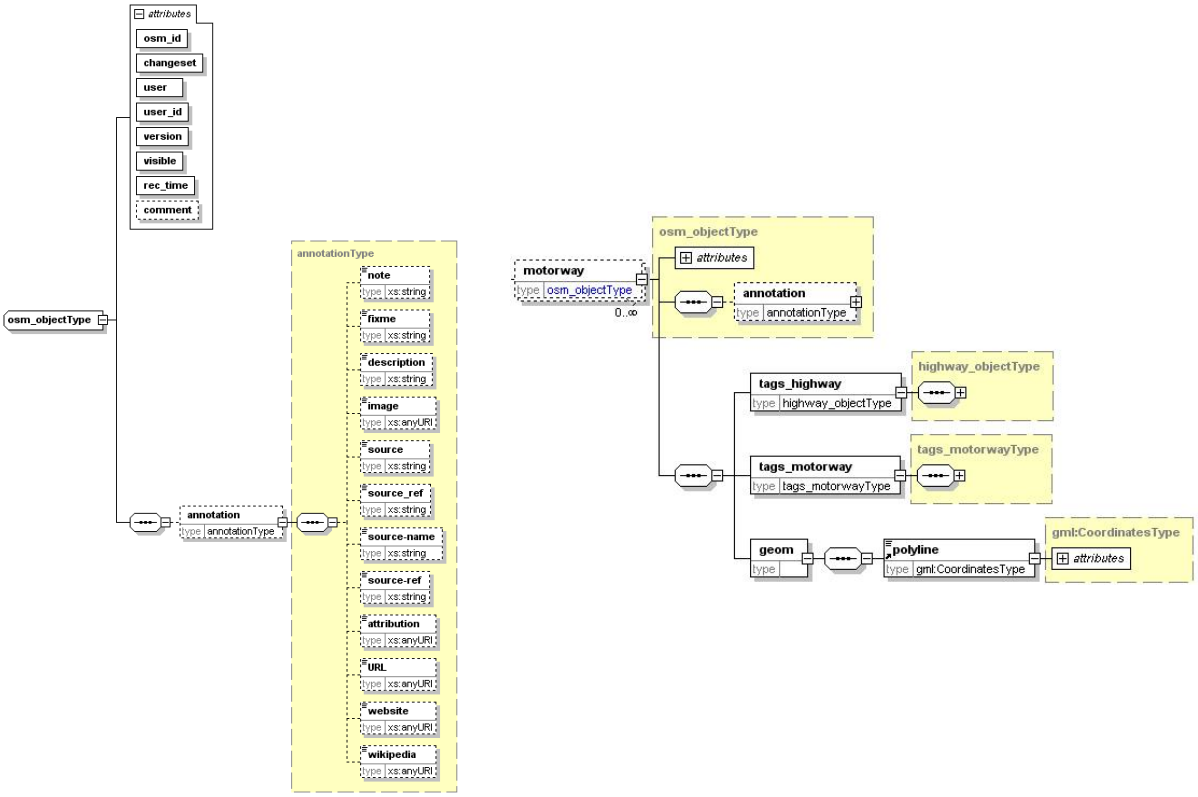


Figure 1. The Keep Right application. The table of contents in left hand side contains the rules monitored by the application.
(Source: Keep right 2010)

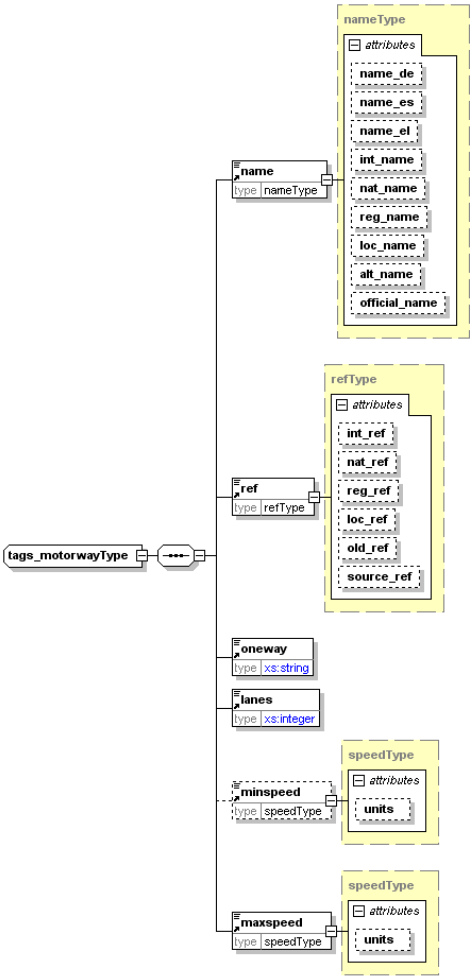
The question that needs to be answered at this point is *"How these specific rules have been chosen?"* and most importantly *"Why only these rules? Do they provide a quality assurance mechanism for the entire OSM dataset? If no, what should we do? "*. In fact, these early efforts add to the argument that there is a strong need for a more holistic approach when it comes to building quality assurance mechanisms and determining the spatial data quality for data generated in GeoWeb applications.

Nonetheless, from the standing point of someone who wishes to use the OSM data there is no clear indication regarding the quality of that product. This stems from the fact that there are no solid specifications that will help to determine the magnitude of the product conformance. Therefore, we

propose an XML Schema able to model the OSM entities as described in the Map Features list (http://wiki.openstreetmap.org/wiki/Map_Features). For example, Figure 2a shows the object type of an abstract OSM object according to the existing rules, while Figure 2b shows the object type of an abstract Motorway object.



(a) (b)



(c)

Figure 2. (a) The XML Schema fragment that shows the attributes of an abstract OSM object (b) The XML Schema fragment for motorways, (c) The XML Schema of the motorway tags.

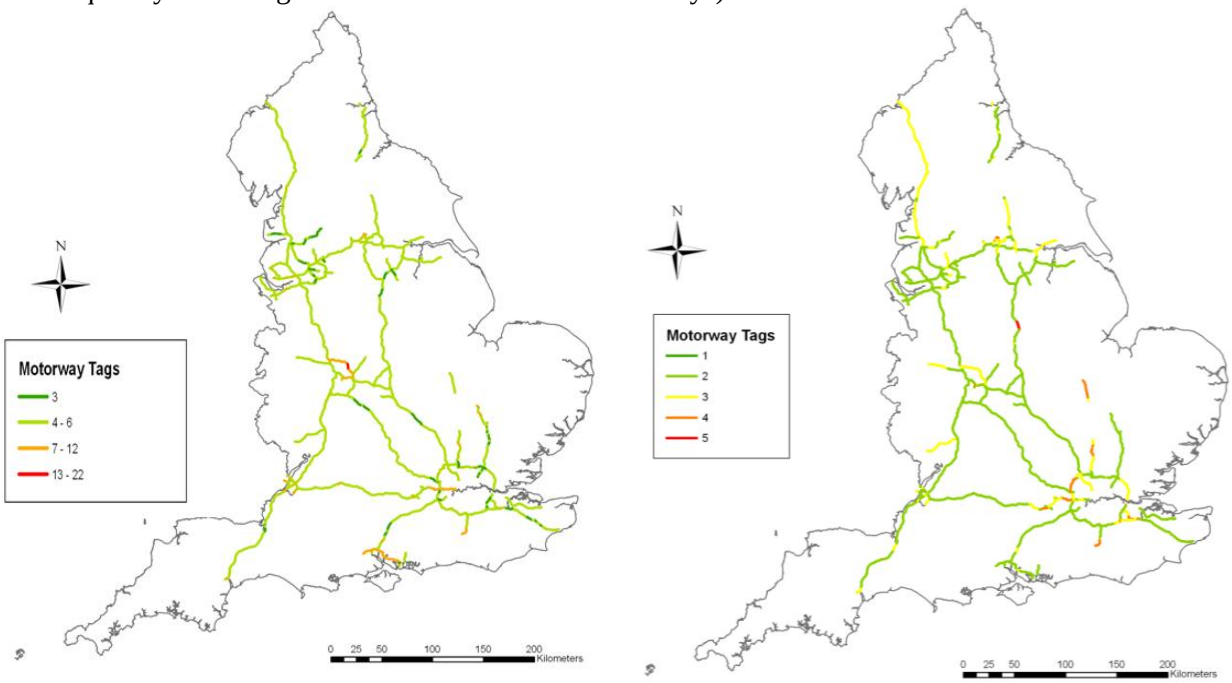
It is important to emphasize that this XML schema has been constructed by following the rules that are described in the OSM wiki pages. Consequently, this XML Schema cannot be regarded as a static document but rather as a set of rules that is constantly synchronized with the rule creating system of OSM.

By formalizing and applying a common language to the rules that affect the creation of the OSM map features it is possible to analyze and measure the level of conformity or violation of the OSM database against those rules. We used the XML schema (and more specifically the XML fragment of each entity as shown for example in Figure 2b for Motorways) in order to examine the quality of OSM features for England. In this effort we followed the methodologies documented in the ISO 19113 (ISO 2002) and ISO 19114 (ISO 2003) standards.

4. Findings and conclusions

Our findings show that wherever there is a systemic way (either a database schema or elements in the user interface of OSM editing applications such as JOSM and Potlach) to handle the data input or to guide user contribution, then the data quality is high. On the contrary, in the cases where there is no guidance other than the wiki pages the quality is considerably lower. For example, Figure 3a shows the total number of tags assigned to each motorway while Figure 3b shows the number of XML

Schema conforming tags that have been assigned to motorways. It can be seen that while OSM contributors have assigned up to 22 tags to a single motorway element, in fact for the majority of the motorways only 2 to 3 tags are valid according to the guidelines published at the wiki pages (and consequently according to the XML Schema for motorways).



(a) (b)

Figure 3. (a) The number of XML Schema conforming tags for each of the motorway entities in England and (b) the total number of tags assigned to each entity.

In other words, it is not uncommon to witness the violation of OSM rules, despite the fact that all these rules have been established with the open and democratic procedure which was described earlier. Finally, it should be noted that since data quality changes whenever there is a change in the data (e.g. due to a transformation), in the ground truth or in the specifications of the product, we have witnessed that constant changes of OSM specifications through the voting system affects negatively the overall quality of the dataset. For example, before the decision to deprecate the "created_by" tag from all the entities in OSM the 22.44% of the Motorways in England violated the conceptual consistency of the XML Schema. This figure climbed to 80.62% when the rule and thus the XML Schema changed.

Therefore, we conclude that for GeoWeb applications, the first step towards the improvement of their data is firstly to conceptually formalize the data sought. This step will enable them to create the necessary environment both in the front- and in the back-end of the application so to help contributors to conform to the chosen schema. In turn, this will improve data quality by diminishing errors and inconsistencies in the dataset. Additionally, whenever there is a change in the specifications the formalisation can facilitate automatic correction of the existing data.

Finally, it is important to note that, such a process does not affect in any way the openness of a Web 2.0, crowd-sourcing application or the excitement and sense of freedom that volunteer geographers feel when contributing to such applications. By doing so, we can have both the openness and the formalization needed to achieve a crowd-sourced dataset of high quality.

References

Flanagin A., Metzer M., 2008. The credibility of volunteered geographic information.

GeoJournal,72: 72:137–148.

Goodchild, M. F., 2007. Citizens as sensors: The world of volunteered geography. GeoJournal, 69(4), 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. [Online] Available at: http://www.ucl.ac.uk/~ucfamha/OSM data analysis 070808_web.pdf [Accessed 20 November 2009]

ISO (International Standardisation Organisation), 2002. ISO 19113:2002 Geographic information — Quality principles.

ISO (International Standardisation Organisation), 2003. ISO 19114:2003 Geographic information — Quality evaluation procedures.

Keep right. 2010. Data consistency checks for OSM [Online] Available at: http://keepright.ipax.at/ [Accessed 20 February 2010].

OSM. 2009. Map Features. [Online] (Updated 12 Oct 2009) Available at: http://wiki.openstreetmap.org/wiki/Map_Features [Accessed 20 November 2009].

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

Biography

Vyron Antoniou is a Captain in the Greek Army and since 1998 serves at Hellenic Military Geographical Service. Currently is a PhD student at UCL in the department of Civil Environmental and Geomatic Engineering. His research interests are in user generated content and National Mapping Agencies; spatial databases; vector data transmission over the Web and web mapping.

Mordechai (Muki) Haklay is a senior lecturer in Geographical Information Science in the department of Civil, Environmental and Geomatic Engineering at UCL, where he is also the Director of the Chorley Institute, which focuses on facilitating interdisciplinary geospatial research at UCL. His research interests are in public access to environmental information, Human-Computer Interaction (HCI) and Usability Engineering for GIS, and Societal aspects of GIS use. He received his PhD in Geography from UCL.

Jeremy Morley is Deputy Director of the Centre for Geospatial Science at the University of Nottingham and was previously a lecturer in Geographic Information Systems at University College London. His interests lie in GIS interoperability; interfaces between GIS web services and mashup technology; sensor webs; and planetary mapping."