Mapping the technical challenges regarding OSM, with examples from the norwegian open and free governmental data

Anne Sofie S. Erichsen

October 18, 2016

Contents

List of Figures

List of Tables

Abstract

A big problem with OSM is finding a good, fast and easy way of correcting data. There exists multiple tools that finds possible errors, but there also has to be a good way of fixing them. A solution to this problem can be can be microtasking. Microtasks are formed basically by dividing a project into smaller tasks, clearly defined, that can be performed independently [Estellés Arolas,].

Chapter 1

Introduction

This paper will inform you on the technical challenges regarding OpenStreetMap and governmental data [Exel et al., 2010]. For a summary of papers you can see on page 8

In theory micotasking seems to solve a lot of OSM problems like overlapping data, deletion of good metadata (read: tags) when running import-scripts and makes it easier to control, both the workflow and quality of the data. Microtasking splits a task into multiple subtasks and distributing these subtasks to humans over the internet. The OSM mindset of schema-less datasets and tags differs from many organizations. With the success of OSM it is time to start taking this mindset serious. OSM also has its weaknesses , but many people believe microtasking solves the majority of them. By using FKB building dataset I will try too look further into mapping governmental data over to the OSM format, also trying to experience if microtasking is the solution of the weaknesses of OSM, like the ones I mentioned above.

Chapter 2

Characteristics of Open Street Map

2.0.1 Introduction

The OpenStreetMap is one of the most impressive projects of Volunteered Geographic Information on the Internet[Neis and Zipf, 2012]. Until recent the mapping of the Earth was preserved highly skilled, well-equipped and organized individuals and groups. One important happening was in 2000 when Bill Clinton removed the selective availability of the GPS signal 4.0.2. This change improved the accuracy of simpler, cheaper GPS receivers so that also ordinary people could start mapping their movements. OpenStreetMap was founded in 2004 at University College London by Steve Coast. The goal was to create a free database with geoxgraphic information of the world [Neis and Zipf, 2012]. Back in 2004 the geographic data was expencieve and hard to get access to.

The OSM project stands out from other data sources mainly because its free to use and its released under a license that allows for pretty much whatever the user wants to as long as the user mention the original creator and the licence[Chilton et al.,]. The most common contribution approach is to record data using a GPS receiver and edit the data using one of the free and available OSM editors [Neis and Zipf, 2012].

Today the world has a need for instant information, particularly in crisis situations [Chilton et al.,]. Here OpenStreetMap is the leading global example of the effictiveness of crowdsourcing of geodata. The project are changing the way individuals and

organisations are thinking about the collection process, purchase and use of geodata [Chilton et al.,]. Crowd sourced geographic data has characteristics or advantages of large data volume, high currency, large quantity of information and low cost [Wang et al., 2013].

2.0.2 Structure

OpenStreetMap uses a topological data structure. This sturcture includes three basic components nodes, ways and relations. Nodes are points with a geographic position stored as coordinates (Lat, long) according to WGS84. Ways are lists of 2 or more nodes, representing a poly-line or polygon, used to represent streets, rivers, among others [Debruyne et al., 2015]. A relation is a multi-purpose data structure that documents a relation between two or more components. To add metadata to geographic objects OpenStreetMap uses Tags. Tags consist of two items, a key and a value of the form key=value. The key is used to describe the topic, category or type of feature, while the value describes the details of the specific form of the key specified. A example of a key-value pair can be building=church, here the key is building and the value is church, this is a building that was built as a church.

The norm in OSM is to try to map new data with existing tags. Good practice is to search for tags, or Map Features, on different OSM wiki-sites. On the tags you like wikipage they recommend different sites, but points out taginfo as the most useful site. Taginfo is a website created for finding and aggregating information about OSM tags, it covers the whole planet and is updated daily. The web page list tags used in the database and also inform on how often they have been used. Also, Taginfo lists other tags which have been used in combination with the tags you searched for. Some countries also have their own taginfo web pages, like Ireland, Great-Britain and France, Norway do not have their own taginfo web page.

Verifiability: From a given scenario, a tag/value combination is verifiable if and only if independent users when observing the same feature would make the same observation every time..

2.0.3 Organizational

Organization and communication

The OpenStreetMap Project is supported by the OpenStreeMap Foundation (OSMF) which is a UK-registered non-profit organization. The foundation was founded in 2006 and consists of members from all over the world, as of December 2015 consist of 350 normal-, 351 associate- and 18 corporate members [OSMF, 2015]. OSMF include a board of seven members and is critical to the ongoing function and growth of the OpenStreetMap project [OSMF,]. The foundation has the responsibility for the servers and services necessary for hosting the OSM project. Also, they support and communicates with the working groups, and delegates tasks that has to be done, like Web site development etc.

A person can contribute to the OSM project without being a member of the foundation. The project has over 3 million registered users [OSM Project, 2016] who are collecting and updating data. The crowdsourced data are then released under the Open Database License, *"a license agreement intended to allow users to freely share, modify, and use this Database while maintaining this same freedom for others"* [ODbL,]. Users can edit maps through different tools made by different OSM contributors. One tools is called iD and is the default web browser editor written by MapBox. There are also desktop editing applications like JOSM and Merkaartor which are more powerful and better suited for advanced users.

Communication is country-mailing lists, wiki-pages and conferences. State of the map is the main OSM conference. Number of mappers and organizations are constantly increasing, what started as a crazy hacker project is now a vital part of the global data ecosystem. The community are constantly developing new ways to contribute. Users can join a mapathon in their hometown, they can sit at home adding data,

The degree to how you can get involved in OSM is so deep, mapping, software processes etc. - keeps people interested. One problem is the communication through the different groups, the energy level is not high enough. Lots of people exited about communication, everyone have a obligation to show the users their different possibilities.

Communication through mailing lists works for the people who subscribes to that list, but with over 150 different list it is impossible for an interested user to stay updated with all the latest achievements. What is happening in the mailing lists has to be available to everyone in the community. A solution to this is called weeklyOSM. It started in 2010 in German, who alone has 50 different mailing lists. The weeklyOSM team are scanning mailing lists, twitter, blogs and so on. There are a international team translating the German blog into different languages. Little by little the rest of the community gets involved, translating the blog into languages

thats not supported.

There are different groups creating the different versions. The goal is to integrate more languages. There are a lot of work every week, hard to find volunteers.

2.0.4 File format, .osm files

The .osm file format is specific to OpenStreetMap and it is not easy to open these files using GIS-software like QGIS. The file format is designed to be easily sent and received across the internet in a standard format. Therefore .osm files are easily obtained, but using the files directly to do analysing and map design is not easy. The .osm files are coded in the XML format. It is recommended to convert the data into other formats when using the files source.

The file is very difficult

Chapter 3

Technical

3.0.1 Existing libraries

The internet consists of hundreds of software libraries and packages. It can be overwhelming for newcomers and hard to find the most suited ones. A good tip is to learn the handful of libraries and packages that most software is derived from, so called root libraries. They are actively maintained and not significantly derived from any other libraries. The libraries do geospatial operations that are hard to implement, so people choose to use the libraries instead. Geospatial datasets are large, often complex and varied. This makes the implementation harder, and some of the reasons for the libraries success. The root libraries are GDAL, OGR, GEOS and PROJ. 4 [Lawhead, 2013]. They are, according to J. Lawhead, "the heart and soul of of the geospatial analysis community". All the libraries are written in C or C++.

Chapter 4

Summary of papers

4.0.1 Quality analysis of open street map data [Wang et al., 2013]

Crowd sourcing geographic data is an opensource geographic data which is contributed by lots of non-professionals and provided to the public. Compared with conventional data collection and update methods, the crowd sourced geographic data has characteristics or advantages of large data volume, high currency, abundance information and low cost and becomes a research hotspot of international geographic information science in the recent years.
The primary problem is to analyse the quality of crowd sourcing geographic data. There are three factors that influence the quality of OSM: First data collected and mapped by non-professionals, secondly the collected data may be from different data sourcing with different precision and thirdly the data collected by different GPS may have different precisions.
The paper assesses three quality elements: 1. Data completeness, 2. Attribute accuracy and 3. Position accuracy.

Enter twice to get spacing between paragraphs.

Data completeness Includes length completeness and name completeness. Length completeness is the geometric quality and data coverage. Formula: $Q_L = \frac{L_{OSM}}{L_R}$
(L = percentage of the length).

Name completeness Name completeness means the completeness of the name attribute.

4.0.2 User generated Street Maps [Weber, 2008]

Until recently the mapping of the Earth was preserved highly skilled, well-equipped and organized individuals and groups. The big change happened in 2000 when Bill Clinton removed the selective availability of the GPS signal, this provided much improved accuracy for simple, low-cost GPS receivers. The wide availability of high-quality location information has enabled mass-market mapping based on affordable GPS receivers, home-computers and the Internet.

OSM background OSM follows the peer production model that created Wikipedia. Its aim is to create a set of map data thats free to use, editable, and licensed under new copyright schemes. Was founded in 2004 at University College London. In May 2008 OSM had more than 33,000 registered users and about 3,500 currently active contributors. OSM decided to follow the route of allowing only registered users to edit the map, this way they can trace the information source. OSM GeoStack is the set of tools that lets users capture, procedure, communicate, aggregate, and consume the geographical information produced in the project.

Editing Tools OSM developers implement tools to facilitate user contributions to the database. In 2008 they had a Flash-based editor called Potlatch. Today JOSM is more common, even though this is used by more experienced OSM contributors. At the end of 2006, Yahoo granted OSM the right to use its satellite imagery to trace roads and other features.

4.0.3 CROWDSOURCING IS RADICALLY CHANGING THE GEODATA LANDSCAPE: CASE STUDY OF OPENSTREETMAP [Chilton et al.,]

Examining the effect of the changing cartography has on data collection using OSM crowdsourcing as a case study. Are parallels to what is happening with data collection in other aspects, like WIKIPEDIA, Flickr and YouTube.

- Today: A need for instant information, particularly in crisis situations
- User Generated Content providers / crowdsourced data collectors are allowed to collect geodata
- Reason: More available satellite imagery, cheaper GPS units, etc
- OSM the leading global example
- OSM was the first online mapping service to accurately map and display the new London Heathrow Terminal 5, on the official opening day.
- OSM project - the achieved coverage, its accuracy, availability and global impact are all changing the way individuals and org are thinking about the collection, purchase and use of geodata.
- Makes OSM stand out from other data sources:
- Completely free of charge
- Is released under a license which allows you to do pretty much what you like
As long as you mention the original creator and the license
- The availability, accuracy and price of OSM data has lead some local authorities in the UK to question the need to have total reliance on being locked into a contract for their geodata with the National Mapping Agency.
- OSM gives the possibility of having really current data available, other services may have a large lead for getting the data from survey to map output.
OSM have specialist maps for cycling, routing, applications, skiing, topography and for maritime use.

4.0.4 The impact of crowdsourcing on spatial data quality indicators

Introduce the concept "Crowd Quality" (CQ) to describe and quantify the quality of crowdsourced geospatial information. Together with the growth in volume, the usage of crowdsourced geospatial info. grew extensively as well.

- Quality: Has a meaning if we have a common understanding of its definition
- ISO19113(2002): Quality is the "totality of characteristics of a product that bear on its ability to satisfy stated and implied needs".
- Van Oort: Identified eleven elements of spatial data quality. The elements are used to describe the quality of geo-data collected and produced.
- Uniform method to produce and process the data –¿ Homogenous and consistent quality
- Crowd Quality (CQ):

- 2 dim.:
1. User-related quality aspects: Quality of information from an individual contributors perspective. This is the typical char. of crowdsourced data.
2. Feature-related quality aspects: Perspective of the spatial feature.

1. 3 components: Local Knowledge, Experience and Recognition.
a. Familiarity to an area can be correlated to the quality of the users contribution
b. Quality of a users contribution is correlated with his overall experience in contributing to the project
c. Online social networks that allow for user contributions, often feedback are established (ratings, recognition for specific contributions). This type of User recognition is largely unknown in crowdsourced geospatial data –¿ Puts a strain in our ability to assess CQ.
2. In crowdsources datasets quality elements can be different for similar features, one user can add his personal attributes when adding a restaurant in OSM, while another restaurant only have one attribute. In traditional datasets the quality elements will be uniform, have the same attributes.
a. Quality elements that are particular interesting, since they are not consistent for crowdsourced data:
1. Lineage, positional accuracy and sematic accuracy.
When data is imported from other sources, these imported features have a very clear lineage with regard to positional accuracy and precision. When imported from GPS data, the positional accuracy is harder to establish.
Sematic accuracy - related to the completeness and internal consistency of the attribute metadata. A schema for attribute metadata is not common in crowdsources geospatial data projects –¿ Create a threat to internal consistency.

Chapter 5

Movies

5.0.1 State of the Map, microtasking

A bot is a tool that carries out repetitive and mundane (dagligdagse) automated edits on a regular basis to help maintain OSM. Most bots deal with tagging, ex xybot. BugBuster deals with removing nodes. Czechreg deals with changing geometries.

Issues with importing: The person need to have experience with the tool and the data, but also need to have the time to manually verify each task. Therefore we have microtasking - A process of splitting a task into multiple subtasks and distributing these subtasks to humans over the internet. He things this is the way of the future. Breaking up tasks tools - OSMLY, MapRoulette, BattleGrid. They can solve our issue, but we need different tools. One to determine conflicting objects, saving the tasks in a backend and then make a usable and easy frontend to work on the tools. OSM Tasking manager, divide up a mapping job.

Microtasked conflation may lead to standard procedures that can help other groups import data. Can help provide a validation step to force someone to look at each and every contribution.

Bibliography

[Chilton et al.,] Chilton, S., Building, F., and Burroughs, T. CROWDSOURCING IS RADICALLY CHANGING THE GEODATA LANDSCAPE : CASE STUDY OF OPENSTREETMAP.

[Debruyne et al., 2015] Debruyne, C., Panetto, H., Meersman, R., Dillon, T., Weichhart, G., An, Y., and Ardagna Afostino, C. (2015). On the Move to Meaningful Internet Systems: OTM 2015 Conferences. *Springer International Publishing*.

[Estellés Arolas,] Estellés Arolas, E. Microtasking: a way of begin using crowdsourcing — My crowdsourcing blog.

[Exel et al., 2010] Exel, M. V., Dias, E., and Fruijtier, S. (2010). The impact of crowdsourcing on spatial data quality indicators. pages 1–4.

[Lawhead, 2013] Lawhead, J. (2013). Learning Geospatial Analysis with Python. *Packt Publishing*, (October).

[Neis and Zipf, 2012] 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(3):146–165.

[ODbL,] ODbL. Open Database License (ODbL) v1.0 — Open Data Commons.

[OSM Project, 2016] OSM Project (2016). Stats - OpenStreetMap Wiki.

[OSMF,] OSMF. About — OpenStreetMap Blog.

[OSMF, 2015] OSMF (2015). Membership/Statistics - OpenStreetMap Foundation Wiki.

[Wang et al., 2013] Wang, M., Li, Q., Hu, Q., and Zhou, M. (2013). Quality analysis of open street map data. *8th International Symposium on Spatial Data Quality*, 40(June):155–158.

[Weber, 2008] Weber, P. (2008). User-Generated Street Maps. pages 12–18.