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

Anne Sofie S. Erichsen

October 31, 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 **??**

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 **??**. 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 Culture

OSMhas no notability rule, an arbitrary amount of detail is possible, but somebody has to maintain it!

2.0.3 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..

2.0.4 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 [OpenStreetMap, 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.

2.0.5 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

OSM import methods

4.0.1 Introduction

The traditional way to contribute data to the OpenStreetMap project is through active users who use their GPS to track roads and their local knowledge to add information about their geographic regions to the OSM database [Zielstra et al., 2013]. Users also digitalize aerial photos. Cheaper GPS receivers and more available satellite imagery with better resolution makes it easier for users to contribute [Chilton et al.,]. The number of active users in different regions varies a lot, making some areas on the OSM map full of data while others are almost empty. This led to a second approach for getting data in to the OpenStreetMap database, bulk imports [Zielstra et al., 2013]. Bulk imports is the process of uploading external data and were meant for initial/preliminary object class uploads, so only if the object were none existing in the area [Zielstra et al., 2013]. Its a good alternative for countries or regions with less active users. Through the years different import methods has been developed. This paper will evaluate the most common methods.

4.0.2 Fully automatic import script

Creating a script that automatically imports big datasets into OpenStreetMap, a bulk import, is not encouraged in the OSM community [Zielstra et al., 2013]. This becomes clear when reading the wikipages about import. A bulk import is suppose to be a supplement to user generated data. The user generated data and the users ability to work is always the priority [OpenStreetMap, c]. A fully automatic import do

automated edits to the OpenStreetMap database with little, if any, verification from a human. Automatic edits is changes that has no or very limited human oversight [edits OpenStreetMap,]. This kind of edits must follow the Automated Edits code of conduct [OpenStreetMap, a]. The policy was created to prevent damaging acts on the database and ignoring it will result in the import be treated as vandalism.

An example of a bulk import was the TIGER import. The Topologically Integrated Geographical Encoding and Reference system (TIGER) data was produced by the US Census Bureau and is a public domain data source. The bulk import was completed in early 2008 [Zielstra et al., 2013], populating the nearly empty map of the United States. The TIGER data was not perfect and had it's faults, but it was better than no data at all [Willis, 2008].

Reading through the OSM talk mailing list we can see that this import is a automatic edit with very limited oversight. *"We are processing 6.502308 OSM objects per second"* [Munro, 2007]. Active OSM users on the mailing tread tries to save their manuall work *"Please don't upload Northampton County, however, since I've mapped my entire town there [..]"* [Mielczarek, 2007]. Users not attenting the mailing list had limited, if any, possibilities of their work beeing saved through the import. The users importing TIGER data tried not to override existing data. They started on empty states but reading though the "TIGER, which states next?" mailing thread the decisions on which state to be imported next was only based on the people attending the conversations. *"I believe I'm the only one out here in Nebraska [..] Feel free to override my edits"* [Bishop, 2007]. The import process did not have any requirements on what they should do with existing data. Users added the TIGER data county by county, state by state. The OSM user Dave Hansen, one of the most active users during this import, created a text file with his upload queue [Hansen, 2007b] where he added states and counties after requests from users [Hansen, 2007c][Hansen, 2007a]. The import team did not have any validation or correction methods or routines. In the aftermath of the import tagging errors have been discovered. For instance, the TIGER data groups residential, local neighbourhood roads, rural roads and city streets into one road class, while OSM uses a more refined data schema with different highway tags for different road classes [Zielstra et al., 2013].

On the TIGER OpenStreetMap wikipage, last updated August 2016, they say it is unlikely that the TIGER data ever will be imported again. The main reason is the growing US mapping community, their mapping is often better than the TIGER data. *"Do not worry about getting your work overwritten by new TIGER data. Go map!"* [OpenStreetMap, 2007]. A new bulk import with updated TIGER data can overwrite existing, more precise data. The TIGER data are of variable quality, poor

road alignment is a huge problem and also wrong highway classification. Many hours of volunteer work could be lost and this is something the community want to avoid. The bulk import in 2007 got the United States on the OSM-Map and saved the mapping community a lot of time finding road names etc. *"TIGER is a skeleton on which we can build some much better maps"* [Willis, 2007]. On the negative side the project kept the US mappers away, they were told for years that their work was no longer needed after the TIGER upload was complete. But the presence of TIGER data ended up requiring a lot of volunteer help, they needed help fixing errors like the poor road alignments and wrong tagging.

Bulk imports are overall not recommended today, but have been helpful as well. In the Netherlands bulk imports have meet little resistance, mainly because the imports are done by dedicated OpenStreetMap mappers who knew the OSM import guidelines. Arguments in favor of bulk imports say that a map that already contains some information is easier to work on and can help lower the entry barriers for new contributors. Another argument is that a almost complete map is more attractive for potential users, that again can encourage more use of OSM data in professional terms [Exel van, 2010]. But a huge minus to bulk imports are the data aging, since the data being imported often already is a few years old and updating it takes time, often years. The TIGER import was data from 2005, but the import finished in 2008 [Zielstra et al., 2013]. Between the time of first import and update the community have fixed bugs, added important metadata, the community would not want to loose that data/information.

Today there are huge amounts of object types in OpenStreetMap. Often bulk import overrides existing data which is one of the "don't do" points on the import guidelines list on the osm wiki page. OpenStreetMap do not have layers, so data on top of data makes it very difficult to organize and find the data.

4.0.3 Fully manually import based on tracing

This method can be very time consuming. The mapping quality depends on the image resolution in the area beeing mapped, it is also hard to add metadata from a image. For instance, its impossible to see the height of a building from a satellite image.

Haiti project, and other Humanitarian OSM project, draw from satellite image dry. A huge problem with this is when during a crisis, many users map the same areas.

During Haiti project a problem was overlapping data, the same road drawn multiple times. This was before tasking manager.

This method is also used today. Humanitarian OSM use the method with the tasking manager. Then the problem with overlapping data are not as likely to occur. This import method do not require any mapping skills

- 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

4.0.4 Guided automatic import

Fully automatic import of huge amounts of data is discouraged in the OSM community, so another approach is guided automatic import. The OSM community encourages people to import only small amounts of data at a time and only after validation and correcting errors [Mehus, 2014]. This method was used when the OSM-community in Norway got approval from the Norwegian map authority to import N50 data [Kihle, 2014]. N50 is the official topographic map of Norway. The import process is described on the OSM wikipage. It says that they will import one municipality at a time. Each municipality dataset will be divided up in a .15 deg times .15 deg grid changeset, each changeset contains from five tousen to tventy tousen elements [OpenStreetMap, 2014]. The N50 import was an community import, but only experienced user were encouraged to import the data [Mehus, 2014].

The norwegian OSM group started importing the N50 data before they had consultet with the OSM imports mailing list, which is required. This was pointed out by DWG member Paul Norman [Mehus, 2014]. DWG is the data working group and they are authorized by the OSMF to detect and stop imports that are against the import guidelines [OpenStreetMap, b].

The N50 release was good news for Norway, since regions, especially in northern parts of Norway, had very little data because of few active OSM mappers. This import would then increase the quality of OpenStreetMap in much bigger parts of Norway kilde. Its a huge dataset so they had to import it with caution, not everything in the dataset is not relevant for OpenStreetMap, this is noted by the OSM user Solhagen hrefhttp://wiki.openstreetmap.org/wiki/User:Solhagenkilde. Another approach is to convert the sosi files to shp through GDAL and import this shp file into JOSM. This approach is not recommended, converting sosi to shape will result in loosing

important data. SOSI uses a hierarchical meta data structure

The data was preprocessed by the OSM user tibnor and uploaded to a google drive folder available for everyone. He started on the preprocessing early 2015. Late 2013 the OSM user gnonthgol* created sosi2osm script for convertion between the norwegian dataformat SOSI and OSM file format kilde. This script is the recommended way of converting the N50 sosi files to the OSM file format. In may 2015 tibnor released a JOSM plugin for the N50 import of rivers/streams to make it faster to check and fix directions. The import process was managed through a wiki page. Here users wrote their name and progression, startdate and enddate of the imported municipality. Elements which needed manual inspection or validation was tagged with "Fix-me" and a description on what to do. They used a python script ("replaceWithOsm.py") to merge N50 data with existing OSM data, adding source=Kartverket N50 tags on the new N50 data. Elements that already exists in OSM, that conflicts with the new data, is marked with FIXME=Merge. Here the user has to search for the conflicting elements and correct the errors manually. Then the data can be uploaded to OSM.

The N50 import was initially stopped in May 2014 by Paul Norman. Some Norwegian users had started importing the data without the proper approach. The Data working group reverted the import because of technical problems [Didriksen, 2014]. This was a step back for the Norwegian OSM team and they had to start over again. It was probably for the best, the DWG group has much experience with automated imports. This process was time consuming, started in 2015 and is still not finished, even though most of the municipalities are done.

4.0.5 Import based on LA-buildings methodology

The N50 import was a good start towards microtasking a huge data import. They divided every municipality into .15 degree .15 degree grids and importet one grid at a time. The LA building import took this mindset to the next level, creating a Tasking manager interface specific tot this import, among other initiatives.

OSM Tasking Manager was created in the aftermath of the Haiti earthquake. This innovation coincided with the growing popularity of microtasking as a solution to manage distributed work

The Guided automatic import from 4.0.4 we saw that dividing datasets into smaller parts makes the import easier to, among others, distribute the workload between

experienced users. OSM Tasking manager takes this approach further so that users can work on the same time with nearby areas. The tasking manager divides the areas into grids and use colors to inform the user if the grid is done, locks it if someone is already working in the grid and also the user easily can see the commit story when they click a grid.

4.0.6 Evaluation

OpenStreetMap is a large community dependent on active users adding data in their geographic regions. The users have different perspectives on importing data through scripts. In empty regions with few active users a bulk import can be good way of developing the OSM map. But in regions with large amounts of data, there is no agreement on whats best. Little research has been done on countries where OSM relied on data imports. The OSM community is undecided on the benefits of bulk imports for the OSM project, especially for areas such as the US where large gorvermental data are freely available [Zielstra et al., 2013].

Chapter 5

OSM comminity

5.0.1 Communication

To communicate with users located all over the world, speaking different languages, is challenging. In OpenStreetMap users work together to create a high-resolution, global representation of the world. All the work done lies in a database which makes the communication and collaboration more challenging. During the Haiti earthquake a big problem was overlapping data. Three independent users could add the same road, with different tags, creating chaos on the map. Especially this event

Generally in Open Source projects the communication between producers

5.0.2 Motivation

Bibliography

[Bishop, 2007] Bishop, D. (2007). [OSM-talk] TIGER, which states next?

[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*.

[Didriksen, 2014] Didriksen, S. (2014). Re: [NUUG kart] Møsvatn.

[edits OpenStreetMap,] edits OpenStreetMap, A. Automated edits - OpenStreetMap Wiki.

[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.

[Exel van, 2010] Exel van, M. (2010). Data Imports In OpenStreetMap Love Em Or Loathe Em? — oegeo.

[Hansen, 2007a] Hansen, D. (2007a). [OSM-talk] TIGER update (need more suggestions).

[Hansen, 2007b] Hansen, D. (2007b). [OSM-talk] TIGER upload queue (maps of entire US).

[Hansen, 2007c] Hansen, D. (2007c). [OSM-talk] TIGER upload queue (maps of entire US).

[Kihle, 2014] Kihle, K. (2014). [NUUG kart] Bruk av datasett fra Kartverket i OpenStreetMap.

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

[Mehus, 2014] Mehus, T. (2014). [Imports] N50 imports from Kartverket (The Norwegian Mapping Authority).

[Mielczarek, 2007] Mielczarek, T. (2007). [OSM-talk] [OSM-dev] TIGER, which states next?

[Munro, 2007] Munro, R. J. (2007). [OSM-talk] TIGER upload status since 0.5 api.

[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.

[OpenStreetMap, a] OpenStreetMap, C. Automated Edits code of conduct - OpenStreetMap Wiki.

[OpenStreetMap, b] OpenStreetMap, D. Data working group - OpenStreetMap Wiki.

[OpenStreetMap, c] OpenStreetMap, I. Import - OpenStreetMap Wiki.

[OpenStreetMap, 2014] OpenStreetMap, N. (2014). Import/Catalogue/N50 import (Norway) - OpenStreetMap Wiki.

[OpenStreetMap, 2016] OpenStreetMap, S. (2016). Stats - OpenStreetMap Wiki.

[OpenStreetMap, 2007] OpenStreetMap, T. (2007). TIGER - 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.

[Willis, 2007] Willis, N. (2007). OpenStreetMap project imports US government maps — Linux.com — The source for Linux information.

[Willis, 2008] Willis, N. (2008). OpenStreetMap project completes import of United States TIGER data — Linux.com — The source for Linux information.

[Zielstra et al., 2013] 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. *Transactions in GIS*, 17(3):315–334.