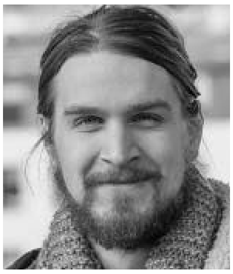
GEOSPATIAL ANARCHY: MANAGING DATASETS THE OPEN SOURCE WAY



Atle Frenvik Sveen
NTNU
atle.f.sveen@ntnu.no

OpenStreetMap (OSM) is the largest and best-known example of geospatial data creation using Volunteered Geographic Information (VGI). A large group of non-specialists joins their efforts online to create an open, worldwide map of the world. The project differs from traditional management of geospatial data on several accounts: both the underlying technology (Open Source components) and the mindset (schema-less structures using tags and changesets).
We review how traditional organizations are currently using the OSM technology to meet their needs and how the mindset of OSM could be employed to traditional management of spatial datasets as well.

Keywords: Volunteered Geographic Information, OpenStreetMap, data management

INTRODUKTION

Earlier, the world of geospatial data was a small place; navigating it did not even require a map. Created by a select few professionals, used by a handful of professionals, and occasionally shared with the general population as printed maps.

Obviously, this is a naive and generalized metaphor, but it serves a purpose in contrasting the trend of Neogeography (Turner 2006). "Neogeography is about people using and creating their own maps, on their own terms and by combining elements of an existing toolset". Closely related to Neogeography is the concept of Volunteered Geographic Information (VGI). VGI describes the phenomenon where private citizens, enabled by the Internet, handheld GPS devices and the graphics capabilities of modern computers, are able to create and share geographic information (Goodchild 2007).

The VGI-movement has made way for several projects aiming to share the generated data with the wider public. The canonical example of successful VGI is undoubtedly OpenStreetMap (OSM). Founded in 2004 by Steve Coast (M. Haklay and Weber 2008) at the University College London, the goal of this online geospatial database is to gather and share geospatial data of the entire world, for everyone to use (Neis and Zipf 2012).

Although some argue that "[...] volunteered and non-specialist data are more affected by inaccuracies and contain less scientific value" (Criscuolo et al. 2016), while others (Mordechai Haklay 2010) compared OSM-data with Ordnance Survey data and found that "OSM information can be fairly accurate", and notes the "impressive update speed" and variations in completeness.

CHARACTERISTICS OF OPENSTREETMAP

Tagging and versioning

In addition to changing the players in the geospatial game, OpenStreetMap arguably also changed the playing field. The traditional ways of creating and organizing geospatial data were clearly challenged. OpenStreetMap presented a fully versioned database (Poore and Wolf 2013) with its core schema-less approach implemented using loosely defined "tags" and geometries represented with lines and nodes. Tags are the OSM counterpart of attributes defined in strict schemas in the traditional relational database mindset, which has been the theoretically correct way of designing database models for decades (Poore and Wolf 2013).

In OSM one or more tags are attached to a geometry to indicate their meaning and functional role. The OSM Wiki specifies that tagging should deliberately be informal, loose and open. The use of existing tags are encouraged, but there are no limitation on the creation of new tags (Ballatore, Bertolotto, and Wilson 2013). Studies have shown that, for a sample of OSM-data, tag values from a controlled vocabulary are extensively used (> 98%) , although correct use of the tags cannot be assured (Mooney and Corcoran 2012).

Who are the contributors and how do they edit?

OSM is open for anyone with a registered account to update and edit. Interestingly, one can observe that the OSM community acts more like a community of close-knit groups, each working on their home country and coordinating their efforts through mailing lists, chat rooms, and Wikis. This way of organizing volunteer-work online closely resembles the Bazaar-model of Open Source software (Raymond 2001), as noted by Haklay et al. (2013).

In addition to local groups, there are some more specialized efforts to help adding data to OSM, most notably the Humanitarian OSM Team (HOT). HOT started as an informal group of OSM volunteers in the wake of the Haiti earthquake in 2010. The individuals joined forces to map the affected areas in OSM to support the aid effort. Today HOT is a registered non-profit organization with full-time staff, working on improving OSM in disaster-affected areas throughout the world (Soden and Palen 2014). HOT attracts attention and commercial support from a multitude of enterprises and organizations. Their success in humanitarian aid is recognized by leading organizations such as the American Red Cross, the Bill & Melinda Gates Foundation, and the World Bank, which all engage in collaborations with HOT (HOT 2016).

Users can add geographical features by tracing aerial and satellite photos (access to imagery is provided by companies such as Microsoft and Mapbox), tracing uploaded GPS tracks, or by editing existing features by adding or altering the tags to add information such as names, types of features etc.

Another method of adding data to OSM is the (sometimes automated) import of existing data with permissive licenses. This data can be governmental datasets released under open licenses, or other open databases of geospatial data. The Netherlands, India, France, parts of Italy, Japan, and parts of Canada are examples of countries where data from other datasets have been added to OSM (Gröchenig, Brunauer, and Rehrl 2014).

Following the release of Norwegian spatial data in 2013, large parts of the national map datasets have been added to OSM. The OSM wiki maintains a list of all known sources for large-scale import of external data (OSM Wiki 2016a).

The OpenStreetMap infrastructure

On the technical side, OpenStreetMap represents an infrastructure of a centralized database in concert with related software-components, most of them available under an Open Source License. The components can be divided in three major parts (OSM Wiki 2016b):

1. Data editing software.
2. Data storage, import and export APIs.
3. Map rendering software

One possible fourth component may be the various visualization tools, but these may also be considered to be a value added resource as a result of the OSM ecosystem and not tightly connected to the core of the OSM initiative. Thus, we abandon styling and cartography from the discussion in this paper.

Users can edit and add data to OSM through several editors; most prominent are the iD web-editor (and the earlier Potlatch editor) and the Java-based JOSM desktop editor. All the editors submit data to the underlying, central, PostgreSQL database through an API.

Data exports from the OSM database are done through the Osmosis library, which produces diff-files (or diffs), files that describes the changes to the underlying database. These diffs can then be fed to other libraries such as osm2pgsql to populate spatial databases such as PostGIS enabling others to replicate the complete database and follow its changesets efficiently.

The third class of OSM components are the map rendering software (renderers), with Mapnik being the best known and used component. These renderers transforms the vector geometries to raster-maps served as map-tiles (see Batty et al. 2010 for an overview of map-tiles) using stylesheets.

Another way of accessing the OSM data is through specialized services for searching and data extraction, such as Nominatim and the Overpass API.

In general, the OSM software stack is considered well documented, easily configurable and backed by a large pool of contributors of both code and technical assistance (Wolf et al. 2011).

APPLYING CONTROLLED ANARCHY IN GOVERNMENTAL INSTITUTIONS

This article has so far shed light on some of the key characteristics of OpenStreetMap, both in terms of user mindset and technical solutions. We argue that both the mindset and technological solutions should be more strongly considered in more traditional data management tasks. Governmental institutions, municipalities and other organizations tasked with gathering and maintaining geospatial datasets should consider implementing the successful concepts we observe from the OpenStreetMap initiative.

The main issues raised from more traditionally geared organizations are the lack of a formal schema, the dilution of the expert role, and to some degree difficult acceptance of new technology.

There is little literature on the topic of implementing “the OSM way” in traditional data management. However, there are examples of organizations using OSM to cover their mapping needs. One such example is the Norwegian University of Science and Technology (NTNU), which in 2009 launched an effort to map the university campus in OpenStreetMap, with the aim of using OSM as the source of the official campus maps. After a competition and encouraged volunteer-effort; 250 changesets of the campus area was registered (Andersen 2009). Although this example shows that organizations can and do use OSM, it is worth noting that NTNU is not a traditional producer of geospatial datasets.

Another, perhaps more relevant example, is the U.S. Geological Survey (USGS). In their work on The National Map, a “collaborative effort among the USGS and other Federal, State, and local partners to improve and deliver topographic information for the Nation", they investigated the feasibility of using the OSM Software stack to facilitate cross-agency co-editing of spatial data. Their experiences from phase one of this work is reported by Wolf et al. (2011).

Their main motivation for adopting the OSM stack was in part to investigate how to let users contribute data, and in part to investigate how to improve collaborative data editing. Using OSM directly was considered, but not pursued due to data lisencing issues. With the software stack beeing Open Source this was their chosen approach. The project reports that, apart from some specific technical issues, the web interface was efficient and easy to use, conflict resolution and versioning works well and the system supports "thousands of simultaneous edit sessions".

On the negative side, the project reported the need for technical staff with an understanding of the main building blocks in the OSM Stack (Linux, Ruby on Rails, and PostgreSQL). Another issue was the OSM approach to quality control; the focus is on implicit quality and the notion that given enough users errors will be corrected (a version of Linus' Law formulated by Raymond 2001). This contrasts the traditional notion of tracking quantitative measures such as accurracy and correctness.

In general, this example shows that the OSM Technology is mature for use in more traditional settings, but it did not explore the application of the OSM mindset. Focus was on how to support a pre-determined schema, i.e. discarding the notion of tags as used by OSM. While the authors note that "this type of convention may not always be possible with all potential partners and volunteers", the project did not shed more light on this topic.

Another governmental organization currently using the OSM stack is the US National Park Service (NPS). NPS has built Places, their "internal data collection system for [..] "core" geospatial data" (National Park Service 2016) on the OSM stack. During an interview with a member of the team (McAndrew 2016) it became evident that the main resons for adopting the stack was the configurability of the components, in addition to being mature and tested in real-world applications. The team also spent a lot of time mapping their existing data to relevant OSM tags, in order to be able to tap into the work done by the OSM community. Nevertheless, some additional tags had to be introduced. Another interesting observation is that some users resisted the idea of letting "anyone" edit "their" data. There is indeed a balance to be struck between a quick feedback loop and correctness.

DISCUSSION

There is no question that VGI in general, and OSM in particular, are more than fleeting trends, they represent shifts in the creation, editing and consumption of geospatial data. This should imply that governmental organizations should examine the way they create and manage geospatial datasets, and assess whether they can improve their internal processes by learning from initiatives like OSM.

In this assessment, there are at least three key aspects that should be considered:

1. The first aspect is the technological platforms and solutions. This relates both to the use of Open Source software, to new concepts for storage and data manipulation, as well as to focus on usability for non-experts. This aspect is arguably the most mature, as there at least some examples of real-life use of the OSM stack, as exemplified by USGS and NPS. In addition, Open Source software for Geospatial (FOSS4G) is proven to be mature (Moreno-Sanchez 2012) and governmental institutions seems to be adopting FOSS4G at an increasing rate.
2. The second aspect is the use, and inclusion of data from VGI initiatives in more "formal" settings. This poses some challenges, but have the potential, if executed correctly, to greatly enhance existing datasets and procedures for managing them (Elwood, Goodchild, and Sui 2012).

3. The third aspect seems to be open for further research, as little work has so far been carried out. The OSM mindset of schema-less datasets and tags as opposed to schemas (i.e. a bottom-up approach) differs drastically from the current workflow in many organizations. This approach undoubtedly raises some issues itself, but without further research and real-world experiments, it is hard to tell. A compelling analogy might be the Open Source workflow (the Bazaar approach described by Raymond 2001), which can be observed influencing software development in traditional software development teams. Some advantages may include; less time spent up-front defining schemas, meaning new datasets can be created and spread faster. Another compelling advantage is that such a system is more capable when it comes to dealing with change; there is no need to revise the schema when a new concept is needed. Possible drawbacks are problems related to programmatically access and usage of data without a defined schema (as described by Atzeni, Bugiotti, and Rossi 2014), as well as concerns about the role of the expert and the reliability of the data.

CONCLUSION

Although the OSM technology stack and the concept of VGI has shown its value both through real-life implementations and in the scientific literature there are still open questions regarding adoption of the OSM mindset in more “formal” settings. We aim at investigating these research questions more in depth, by carrying out small-scale, real world, implementations and investigate if this mindset has any advantages, and if so, identifying what they are and how they can be utilized.

We are interested in cooperation with organizations willing to participate in such experiments and who are open to challenging the way they handle their geospatial data.

REFERENCER

- Andersen, Rune M. 2009. *"Resultater Fra Kartkonkurransen."* Ntnukart. https://ntnukart.wordpress.com/2009/11/18/resultater-fra-kartkonkurransen/.
- Atzeni, Paolo, Francesca Bugiotti, and Luca Rossi. 2014. *"Uniform Access to NoSQL Systems."* Information Systems 43: 117–33. doi:10.1016/j.is.2013.05.002.
- Ballatore, Andrea, Michela Bertolotto, and David C. Wilson. 2013. *"Geographic Knowledge Extraction and Semantic Similarity in OpenStreetMap."* Knowledge and Information Systems 37 (1): 61–81. doi:10.1007/s10115-012-0571-0.
- Batty, Michael, Andrew Hudson-Smith, Richard Milton, and Andrew Crooks. 2010. *"Map Mashups, Web 2.0 and the GIS Revolution."* Annals of GIS 16 (1): 1–13. doi:10.1080/19475681003700831.
- Criscuolo, Laura, Paola Carrara, Gloria Bordogna, Monica Pepe, Francesco Zucca, Roberto Seppi, Alessandro Oggioni, and Anna Rampini. 2016. *"Handling Quality in Crowdsourced Geographic Information."* European Handbook of Crowdsourced Geographic Information, 57–74. doi:http://dx.doi.org/10.5334/bax.e.
- Elwood, Sarah, Michael F. Goodchild, and Daniel Z. Sui. 2012. *"Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice."* Annals of the Association of American Geographers 102 (3). Taylor & Francis Group: 571–90. doi:10.1080/00045608.2011.595657.
- Goodchild, Michael F. 2007. *"Citizens as Sensors: The World of Volunteered Geography."* GeoJournal 69 (4): 211–21. doi:10.1007/s10708-007-9111-y.
- Gröchenig, Simon, Richard Brunauer, and Karl Rehrl. 2014. *"Digging into the History of VGI Data-Sets: Results from a Worldwide Study on OpenStreetMap Mapping Activity."* Journal of Location Based Services 8 (3). Taylor & Francis: 198–210. doi:10.1080/17489725.2014.978403.
- Haklay, M., and P. Weber. 2008. *"OpenStreetMap: User-Generated Street Maps."* IEEE Pervasive Computing 7 (4): 12–18. doi:10.1109/MPRV.2008.80.
- Haklay, Mordechai. 2010. *"How Good Is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey Datasets."* Environment and Planning B: Planning and Design 37 (4): 682–703. doi:10.1068/b35097.
- Haklay, Mordechai, Sofia Basiouka, Vyron Antoniou, and Aamer Ather. 2013. *"How Many Volunteers Does It Take to Map an Area Well? The Validity of Linus' Law to Volunteered Geographic Information."* http://dx.doi.org/10.1179/000870410X12911304958827. Taylor & Francis. doi:10.1179/000870410X12911304958827.
- HOT. 2016. *"HOT Partnerships."* Accessed September 29. https://hotosm.org/partnerships.
- McAndrew, James. 2016. *"Interview."* Personal communication.
- Mooney, Peter, and Padraig Corcoran. 2012. *"The Annotation Process in OpenStreetMap."* Transactions in GIS 16 (4): 561–79. doi:10.1111/j.1467-9671.2012.01306.x.
- Moreno-Sanchez, Rafael. 2012. *"Free and Open Source Software for Geospatial Applications* (FOSS4G): *A Mature Alternative in the Geospatial Technologies Arena."* Transactions in GIS 16 (2). Blackwell Publishing Ltd: 81–88. doi:10.1111/j.1467-9671.2012.01314.x.
- National Park Service. 2016. *"Places."* National Park Service Website. Accessed September 28. https://www.nps.gov/maps/tools/places/.
- Neis, Pascal, and Alexander Zipf. 2012. *"Analyzing the Contributor Activity of a Volunteered Geographic Information Project — The Case of OpenStreetMap."* ISPRS International Journal of Geo-Information 1 (3): 146–65. doi:10.3390/ijgi1020146.
- OSM Wiki. 2016a. *"Import/Catalogue."* http://wiki.openstreetmap.org/w/index.php?title=Import/Catalogue&oldid=1344831.
- 2016b. *"OpenStreetMap Component Overview."* http://wiki.openstreetmap.org/w/index.php?title=Component_overview&oldid=1315501.
- Poore, Barbara S., and Eric B. Wolf. 2013. *"Metadata Squared: Enhancing Its Usability for Volunteered Geographic Information and the GeoWeb."* In Crowdsourcing Geographic Knowledge, 43–64. Dordrecht: Springer Netherlands. doi:10.1007/978-94-007-4587-2_4.
- Raymond, Eric Steven. 2001. *"The Cathedral and the Bazaar."* In The Cathedral and the Bazaar: Musings on Linux and Open Source by an Accidental Revolutionary, 19–65. O'Reilly. https://people.eecs.berkeley.edu/~kubitron/cs162/hand-outs/cathedral-bazaar.pdf.
- Soden, Robert, and Leysia Palen. 2014. *"From Crowdsourced Mapping to Community Mapping: The Post-Earthquake Work of OpenStreetMap Haiti."* In COOP 2014 - Proceedings of the 11th International Conference on the Design of Cooperative Systems, 27-30 May 2014, Nice (France), 311–26. Cham: Springer International Publishing. doi:10.1007/978-3-319-06498-7_19.
- Turner, Andrew J. 2006. *Introduction to Neogeography.* OReilly Short Cuts.
- Wolf, Eric B, Greg D Matthews, Kevin Mcninch, and Barbara S Poore. 2011. *"OpenStreetMap Collaborative Prototype, Phase 1."* U.S. Geological Survey Open-File Report 2011-1136, 23.