Geospatial Data Management in the Future

Combining data management techniques from Open Source communities and governmental bodies to enable cross-data-exploration and sense making.

Ph.D. Project Description – Atle Frenvik Sveen (atle.f.sveen@ntnu.no)

1. Background

Geospatial Data has been created and managed since the first maps where made (Garfield, 2013). The impact of the digital revolution on this field have far-ranging consequences. A map is but one of several representations of the underlying digital data. The digitalization of the map-making process thus involves several shifts. One is the de-coupling of the printed map from the actual data, another is the fact that geospatial data can be used for more than printing maps.

Open Data is another consequence of digitalization. There is an increasing political pressure to make digital data produced and maintained by governments available to the public (Cox & Alemanno, 2003; Ginsberg, 2011; Yang & Kankanhalli, 2013). Political accountability, business opportunities, and a more general trend towards openness are all cited as reasons behind this movement (Huijboom & Broek, 2011; Janssen, Charalabidis, & Zuiderwijk, 2012; Sieber & Johnson, 2015). In practice this means that geospatial data from a range of sources are becoming available for everyone to use for whatever purpose they see fit.

A third trend is crowdsourcing, or Volunteered Geographic Information (VGI) (M. F. Goodchild, 2007). This concept bears some resemblance to Free and Open Source Software. The underlying concept is that amateurs collaborate on tasks such as writing online encyclopedias, writing computer software, or, as in the case of geospatial data, create a database of map data covering the world: OpenStreetMap (OSM) (M. Haklay & Weber, 2008).

How do these trends shape the landscape of Geospatial Data Management? The divide between a printed map and the underlying data makes a case for the use of Geographic Information Systems (GIS) to perform analyses and edit the data to infer new information and to influence decision-making processes (M. Goodchild, 2003). While these tools have been available for decades, the Open Data movement means that data is now available to a large group of users, and that data from various sources can be combined and managed in the same system, possibly far removed from the organization that produced the data.

Storage of geospatial data is a topic that is well covered in literature, with specific focus on spatial databases, of both the relational and NoSQL types (Güting, 1994; Scholz, 2011; Shekhar et al., 1999). Geospatial data distribution is also a topic that is well covered through the focus on Spatial Data Infrastructures (SDIs) and the Open Geospatial Consortium (OGC) standards [SITAT]. The quality, reliability, and usages of VGI are also topics that have received a great deal of scientific interest (Elwood, Goodchild, & Sui, 2012; Mordechai Haklay, 2010).

What is lacking is a combined overview and a set of best practices. What characterizes a system built to handle an automated gathering of geospatial data published in a myriad of formats, with different

metadata standards (or no metadata at all), with different update frequencies, and different licenses? A thorough investigation of these problems will enable a better understanding of what data is of interest, how it should be shared, and how the promised value of Open Geospatial Data can be extracted.

2. Objectives

The overall objectives of this project are (1) to establish guidelines on how to store and manage geospatial data from disparate sources, with different structure and quality, and (2) to explore how this data can be utilized for value generation and decision support. The overarching theme of both objectives are how the Open Source mindset can be utilized.

The following lists the most important tasks identified for the fulfillment of the overall objectives.

1. Investigate the current Geospatial Data Ecosystem, to understand the history and state-of-the-art of geospatial data storage and management, as well as establishing an overview of existing and emerging data sources.
   - *Through literature surveys and active participation in the research community, as well as investigation of current practices in leading geospatial companies.*

2. Develop new, automated, methods for storing and managing geospatial data and benchmarking of both existing and new methods.
   - *Prototype implementation and benchmarking.*
   - *Establish guidelines and best-practices*

3. Find new methods for value generation and analysis of geospatial datasets suited for decision support.
   - *This will be carried out in cooperation with the Data Warehouse Division at Norkart to ensure a strong link to the industry.*

4. Disseminate the research through relevant channels.
   - *International peer-reviewed journals*
   - *Presentation and contribution to conferences*
   - *Popular science channels where appropriate*

5. Establish collaboration with other relevant researchers and research groups as well as users and contributors to FOSS4G and OSM.
   - *Participation at conferences, gatherings, and online communities*

3. Scope

The focus of the project is to investigate how an Open Source mindset can be utilized for geospatial data storage and management, and further for utilization of the data. This enables us to use existing libraries and components and thus benchmark solutions that are readily usable in the industry. These priorities also establish a natural line of progression of the work. Initially a literature survey on Open Geospatial

Data and related areas will be conducted. This builds a foundation for setting up experiments to evaluate storage and gathering strategies. In turn, this offers a building block for further studies on how to utilize the data in new and better ways.

Open Data is concerned just as much with interpersonal, organizational, juridical, and political issues as they are with technology. While these aspects are both important and offer a range of interesting research questions in themselves, our main focus is the technical challenges this topic presents. While storage and management of both data in general and geospatial data in particular is viewed by many as an exercise in data structures and algorithms, we focus on a higher level of abstraction. That is, we will focus on how existing data structures and algorithms can be applied in new ways.

4. Research method

As stated, the main objectives of this project are to establish guidelines and best-practice descriptions supported by working implementations. All these activities are concerned with computer programs in some way or another. Several different research methods can, and probably should, be applied to investigate the performance of computer systems. Most computer programs are deterministic, and thus lends themselves to quantitative research methods. Algorithms, the basic foundation of computer programs, can in many cases be proved correct using mathematical induction, while more complex algorithms and programs may depend on so many external variables that statistical analysis is necessary to reason about the results. The data that serves as input to statistical analysis may be performance metrics measured in terms of execution time, throughput, processing power used, or even the power needed to perform the computation.

In addition to measurements and reasoning based on these quantitative measurements, the fields also require a certain degree of qualitative methods. This aspect becomes evident when we include the human users of such programs in the equation. The construction of user-interfaces, ranging from graphical user interfaces, via command-line interfaces, to configuration files and application programming interfaces, has to take the user into account. These issues are referred to as Usability issues, and can be inspected and evaluated using a range of qualitative methods, ranging from formal inspections, via empirical methods, to informal inspections based on heuristics (Nielsen, 1994).

To describe the research methods employed in the project we use the proposed work on “Efficient Storage Strategies for Heterogeneous Geospatial Data” as a case. We start out with the hypothesis that there exists some optimal way of constructing a computer program to deal with the efficient storage of geospatial data from disparate sources. In order to test this, we plan to first survey literature and the industry for existing solutions, describe their features, and compile a set of metrics. Based on this we will implement a novel solution, along with an implementation that resembles the current state-of-the-art. The next step is to decide a set of metrics to measure the efficiency of these systems. These metrics may include ways to measure write- and read-speeds, error-resilience, and storage requirements. Then the two solutions are benchmarked using a standardized set of input data. After obtaining and analyzing these measurements we are in a position to determine if our novel solution is any better than the current state-of-the-art. If not, we need to revise our implementation and re-run the benchmarks. If our implementation is performing better according to our metrics we may bring it to the next stage, where we assess usability issues by conducting user interviews, usability inspections and other related

investigations. A final stage would be to try out our implementation in an industry-setting, as a replacement for existing programs.

5. Ethical issues

While all research that is disseminated openly may be used by anyone, for any purpose (much like Free Software or Open Data) there is no need to encourage malicious use of the research conducted in this project. Geospatial data has a long history of being used for military purposes (Hewitt, 2011), which someone may deem ethically questionable. Other malicious activities, such as human trafficking, rainforest destruction, illegal use of motorized vehicles in national parks, disturbance of wildlife, poaching, drug smuggling, human trafficking, burglaries, stalking, and surveillance may also benefit from improved use of geospatial data. However, we are not limiting our work because of this, as almost all technological advancements may be used for malicious purposes.

A more concrete issue is the combination of data from separate sources. This may lead to inferred data, which may be problematic for privacy concerns (Wu, Zhu, Wu, & Ding, 2014). Techniques such as strict access control and anonymization of data are usual measures against this, but these techniques are not watertight (Cormode & Srivastava, 2009; Machanavajjhala & Reiter, 2012). The safest approach is to use data that does not contain personal information at all, and this is the plan for this project. If datasets that contains personal information are deemed necessary to use, the above-mentioned precautions will be taken. This applies specifically to cadastral data, which may be relevant to use in the project. If such data is used there is existing Norwegian legislation regarding the use of cadastral data, which will be followed.

6. Expected results

There are two main results we hope to obtain from this project. The fist is a better understanding of how geospatial data can be gathered from disparate sources and stored in an efficient manner that can be utilized. The other main result is to find new areas, products, and methods that be carried out by using this data. Establishing systems for assessing quality and fitness for use of the data is also an important aspect.

These findings will both be of a theoretical nature; i.e. disseminated through scientific, peer-reviewed papers and conferences as well as actual implementations of the methods. The implementations will be carried out in cooperation with the Data Warehouse Division of Norkart AS. This ensures that the real-life implementations will serve as points-of-reference, and that new approaches will be adapted to a real-life scenario.

One of the goals of the Open Data movement is economic benefits, and several countries have conducted studies that estimate the economic benefit of opening data in general or geospatial data in particular. In Norway, the estimated benefit of Open Geospatial Data is in the range 32 – 174 MNOK (3 – 18 million €) (Vennemo, Ibenholt, Magnussen, Moen, & Riis, 2014). Although there are a lot of uncertainty in these estimates (Koski & Tutkimuslaitos, 2015) there is both a political and commercial interest in leveraging Open Geospatial Data. As such, this project is directly relevant for the geospatial industry.

The close cooperation with Norkart AS means that the findings of the project will be made available to an innovative company that has both the resources, customers, and vision to leverage the findings and carry them from the initial, investigative, phase to a product or service. This is also one of the goals of the Industrial Ph.D. scheme.

7. Work plan

There is much wisdom in the words of Winston Churchill; “Plans are of little importance, but planning is essential”. While one could and should try to plan for a project with a span of four years there is only one certainty: the plan will change. With that in mind the following outlines the tasks that will be performed, publications that will be authored, and conferences that will be attended in order to complete the project.

Due to the organization of the project as an Industrial Ph.D. scheme there is a requirement to spend 25% of the time working for Norkart AS. The allocation of this time will be flexible and should not interfere with important milestones and deadlines in the Ph.D. project.

Semester	**Activity**
Fall 2016	- Courses worth 10 credits - Start literature survey - IFEL8000 completed
Spring 2017	- 1 Paper (The Open Geospatial Data Ecosystem) submitted - Courses worth 15 credits - Project plan finished
Fall 2017	- Conference presentation: FOSS4G Boston (OpenStreetMap + Micro-tasking) - Courses worth 7.5 credits - Start implementation + benchmarking on Paper (Efficient Storage Strategies for Heterogeneous Geospatial Data)
Spring 2018	- 1 Paper (Efficient Storage Strategies for Heterogeneous Geospatial Data) submitted - 1 Conference presentation
Fall 2018	- Continue work on Storage strategies - 1 Paper (On OpenStreetMap/Micro-tasking) submitted - 1 Paper (On Data Usage/BigData/BI) submitted
Spring 2019	- 1 Paper (Efficient Storage Strategies for Heterogeneous Geospatial Data, part II) submitted - 2 conference presentations
Spring 2020	- Completion, summary, and write-up - 1 level 2 paper submitted
Fall 2020	- Ph.D. Defense

Date: 18.04.2017

Sign: _____ Sign: _____

Prof. Terje Midtbø
Thesis Supervisor

Atle Frenvik Sveen
Ph.D. Candidate

References

Cormode, G., & Srivastava, D. (2009). Anonymized data. *Proceedings of the 35th SIGMOD International Conference on Management of Data - SIGMOD '09*, 1015. https://doi.org/10.1145/1559845.1559968

Cox, P., & Alemanno, G. (2003). Directive 2003/98/EC of the european parliament and of the council of 17 november 2003 on the re-use of public sector information. *Official Journal of the European Union*, *46*, 1–156.

Elwood, S., Goodchild, M. F., & Sui, D. Z. (2012). Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice. *Annals of the Association of American Geographers*, *102*(3), 571–590. https://doi.org/10.1080/00045608.2011.595657

Garfield, S. (2013). *On The Map: Why the world looks the way it does*. Profile Books.

Ginsberg, W. R. (2011). *Obama Administation's Open Government Initiative: Issues for Congress*. Congressional Research Service. Retrieved from https://fas.org/sgp/crs/secrecy/R41361.pdf

Goodchild, M. (2003). Geographic Information System (GIS). In *Encyclopedia of Computer Science* (pp. 748–750). Chichester, UK: John Wiley and Sons Ltd. Retrieved from http://dl.acm.org/citation.cfm?id=1074100.1074424

Goodchild, M. F. (2007). Citizens as sensors: The world of volunteered geography. *GeoJournal*, *69*(4), 211–221. https://doi.org/10.1007/s10708-007-9111-y

Güting, R. H. (1994). An introduction to spatial database systems. *The VLDB Journal*, *3*(4), 357–399. https://doi.org/10.1007/BF01231602

Haklay, M. (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. https://doi.org/10.1068/b35097

Haklay, M., & Weber, P. (2008). OpenStreetMap: User-Generated Street Maps. *IEEE Pervasive Computing*, *7*(4), 12–18. https://doi.org/10.1109/MPRV.2008.80

Hewitt, R. (2011). *Map of a nation: A biography of the Ordnance Survey*. Granta Books.

Huijboom, N., & Broek, T. Van Den. (2011). Open data: an international comparison of strategies. *European Journal of ePractice*, *12*(March/ April 2011), 1–13. https://doi.org/1988-625X

Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, Adoption Barriers and Myths of Open Data and Open Government. *Information Systems Management*, *29*(4), 258–268. https://doi.org/10.1080/10580530.2012.716740

Koski, H., & Tutkimuslaitos, E. (2015). *The Impact of open data – a preliminary study*.

Machanavajjhala, A., & Reiter, J. P. (2012). Big privacy: protecting confidentiality in big data. *XRDS: Crossroads, The ACM Magazine for Students*, *19*(1), 20–23. https://doi.org/10.1145/2331042.2331051

Nielsen, J. (1994). Usability inspection methods. *Conference Companion on Human Factors in Computing Systems - CHI '94*, *25*(1), 413–414. https://doi.org/10.1145/259963.260531

Scholz, J. (2011). Coping with Dynamic, Unstructured Data Sets–NoSQL a Buzzword or a Savior? In *Proceedings REAL CORP* (pp. 121–129).

Shekhar, S., Chawla, S., Ravada, S., Fetterer, A., Liu, X., & Lu, C. T. (1999). Spatial databases accomplishments and research needs. *IEEE Transactions on Knowledge and Data Engineering*, *11*(1), 45–55. https://doi.org/10.1109/69.755614

Sieber, R. E., & Johnson, P. A. (2015). Civic open data at a crossroads: Dominant models and current challenges. *Government Information Quarterly*, *32*(3), 308–315. https://doi.org/10.1016/j.giq.2015.05.003

Vennemo, H., Ibenholt, K., Magnussen, K., Moen, E., & Riis, C. (2014). *Verdien av gratis kart- og eiendomsdata*. Retrieved from https://www.regjeringen.no/globalassets/upload/kmd/plan/verdien-av-gratis-kart-og-eiendomsdata.pdf

Wu, X., Zhu, X., Wu, G. Q., & Ding, W. (2014). Data mining with big data. *IEEE Transactions on Knowledge and Data Engineering*, *26*(1), 97–107. https://doi.org/10.1109/TKDE.2013.109

Yang, Z., & Kankanhalli, A. (2013). Innovation in Government Services : The Case of Open Data. *IFIP International Federation for Information Processing 2013*, 644–651. https://doi.org/10.1007/978-3-642-38862-0_47