Knowledge-based geospatial data integration and visualization with Semantic Web technologies

Weiming Huang

GIS Centre, Lund University, Lund, Sweden
weiming.huang@nateko.lu.se

Abstract. Geospatial information is indispensable for various spatially-informed analysis and decision-making, e.g. traffic analysis and built environment processes. Geospatial data often must be integrated for meaningful analysis, whereas such integration is challenging due to siloed data organization, semantic heterogeneity and multiple representation of geospatial data. Moreover, the visualization of geospatial data is one of the most prominent ways of utilizing geospatial data, however how to properly visualize the data is sometime difficult, as it pertains to a wide range of visualization (cartographic) knowledge. Semantic Web technologies unveil a promising way to mitigate these issues, as they provide means of data integration on the Web, and knowledge representation capacity to formally represent the visualization knowledge. In this PhD project, we investigate the potential values of Semantic Web technologies for geospatial data integration (particularly for geospatial data with multiple representation) and visualization in several cases, where the integration and visualization knowledge is formalized using Semantic Web technologies. All the case studies embody real-world meaning and entail data integration and visualization challenge, which have been addressed by state-of-the-art solutions inadequately. Preliminary results demonstrate great yet not fully unlocked potential of Semantic Web technologies for geospatial data, and also disclose challenges that need to be addressed.

Keywords: geospatial data, data integration, data visualization, Semantic Web, ontologies, semantic rules, SHACL, web maps, spatial analysis

1 Problem statement

Over the last decades, the massive use of geospatial information in various real-world applications (e.g. traffic analysis and built environment processes) gradually reveals the indispensable role of geospatial information for spatially-informed analysis and decision making [1]. At the meantime, geospatial information is one of the most powerful information integrators to bridge diverse sources of information [2]. Such natures of geospatial information entail the need of geospatial data integration and geospatial knowledge outreach.

Geospatial data integration includes the integration of multi-source geospatial data and the integration between geospatial and other types of data that can be grounded

geographically. Currently, geospatial data is maintained and delivered mainly by spatial data infrastructures (SDIs). However, the data in SDIs is inadequately integrated and harmonized, particularly the integration of geospatial data and other types of data is rare. Geospatial data integration is complex, and a prominent intricacy is, among others, the multiple (geometric) representations of geospatial data, which is a specific data integration problem in the geospatial domain [3]. The multiple representations delineate the geographic space with several abstraction levels (e.g. a building can be represented as a point or a polygon), and thereby enable visualization and analysis at different scales. However, this often arises difficulties when incorporating geospatial data for spatially analysis.

Moreover, the knowledge concerning how to appropriately use geospatial data is important. There have been many endeavors for geospatial knowledge formalization, whereas today experts from other domains still often have to look into the literature, or cooperate with geospatial experts to accomplish meaningful use of geospatial data. Visualization, as one of the most predominating ways of utilizing geospatial data, also entails much semantic intricacies, as visualizing geospatial data in a sensemaking and cartographically satisfactory way pertains to a wide range of cartographic knowledge, which is hard to transfer, interpret, and reuse, especially by non-geospatial experts.

The increasing appreciation of Semantic Web technologies in the geospatial domain [4] unveils a promising means to unravel the above discussed difficulties of geospatial data integration and visualization. Semantic Web technologies provision mechanisms for integrating and interlinking geospatial data on the Web in a distributed manner; they allow for lifting semantic harmonization level with formally defined ontologies; and the knowledge representation capacity of this technology stack provides a promising way to represent and share geospatial (visualization) knowledge on the Web to foster wider use of such knowledge and spatially enable the Web [5]. However, current outcomes of Semantic Web for geospatial data are insufficient in terms of, among others, handling multiple representation (as the concepts used for data with different representations are the same, but the data should not be applied in the same means [3]), and formalizing and representing geospatial knowledge on the Web.

Therefore, this PhD project investigates the potential of Semantic Web technologies for geospatial data integration (particularly the handling of multiple representations), and formalizing geospatial (visualization) knowledge for knowledge outreach on the Web.

2 Relevancy

Geospatial information plays an indispensable role in a vast number of spatial analysis and spatially-informed decision making, thus sharing, integrating geospatial data on the Web are important, so is the outreach of geospatial knowledge.

From another perspective, Semantic Web technologies (especially the parts concerning linked data and ontologies) are increasingly adopted and applied in the geospatial domain. A recent survey conducted in 2018 by EuroSDR (European Spatial Data Re-

search) demonstrated that linked data has been seen as one of the most important research issues and major movers toward future SDI [6]. Linked data was also voted as one of the most important SDI research topics during the AGILE (Association of Geographic Information Laboratories in Europe) 2018 workshop 'SDI research and strategies towards 2030'[1]. In this context, it is relevant to investigate the potential benefits of employing Semantic Web technologies for delivering geospatial data and knowledge. This is in line with several international and national initiatives, e.g. the INSPIRE (infrastructure for spatial data in Europe) investigation on geospatial linked data, and Swedish national study on linked geodata [7].

3 Related work

The application of Semantic Web technologies has developed considerably in geospatial domain in the last decade, as they address several long-standing challenges of e.g. data integration, semantic interoperability and knowledge formalization, and provide a promising way to connect SDIs with the mainstream IT to augment the application of geospatial data [5]. Consequently, the amount of geospatial data released as linked data is rapidly growing, and some of them are serving as central hubs in the Linked Open Data (LOD) cloud, e.g. GeoNames[2]. Furthermore, a number of geo-ontologies have been designed, the geospatial linked data query language GeoSPARQL has been standardized by Open Geospatial Consortium (OGC) [8], and a number of RDF stores have become spatially-enabled (e.g. Stardog[3] and Virtuoso[4]). These theoretical and technical advancements have created an increasingly mature environment for incorporating geospatial data and knowledge in the Semantic Web.

3.1 Geospatial data integration with Semantic Web technologies

Increasing geospatial data has been published or planned to be delivered as linked data; this trend is particularly prominent for authoritative geospatial data. For instance, Ordnance Survey, the national mapping agency (NMA) in the UK, has released several geospatial datasets maintained by them as linked data [9]. In the Netherlands, Kadaster released several key datasets, e.g. building data, addresses, as linked data on the Web, together with other governmental open data [10]. In Europe, the Joint Research Centre (JRC) of the European Commission investigated the potentials of publishing the INSPIRE-compliant geospatial data as linked data through the ARE3NA activity[5].

Semantic Web and linked data are used for geospatial data integration to (partially) resolve semantic heterogeneity of multi-source data and consolidating distributed information. Such work has been accomplished mostly in the environment of SDIs. For

1 https://kcopendata.eu/sdi2030/
2 https://www.geonames.org/
3 https://www.stardog.com/
4 https://virtuoso.openlinksw.com/
5 https://inspire.ec.europa.eu/news/linking-inspire-data-draft-guidelines-and-pilots

instance, Janowicz et al. [11] proposed a framework for semantically enabling SDIs, in which both geospatial data and activities (discovery, registration, processing and visualization) are semantically annotated. Lutz et al. [12] leveraged ontologies and logical reasoning for overcoming semantic heterogeneity in SDIs to foster better geospatial data exchange and reuse. van den Brink et al. [3] identified that many vocabularies have been defined within domains, whereas other domains are seldom taken into account; thus they proposed a methodology and tools for non-automatic, community-driven ontology matching for data harmonization to facilitate data reuse between datasets in the geospatial domain. Despite these promising results, we still need more advanced techniques to e.g. handle multiple representations of geospatial data for cross-detailed-level integration with subtle semantic relations (as illustrated in Section 7).

3.2 Geospatial knowledge representation using Semantic Web technologies

The capacity of knowledge representation of Semantic Web leveraging ontologies and rules has been recognized in the geospatial domain for many years and used in a number of studies. These studies span several research subjects of e.g. visualization, geo-processing and information retrieval. For instance, Hofer et al. [13] developed a knowledge base to support the composition of geo-processing workflow, in which ontologies were used to formalize the geo-operators, and SWRL rules were used for formulating the rules associated with the geo-operators chaining. Keßler et al. [14] employed ontologies and SWRL rules for context-aware geographic information retrieval, where they used ontologies for organizing the semantically annotated data and rules for deriving inference for context detecting. Gould and Mackaness [15] formalized the knowledge for on-demand map generalization using ontologies to facilitate the knowledge to be shared, expanded and reused in mapping systems.

With regard to the visualization of geospatial data, Scheider and Huisjes [16] distinguished extensive and intensive properties using machine learning techniques and formalized different types of properties using ontologies to help map making, as the cartographic rules applied to the two types of properties are fundamentally different. Carral et al. [17] designed an ontology for cartographic map scaling at the dataset level. Varanka and Usery [17] proposed to semantically represent map features using Semantic Web technologies to form the knowledge base of maps. Grounded upon this idea, we believe more knowledge concerning how the raw data is converted to visualizations (visualization knowledge) can be formalized and shared on the Semantic Web. This is also in line with the OGC investigation on semantic data portrayal with the ambition of creating a web of knowledge for data portrayal [19].

4 Research question

The overall research question is *what are the benefits of Semantic Web technologies for geospatial data integration and visualization?*

Under this hood, we formulate several specific research questions focusing on real-world problems that can potentially better addressed by Semantic Web technologies:

1) Geospatial data is often repetitively generated despite relations between the objects. Is it possible to link geospatial objects to existing objects in the Semantic Web to diminish data repetition and inconsistency?
2) The knowledge concerning how to visualize geospatial data is important. Is it possible to use Semantic Web technologies to formalize such knowledge, and thus share it on the Web?
3) Multiple representation of geospatial data sometimes renders data integration complex and problematic. Is it possible to leverage Semantic Web technologies to formalize the knowledge of multiple representation and assist cross-detailed-level data integration?
4) Geospatial data interlinking is imperative to further unlock the potential of the Semantic Web. Therefore, it is relevant and important to investigate how to advance geospatial data interlinking.

5 Hypotheses

In order to answer the research questions, we formulated a set of hypotheses that are used to operationalize the work, including:

1) Using linked data and ontologies can facilitate the multi-source geospatial data integration on the Web, especially instead of repetitively generating multi-source geospatial data, one could link the data to reference data (e.g. authoritative geospatial data from NMAs) to obtain more accurate location information for better visualization and analysis.
2) Coupling ontologies and semantic rules can (partially) formalize the geospatial data visualization knowledge into knowledge bases, thereby enable semantic reasoning to derive proper visualization means for the data, which can make the visualizations appropriate tools for decision making.
3) Combining ontologies and semantic constraints (SHACL) can represent complex and subtle semantic relations raised by multiple representation of geospatial data, and thus facilitate the use of geospatial data in other domains that perceive the geographic space differently.
4) For automating geospatial data integration at instance level, the knowledge graph embedding technique is useful, but geometric (location) information is also important. Thus, combining geometric information with knowledge graph embedding technique can help geospatial data integration and interlinking.

6 Approach and evaluation plan

As geospatial information can be used in various real-world applications, the value of the data integration and visualization can be revealed in solving real-world or even long-standing problems that have not been solved before or have been inadequately

solved. Therefore, the approach for addressing the research questions and testing the hypotheses is mainly case studies, i.e. spatially-informed analysis or decision making. The evaluations are/will be mainly performed by comparing Semantic Web-based solutions to traditional solutions.

Specially, to test hypothesis 1, we use the case of web maps for natural reserved areas, as this type of geospatial objects often have intrinsic connections with other geospatial objects, whereas state-of-the-art approach of data modelling neglects such relations. Linked data can be used to relatively position the natural reserved areas to reference objects (e.g. roads, rivers, cadastres).

To test hypothesis 2, we use the case of heritage building protection mapping, where we use ontologies and SPIN (SPARQL) rules to formalize the visualization knowledge and distributed linked data retrieval. The rationale of using SPIN rules rather than e.g. SWRL rules is that it is often that the visualization rules include non-monotonic semantics, e.g. a rule stating that render the object in a certain way if the value of an attribute does not exist (closed world assumption). Also, SPIN rules have a formalized vocabulary and can be more readily shared on the Web.

To test hypothesis 3, we use a case study of evaluating urban infrastructure's suitability for bicycling, where we utilise SHACL constraints to represent subtle and complex semantic relations raised by multiple representations between the geospatial domain and the traffic domain. With SHCAL constraints, the knowledge concerning using which level of representations for which scenarios can be explicated and formally represented. Such formalized knowledge can guide cross-detailed-level data integration and also facilitate wide-use of such knowledge.

To test hypothesis 5, we plan to utilise the geospatial data available in the LOD cloud, and also we will generate geospatial data with multiple representations from authoritative geospatial datasets, and compare the method with state-of-the-art methods for linked data interlinking, and object matching methods in the geospatial domain.

7 Results

To validate hypothesis 1, we developed a relative positioning approach based on linked data and ontologies. That is, instead of absolutely positioning all the geospatial features repetitively, we relatively position geospatial (thematic) features (objects) to background data (i.e. geospatial data with multiple representations from Swedish mapping agency). Ontologies were designed for storing the relative positioning information, linked data was used to link the relatively positioned features to reference features. This approach accomplished self-adapting web maps for better visualization performance, which had seldom been addresses by other methods [20].

To validate hypothesis 2, we designed a knowledge base encapsulating ontologies and semantic rules (SPIN rules) to represent the knowledge concerning cartographic scale, data portrayal, and geometry source. The approach accomplished visualizing distributed and multi-scale geospatial data in a cartographically satisfactory way, which can be hardly implemented using current OGC technology stack [21].

To validate hypothesis 3, we employ the study case of evaluating the suitability of urban road network infrastructure for bicycling, in which geospatial data needs to be integrated with field-collected data. Such integration is complex, as the traffic researchers (who develop indexes for the suitability evaluation) perceive the geographic space differently than the modelling of geospatial data. The indexes treat the junctions in the road network as a whole (using a single point feature to represent a junction), this corresponds to the data modelling approach of the less detailed geospatial road network. However, the indexes need the dedicated bicycling paths information, which is only available in the most detailed geospatial road network (in the most detailed road network, the junctions are modelled with detailed structure, mainly including polylines and points). Such cross-detailed-level and cross-domain data integration cannot be implemented merely using ontologies, thus we impose SHACL constraints for this type of data integration. The constraints ensure the semantic correctness of utilising data from different detailed levels.

In addition, we investigate some popular and well-known RDF stores, i.e. RDF4J, Jena, Stardog, Virtuoso and GraphDB for their geospatial query capacity, particularly focusing on GeoSPARQL-compliance and query performance. This is important as it gives insights concerning where to deploy the proposed approach. The assessment and benchmarking are conducted in two scenarios. In the first scenario, geospatial data comprises a part of a large scale data infrastructure and is integrated with other types of data. In the second scenario, we benchmark the RDF stores in a dedicated SDI environment with purely geospatial data. The results show that GeoSPARQL-compliance has considerably developed with reasonable query efficiency, while query correctness still remains a challenge, as different stores sometimes return different results for the same query.

8 Reflections

To date, we have collected positive evidence showing that Semantic Web technologies have great and yet not fully unlocked potential benefits for geospatial data integration and visualization. The supervision team of this PhD project is mainly from the geospatial domain, thus we are familiar with the real need and challenges that are potentially could be better solved with Semantic Web technologies, and we have tight connections with the authorities that are interested in employing Semantic Web technologies for geospatial data, e.g. Swedish mapping agency, Swedish Geological Survey, Swedish Traffic Administration, etc. This project is also conducted closely cooperating with experts from other domains that are in need of geospatial data and knowledge, e.g. traffic researchers. Furthermore, we have close collaboration with Semantic Web researchers with theoretical and technical advice. In summary, this project has good connections and background knowledge to investigate the research questions. Also, the values revealed from the case studies embody real-world usefulness of Semantic Web technologies, and this will potentially draw more extensive attention from various domains.

Despite the promising results, there are still several challenges, e.g. the geospatial data interlinking on the Web still remains a challenging and sometimes expensive task, while it is imperative for unlocking the values of Semantic Web for geospatial data. We plan to address this issue in the next step. This work will benefit from both the advancements in the Semantic Web (e.g. knowledge graph embedding technique), and the outcomes of geospatial feature matching that has been studied for decades.

Acknowledgements

This PhD project is under the supervision of Prof. Lars Harrie and Dr. Ali Mansourian at Lund University.

References

1. Kuhn, W.: Core concepts of spatial information for transdisciplinary research. International Journal of Geographical Information Science 26, 2267-2276 (2012)
2. Janowicz, K., Scheider, S., Pehle, T., Hart, G.: Geospatial semantics and linked spatiotemporal data–Past, present, and future. Semantic Web 3, 321-332 (2012)
3. van den Brink, L., Janssen, P., Quak, W., Stoter, J.: Towards a high level of semantic harmonisation in the geospatial domain. Computers, Environment and Urban Systems 62, 233-242 (2017)
4. Wiemann, S., Bernard, L.: Spatial data fusion in spatial data infrastructures using linked data. International Journal of Geographical Information Science 30, 613-636 (2016)
5. Schade, S., Smits, P.: Why linked data should not lead to next generation SDI. In: Geoscience and Remote Sensing Symposium (IGARSS) 2012 IEEE International, pp. 2894-2897. IEEE (2012)
6. EuroSDR. EuroSDR Annual Report 2018. Available online: http://www.eurosdr.net/sites/default/files/images/inline/eurosdr_annual_report_2018.pdf (accessed June 12, 2019)
7. Blomqvist E., & Östman, A. Länkade Geodata - Omvärldsanalys. Report in the project Länkade Geodata. (2014)
8. Perry, M., and John Herring.: OGC GeoSPARQL-A geographic query language for RDF data. Open Geospatial Consortium technical report (2012)
9. Goodwin, J., Dolbear, C., Hart, G.: Geographical linked data: The administrative geography of great britain on the semantic web. Transactions in GIS 12, 19-30 (2008)
10. Folmer, E., Beek, W., Rietveld, L.: Linked Data Viewing as part of the Spatial Data Platform of the Future. International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 42, 49-52 (2018)
11. Janowicz, K., Schade, S., Bröring, A., Keßler, C., Maué, P., Stasch, C.: Semantic enablement for spatial data infrastructures. Transactions in GIS 14, 111-129 (2010)
12. Lutz, M., Sprado, J., Klien, E., Schubert, C., Christ, I.: Overcoming semantic heterogeneity in spatial data infrastructures. Computers & Geosciences 35, 739-752 (2009)
13. Hofer, B., Mäs, S., Brauner, J., Bernard, L.: Towards a knowledge base to support geoprocessing workflow development. International Journal of Geographical Information Science 31, 694-716 (2016)

14. Keßler, C., Raubal, M., Wosniok, C.: Semantic rules for context-aware geographical information retrieval. In: European Conference on Smart Sensing and Context, pp. 77-92. Springer (2009)
15. Gould, N., Mackaness, W.: From taxonomies to ontologies: formalizing generalization knowledge for on-demand mapping. Cartography and Geographic Information Science 1-15 (2015)
16. Scheider, S., Huisjes, M.D.: Distinguishing extensive and intensive properties for meaningful geocomputation and mapping. International Journal of Geographical Information Science 33, 28-54 (2019)
17. Carral, D., Scheider, S., Janowicz, K., Vardeman, C., Krisnadhi, A.A., Hitzler, P.: An ontology design pattern for cartographic map scaling. In: Extended Semantic Web Conference, pp. 76-93. Springer (2013)
18. Varanka, D.E., Usery, E.L.: The map as knowledge base. International Journal of Cartography 4, 201-223 (2018)
19. Fellah, S.: Testbed-12 Semantic Portrayal, Registry and Mediation Engineering Report. Open Geospatial Consortium technical report (2017)
20. Huang, W., Mansourian, A., Abdolmajidi, E., Xu, H., Harrie, L.: Synchronising geometric representations for map mashups using relative positioning and Linked Data. International Journal of Geographical Information Science 32, 1117-1137 (2018)
21. Huang, W., Harrie, L.: Towards knowledge-based geovisualisation using Semantic Web technologies: a knowledge representation approach coupling ontologies and rules. International Journal of Digital Earth, Advance online publication (2019)