Integrating Big Spatio-Temporal Data Using Collaborative Semantic Data Management

Matthias Frank(✉)

Information Process Engineering, FZI Forschungszentrum Informatik,
Haid-und-Neu-Str. 10-14, 76131 Karlsruhe, Germany
frank@fzi.de

Abstract. Good decision support of geographical information systems depends on the accuracy, consistency and completeness of the provided data. This work introduces the hypothesis that the increasing amount of geographic data will significantly improve the decision support of geographical information systems, providing that a smart data integration approach considers provenance, schema and format of the gathered data accordingly. Sources for spatial data are distributed and quality of the data is varying, especially when considering uncertain data like volunteered geographic information and participatory sensing data. In our approach, we address the challenge of integrating Big Data in geographical information systems by describing sources and data transformation services for spatio-temporal data using a collaborative system for managing meta data based on Semantic MediaWiki. These machine interpretable descriptions are used to compose workflows of data sources and data transformation services adopted to the requirements of geographical information systems.

1 Introduction and Motivation

Geographical information systems (GISs) are important tools for decision support in various fields like civil planning, emergency management, agriculture or environment and nature protection. While the amount of data with a geographic context is increasing due to improved sensor technology and data created by mobile devices or users of social web applications, the reliability of these data may be uncertain and has to be evaluated when used in GIS. In addition, data values have different schemas to describe locations, like addresses, relative spatial relationships or different coordinates reference systems. Measured quantities and units that might be used for data values may also vary across data sources, leading to schema heterogeneity. In this approach, we use semantic Web technology to describe *(i) data sources and data transformation services* for GIS in a machine interpretable way. These semantic descriptions lay the foundation for *(ii) composing workflows* of data sources and data transformation services that fulfill different requirements of GISs on demand, even if these requirements

Ph.D. supervisor: Prof. Dr. York Sure-Vetter.

A. Bozzon et al. (Eds.): ICWE 2016, LNCS 9671, pp. 507–512, 2016.
DOI: 10.1007/978-3-319-38791-8_38

are not known at design time. In addition, provenance information and uncertainty of data are modeled using semantic Web technologies, which allows to *(iii) include volunteered geographic information (VGI)* and participatory sensing data. We hypothesise that the increasing amount of geographic data will significantly improve the decision support of GIS, providing that a smart data integration approach considers provenance, schema and format of the gathered data accordingly. This leads to the following research questions:

RQ1 How does the integration of uncertain data like VGI and participatory sensing data improve the decision support in GIS in terms of accuracy, consistency and completeness?
RQ2 Are domain experts provided with the necessary provenance information in order to compare the processed values of heterogeneous data sources within the result set?

In Sect. 2, we discuss related work and state our aims and objectives in Sect. 3. We introduce our research methodology in Sect. 4 and conclude in Sect. 5.

2 Related Work

In this section we discuss related work on the topics of *(i) data transformation and interoperability* of GIS, *(ii) semantic workflow composition* and *(iii) integration of VGI* and participatory sensing data.

2.1 Data Transformation and Interoperability of GIS

Transforming data from heterogeneous data sources into a unified schema and the interoperability of distributed systems is still an ongoing research topic where web services are commonly used for converting data. Kaempgen et al. [5] presented OLAP4LD, a framework for developers of applications over Linked Data sources reusing the Resource Description Framework (RDF) Data Cube Vocabulary. The Quantities, Units, Dimensions and Data Types Ontologies (QUDT)[1] can be used as a common standard for describing units and their conversation. In the context of GIS, transformation of spatial data across different coordinates reference systems was addressed by Atemezing et al. [1] which have published a dataset dedicated to the description of coordinates reference systems defined and maintained by the French national mapping agency. Similar requirements are also given for the Gauss-Krueger coordinates reference system used by national agencies in Germany. Li et al. [10] reported on their efforts to design and develop a geospatial cyberinfrastructure for urban economic analysis and simulation using a service-oriented architecture to allow widespread sharing and seamless integration of distributed geospatial data. For the interoperability of spatial data observed by sensors, the World Wide Web Consortium (W3C) Semantic Sensor Network Incubator Group introduced the Semantic Sensor Network (SSN) ontology[2] for describing sensors and observations.

[1] http://www.qudt.org/.
[2] http://purl.oclc.org/NET/ssnx/ssn.

2.2 Semantic Workflow Composition

Another challenge is to compose the workflow of data sources and transformation services that fulfills the requirements of any GIS. Gil et al. [4] have formalized an approach how the selection of application components and data sources can be automated in general using semantic Web technologies. Kopeck et al. [6] presented research in lightweight machine-readable service descriptions and semantic annotations for Web application programming interfaces (APIs), building on the Hypertext Markup Language (HTML) documentation that accompanies the APIs. Lanthaler [7] described an approach to build hypermediadriven Web APIs based on Linked Data technologies and developed Hydra [8], a small vocabulary to describe Web APIs. Lanthaler and Guetl [9] also introduced an approach to create machine-readable descriptions for RESTful services and show how these descriptions along with an algorithm to translate SPARQL Protocol and RDF Query Language (SPARQL) queries to Hypertext Transfer Protocol (HTTP) requests can be used to integrate RESTful services into a global read-write Web of Data. Calbimonte et al. [2] annotated sensor data and observations using an ontology network based on the SSN ontology and showed how to provide a highly flexible and scalable system for managing the life-cycle of sensor data in the context of the semantic Web of Things. Gemmeke et al. [3] have shown that semantic technologies can help to cope with data format heterogeneity, distribution of the data sets and interoperability issues in the medical domain, for example when processing medical images. Similar challenges have to be addressed in the domain of GISs when processing raster data created by satellites, drones or surveillance cameras.

2.3 Integration of Volunteered Geographic Information

In our approach, we also want to integrate uncertain data like VGI, Linked Open Data (LOD) and participatory sensing data to support decision with GISs. Lopez-Pellicer et al. [11] proposed a refinement of Linked Data practices, named Geo Linked Data, which defines a lightweight semantic infrastructure to relate URIs that identify real world entities with geospatial Web resources, such as maps. Stadtler et al. [12] elaborated on how the collaboratively collected OpenStreetMap[3] data can be interactively transformed and represented adhering to the RDF data model. They described how this data is interlinked with other spatial data sets, how it can be made accessible for machines according to the Linked Data paradigm and for humans by means of several applications, including a faceted geo-browser. The spatial data, vocabularies, interlinks and some of the applications are openly available in the LinkedGeoData[4] project.

3 Aims and Objectives

When integrating geographic data from different sources with different units of measurement, property definitions or coordinates reference systems, the data

[3] http://www.openstreetmap.org.
[4] http://linkedgeodata.org.

have to be transformed into a homogenous schema in order to receive a unified view of all sources. This requires not only a structure but also explicit and well defined semantics for meta data that represents the relations of the data that should be integrated as well as a mechanism for automated preprocessing and reliability quantification of uncertain data like VGI and participatory sensing data. By describing heterogeneous data sources for GIS and the input and output of available data transformation services semantically, we are able to implement services that dynamically compose workflows for processing these data and fulfill the requirements of different GISs. Data and index stuctures for the extraction and aggreation of relevant structures and the accessibility for analytics have to be covered by the composed workflows. For a complete representation of data, missing values have to be predicted by suitable algorithms. This approach enables domain experts to select any combination of data sources and receive a complete result set in different data schemas without the need of considering the format and completeness of the original sources. On the other hand, all provenance information has to be retained in order to make the values of different sources as well as predicted values comparable. Using semantic Web technology for managing geo temporal data does also enable semantic analytics for unstructured and remote sensing data. In our work we investigate a semantic workflow composition for integrating Big Data in GIS.

4 Research Methodology

In our approach, we design an infrastructure that gathers Big Data which may be needed to support decisions with GISs. The infrastructure itself has to be generic in order to fulfill the requirements of different GISs that may consume the gathered data for further processing. For the evaluation, this infrastructure is used for different use cases in order to prove the flexibility and answer the research questions stated in Sect. 1. To build up our infrastructure, we gather data from the regional environment authorities of Baden-Württemberg[5], the German weather service[6], mobile measurements on a urban railway[7] and remote sensing data from satellites operated by European Space Agency and National Aeronautics and Space Administration. More data sources should be integrated in a later stage of the project. Data of areas of interests which are not covered by the data gathered already will be collected by drones equipped with the necessary sensors where needed. The first step is to build a collaborative system based on Semantic MediaWiki (SMW) for managing meta data. This system does import and reuse commonly used vocabulary in the domain of GISs like SSN, QUDT and GeoVocab[8] and is used to describe data sources and transformation services. This information is used to dynamically build workflows consisting of the data sources and transformation services needed to fulfill the requirements defined

[5] https://www.lubw.baden-wuerttemberg.de/lubw.
[6] http://www.dwd.de.
[7] http://www.aero-tram.kit.edu/.
[8] http://geovocab.org/.

by the consuming GIS. We assume that these requirements are not know at design-time, therefore our infrastructure has to cover them on demand. For a first demonstration, we have used open refine[9] with the rdf plugin[10] in order to *(i) transform temperature data of weather stations* manually. We have used the SPARQL Inferencing Notation[11] to define unit conversation of thermodynamic temperatures using the information of QUDT. The prepared sources and services are then registered in SMW using suitable SMW-templates. The data of our SMW is stored in an Apache Jena[12] triple store which provides a Fuseki SPARQL endpoint. This endpoint provides the API for machine interpretable descriptions which we intent to use for *(ii) dynamically composed workflows of data sources and data transformation services* later in our work. The *(iii) integration of VGI and participatory sensing data* is not yet realized in this stage of our work.

5 Conclusions and Contribution to Web Engineering

In order to evaluate our contribution, we are going to investigate the output created by the dynamically created workflows of our infrastructure with regard to completeness, number and quality of integrated sources and the explanatory power of the provenance information within the result set. This will be made for a specific use case, which requires a set of records for thermodynamic temperature values from heterogeneous sources like weather stations, satellite data and thermal sensors of drones. With the gathered data as our training set, we plan to perform predictions for sub urban heat islands within the city of Karlsruhe, Germany, and evaluate the prediction with our test data set which is classified as measurements from sub urban heat islands in the same city. By varying the sources used as input for the predictions, we are going to evaluate the impact of these sources on the decision support in GISs. With our approach, we show how the decision support of a new generation of GISs can be improved by making big geo-temporal data including VGI and participatory sensing data available for analytics. We introduced the principles of this approach with the use case of thermal data gathered from various sources like weather stations, satellite observations and mobile sensors. By describing these data sources and appropriate data transformation services in a SMW we created machine interpretable descriptions. We intent to use this data for dynamically composed workflows of data sources and data transformation services later in our work.

Acknowledgements. This work was supported by the German Ministry of Education and Research (BMBF) within the BigGIS project (Ref. 01IS14012A). I thank Dr. Stefan Zander, Dr. Benedikt Kämpgen and Prof. Dr. Rudi Studer for guidance and insights.

[9] http://openrefine.org/.
[10] http://refine.deri.ie/rdfExport.
[11] http://spinrdf.org/.
[12] https://jena.apache.org/index.html.

References

1. Atemezing, G.A., Abadie, N., Troncy, R., Bucher, B.: Publishing reference geodata on the web: Opportunities and challenges for ign france. TC-SSN 2014 - Terra Cognita - Semantic Sensor Networks (2014)
2. Calbimonte, J.P., Sarni, S., Eberle, J., Aberer, K.: Xgsn: An open-source semantic sensing middleware for the web of things. In: Joint Proceedings of the 6th International Workshop on the Foundations, Technologies and Applications of the Geospatial Web, TC 2014, and 7th International Workshop on Semantic Sensor Networks, SSN 2014. CEUR Workshop Proceedings, vol. 1401, pp. 51–66. CEUR-WS.org (2015)
3. Gemmeke, P., Maleshkova, M., Philipp, P., Götz, M., Weber, C., Kämpgen, B., Nolden, M., Maier-Hein, K., Rettinger, A.: Using linked data and web apis for automating the pre-processing of medical images (2014)
4. Gil, Y., González-Calero, P.A., Kim, J., Moody, J., Ratnakar, V.: A semantic framework for automatic generation of computational workflows using distributed data and component catalogues. J. Exp. Theor. Artif. Intell. **23**(4), 389–467 (2011)
5. Kämpgen, B., Harth, A.: OLAP4LD – a framework for building analysis applications over governmental statistics. In: Presutti, V., Blomqvist, E., Troncy, R., Sack, H., Papadakis, I., Tordai, A. (eds.) ESWC Satellite Events 2014. LNCS, vol. 8798, pp. 389–394. Springer, Heidelberg (2014)
6. Kopecký, J., Vitvar, T., Pedrinaci, C., Maleshkova, M.: Restful services with lightweight machine-readable descriptions and semantic annotations. In: Wilde, E., Pautasso, C. (eds.) REST: From Research to Practice, pp. 473–506. Springer, New York (2011)
7. Lanthaler, M.: Creating 3rd generation web apis with hydra. In: 22nd International World Wide Web Conference, WWW 2013, Rio de Janeiro, Brazil, May 13–17, 2013, Companion Volume, pp. 35–38. International World Wide Web Conferences Steering Committee/ACM (2013)
8. Lanthaler, M., Guetl, C.: Hydra: A vocabulary for hypermedia-driven web apis. In: Proceedings of the WWW2013 Workshop on Linked Data on the Web, Rio de Janeiro, Brazil, 14. CEUR Workshop Proceedings, vol. 996. CEUR-WS.org (2013), May 2013
9. Lanthaler, M., Gütl, C.: Seamless integration of restful services into the web of data. Adv. MM **2012**, 586542: 1–586542: 14 (2012)
10. Li, W., Li, L., Goodchild, M.F., Anselin, L.: A geospatial cyberinfrastructure for urban economic analysis and spatial decision-making. ISPRS Int. J. Geo-Information **2**(2), 413–431 (2013)
11. Lopez-Pellicer, F.J., Silva, M.J., Chaves, M., Javier Zarazaga-Soria, F., Muro-Medrano, P.R.: Geo linked data. In: Bringas, P.G., Hameurlain, A., Quirchmayr, G. (eds.) DEXA 2010, Part I. LNCS, vol. 6261, pp. 495–502. Springer, Heidelberg (2010)
12. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: Linkedgeodata: A core for a web of spatial open data. Semant. Web **3**(4), 333–354 (2012)