Smart Web Services for Big Spatio-Temporal Data in Geographical Information Systems

Matthias Frank, Stefan Zander

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

Abstract. The informative value of analytic processes by geographical information systems depends on the accuracy, consistency and completeness of the gathered data fed into the system. By feeding Big Data into it, such requirements are hard to maintain, as the provenance, veracity, velocity, structural and semantic heterogeneities of the gathered spatio-temporal data have to be addressed. Exploitation and integration of Big Data in such ways is an ongoing challenge. We present fundamentals of a well-defined and collaborative information integration approach based on semantic web technology, established ontologies and linked APIs that specifically emphasizes a spatio-temporal relation and enable a new generation of geographical information systems. We employ the concept of smart web services for dynamically composed workflows in order to cope with the characteristics of Big Data value streams and generate more elaborated data.

1 Introduction

Geographical information systems (GISs) are important tools for decision support based on spatio-temporal data. These tools are used in various fields like civil planning, emergency management, agriculture or environment and nature protection. In this paper, we introduce a novel approach of how GISs can exploit and integrate Big Data based on semantic web technology, established ontologies and linked application programming interfaces (APIs). Due to improved and pervasive sensor technology and data created by mobile devices and users of social Web applications, the amount of spatio-temporal data is increasing. At the same time, the reliability of these data may be uncertain and need to be taken into consideration when used in GIS. In addition, spatio-temporal data from different sources may use different schemas to describe locations, like addresses, relative spatial relationships or different coordinates reference systems. The quantities measured and units used for data values may also vary across heterogeneous and uncontrolled data sources. Due to these developments, GIS are facing challenges in all four dimensions of Big Data:

– **Volume**: The prevalence and omnipresence of sensor technology and ubiquitous data sources inposes challenges regarding data volumes to be integrated.

- **Variety**: Unstructured data are new kinds of data for GIS, which require innovative methods of data interpretation for analyzing, interpolating, predicting and visualizing.
- **Velocity**: In order to permanently integrate acquired sensor data in GIS, the common batch processing of these systems have to be technically and conceptually reorganized in order to enable real-time analysis and activity recommendations.
- **Veracity**: The integration of volunteered geographic information (VGI) and other user-created content as well as integration of remote sensing analyzed image processing data, which may be incomplete prevent the assumption that collected data are complete and correct at any given point on time.

By feeding Big Data into GISs, we have to take these characteristics into consideration with a special focus on the requirements imposed by GISs, including the provenance information of data. In our approach, we use semantic Web technology to *i) describe data sources and data transformation services* for GIS in a machine interpretable way. This enables smart web services to *ii) compose a workflow* that generates the result set of more elaborated data with respect to accuracy, consistency and completeness. The informative value of analytic processes by GISs depends on the data generated by the composed workflow used as input. When integrating heterogeneous sources of spatio-temporal data with different units of measurement, property definitions or coordinates reference systems, the data have to be transformed into a unified result set across all sources. The schema of this result set has to be flexible and defined on demand in order to match the requirements of different use cases. This requires a structure for meta data that represents the relations of the data that should be integrated. By describing heterogeneous data sources for GIS together with input and output parameters of available data transformation services semantically, workflows for processing these data can be composed dynamically in order to fulfill use case specific requirements. The approach presented in this paper enables domain experts to select a combination of data sources and define the data structure needed for a specific use case with respect to the quantities, units, granularity, precision of measurements, period and area under investigation. On the other hand, all provenance information has to be retained in order to make the values of different sources comparable. Using semantic Web technology for managing spatio-temporal data also enables semantic analysis for unstructured and remotely sensed data. We investigate a semantic workflow composition approach for integrating Big Data in GISs and hypothesise that the increasing amount of geographic data will significantly improve the scientific findings of GISs closer to reality. However, the level of improvement strongly depends on a common understanding of concepts across heterogeneous data sources. This leads to the following research questions:

RQ1 Which type of provenance information is relevant in order to make the processed values of heterogeneous data sources comparable within a unified result set?

RQ2 How does a collaborative approach of describing sources and services for spatio-temporal data scale for Big Data processing workflows in GIS?

Our approach to answer these research questions is based on the related work presented in Section 2. We present fundamentals of a well-defined and collaborative information integration approach in Section 3, show a concrete use case in Section 4 and discuss the preliminary results in Section 5.

2 State of the Art

In addition to the related work discussed in [5], we discuss more related work on the topics of *i) data transformation and interoperability* of GIS and *ii) the concept of smart web services* we intend to use to address transformation and interoperability issues in this section.

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. As an example, Stolz and Hepp [12] proposed to integrate currency conversion functionality from open Web APIs into the Linked Open Data (LOD) cloud in a conceptually clean, scalable way. Harth et al. [7] used Karma[1] for a dynamic integration of a reasonable amount of static and dynamic linked data. Cruz et al. [3] have created a semantic framework for Geospatial and temporal data Integration, Visualization, and Analytics. 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. For GIS, the Open Geospatial Consortium (OGC)[3] defines standards for interoperability. One of their initatives is the Sensor Web Enablement (SWE)[4] which supports services for web integration of sensors like the Sensor Observation Service (SOS)[5] which is a web service to query real-time sensor data and sensor data time series. Observations and Measurements (O&M) is the response model used for SOS, for example the Water Model Language (WaterML)[6] for the representation of water observations data. Lefort et al. [10] have introduced an approach of how to combine the SSN ontology with the Resource Description Framework (RDF) Data Cube vocabulary to a meaningful ontology and applied that ontology on the homogenised daily temperature dataset for the monitoring of climate variability and change in Australia. The

[1] `https://usc-isi-i2.github.io/karma/`
[2] `http://purl.oclc.org/NET/ssnx/ssn`
[3] `http://www.opengeospatial.org/`
[4] `http://www.opengeospatial.org/ogc/markets-technologies/swe`
[5] `http://www.opengeospatial.org/standards/sos`
[6] `http://www.opengeospatial.org/standards/waterml`

Quantities, Units, Dimensions and Data Types Ontologies (QUDT)[7] can be used as a common standard for describing units and their conversion.

2.2 Smart Web Services for Interoperability

Vettor et al. [1] have shown that a service oriented architecture can help to solve heterogeneity issues by attaching explicit semantics to data in a company's information system. Lanthaler and Guetl [9] discussed some of the challenges and choices that need to be made when designing RESTful Web APIs and described an alternative, domain-driven approach to design Web APIs. Based on the semantic description of data sources and data transformation services we intent to employ the concept of Smart Web Services introduced by Maleshkova et al. [11]. For processing symbolic data, Kaempgen et al. [8] extended the drill-across operation over data modeled in the RDF Data Cube vocabulary[8] to consider implicit overlaps between datasets in Linked Data, defined convert-cube operation over values from a single dataset and generalised the two operations for arbitrary combinations of multiple datasets with the merge-cubes operation. Dimou et al. [4] introduced an approach that takes advantage of widely-accepted vocabularies, originally used to advertise services or datasets, such as Hydra or dcat, to define how to access Web-based or other data sources. Gil et al. [6] gave an overview of the Organic Data Science framework, an approach for scientific collaboration that opens the science process and exposes information about shared tasks, participants, and other relevant entities based on Semantic MediaWiki (SMW)[9]. Cherfi et al. [2] proposed and discussed the main constituents of an ontology of quality federating all the aspects of information system components quality.

The work presented in this section expresses that exploitation and integration of Big Data in a way that addresses provenance, veracity, velocity, structural and semantic heterogeneities of spatio-temporal data, especially for GIS, is an ongoing challenge. Based on the related work introduced in this section, we present our collaborative information integration approach for spatio-temporal data in GIS in Section 3.

3 Approach

In our approach, we present fundamentals of a well-defined and collaborative information integration of spatio-temporal data used in GIS. We address the different requirements on more elaborated data for data analysis by describing sources of spatio-temporal data and APIs using semantic web technology and established ontologies. We employ the concept of smart web services for dynamically composed workflows in order to cope with the characteristics of Big Data

[7] http://www.qudt.org/
[8] http://www.w3.org/TR/vocab-data-cube/
[9] http://semantic-mediawiki.org/

value streams. For contributing to the flexibility and usability of GIS, we pose the following requirements:

R1 *Collaborative*: Users should be able to add sources of spatio-temporal data and relevant APIs to the GIS.
R2 *Semantical*: All sources of spatio-temporal data and relevant APIs have to be described in a machine interpretable way.
R3 *Efficient*: Value streams of spatio-temporal observations have to be transmitted and processed with the least possible amount of overhead in order to cope with high volume data.

The first step is to build a *collaborative* system based on SMW for managing meta data. This system does import and reuse commonly used vocabulary in the domain of GISs like SSN, QUDT and GeoVocab[10], which does also cover the Basic Geo Vocabulary[11], to *semantically* describe data sources and the Hydra vocabulary[12] for APIs of transformation services. This information is used to dynamically build workflows consisting of the data sources and (smart) transformation services needed to fulfill the data conditions requested by the data consumer. For an *efficient* processing of the spatio-temporal data, only the meta data of these value streams are modeled with the flexibility of semantic web data formats like RDF, while the observed values are transmitted and transformed with the least possible amount of overhead. The scope of our work within the general architecture of a GIS is shown in Figure 1.

Users create description pages of available data sources or APIs in SMW and the system provides a representation of this information in RDF using our specified ontology annotated with common vocabularies. Smart web services can therefore use this information instantly. We provide a RESTful API that enable domain experts to query the quantities, units, granularity, precision of measurements, period and area under investigation for a specific use case. The response of this API call can be the plain observations within the result set (e.g. JSON, CSV, XML), the observations with semantically described meta data (e.g. JSON-LD, RDF/XML) or a hyperlink to a relational data base that holds the values of the result set, depending on the needs of the data consumer. As the RESTful API used for the unified data access is described semantically itself, all parameters and allowed values can be queried by a consuming application which allows for continuous integration of more functionality. A concrete use case of our approach is described in Section 4.

4 Use Case

Analyzing the characteristics of suburban heat islands (SUHIs), that means heat island that occur within cites on hot days, requires a set of records for thermodynamic temperature values from heterogeneous sources for the area and period of

[10] http://geovocab.org/
[11] `https://www.w3.org/2003/01/geo/`
[12] `http://www.hydra-cg.com/spec/latest/core/`

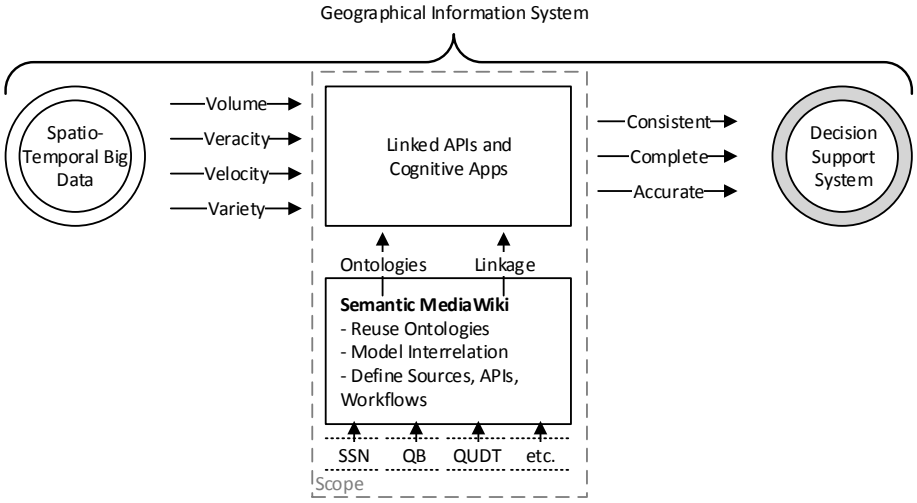

Fig. 1. High-level architecture of the intended GIS

investigation. As a first demonstration of our approach, we gather data from the regional environment authorities of Baden-Württemberg[13], the German weather service[14], mobile measurements on an urban railway[15] and remote sensing data from satellites operated by European Space Agency and National Aeronautics and Space Administration. With the gathered data as our training set, we plan to perform predictions of SUHIs within the city of Karlsruhe, Germany, and evaluate the prediction with our test data set which is classified as measurements from SUHIs 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.

At this stage of our work we have developed the ontology that describes the sources of spatio-temporal data in SMW reusing the vocabularies of SSN and RDF data cube to model the meta data of observations, QUDT for units and quantities and the Basic Geo Vocabulary for the coordinates reference system. Using this collaborative information integration approach, we are able to include the heterogeneous sources of thermodynamic temperature values for our use case and provide meaningful, machine interpretable meta data. We have implemented a RESTful-API that can provide unified result sets for downstream decision support systems by exploiting the semantic descriptions of data sources and data transformation services in SMW.

Our current work focuses on employing the concept of linked APIs and smart web services that use our semantic meta data in order to *dynamically compose*

[13] `https://www.lubw.baden-wuerttemberg.de/lubw`
[14] `http://www.dwd.de`
[15] http://www.aero-tram.kit.edu/

the workflows that cope with the characteristics of Big Data value streams and generate more elaborated data. As an example, the unit conversion of thermodynamic temperatures in our use case may be invoked automatically based on the meta data provided by our SMW, the unit conversion rules defined in QUDT and a rule engine like SPARQL Protocol and RDF Query Language (SPARQL) Inferencing Notation[16] that executes these rules. For the evaluation of the automatically processed data, like the *transformation of temperature data of weather stations*, we have also used open refine[17] with the RDF plugin[18] to manually create test data.

5 Discussion and Conclusion

Based on semantic web technology, established ontologies and linked APIs, we have presented fundamentals of a well-defined information integration approach. Our work supplements the work introduced in Section 2 with respect to a collaborative and continuous integration of data sources and APIs that specifically emphasis a spatio-temporal relation. Exploitation and integration of big spatio-temporal data in a new generation of GIS strongly depend on a common understanding of concepts across heterogeneous data sources. We have shown how to address this issue by combining the dynamics of an collaborative approach with the expressive power of established ontologies. For the evaluation of our approach we are going to observe experimental results in the SUHI use case that show how collaborative created descriptions of spatio-temporal data sources and data transformation services can be used to generate a machine interpretable ontology and employ the concept of smart web services for dynamically composed workflows. We believe that these preliminary results will indicate that our approach enables users even without a web engineering background to easily add sources and services for an existing GIS. For a meaningful evaluation of the research questions defined in Section 1, we have to create more examples of dynamically composed workflows from big spatio-temporal data. The values transformed from different sources to a unified result set have to be investigated together with domain experts with respect to their comparability among each other depending on the provenance, accuracy, consistency and completeness. With an increasing number and amount of data sources and spatio-temporal values that has to be processed, we have to prove that our approach does also scale for Big Data in GIS workflows.

Acknowledgements. This work was supported by the German Ministry of Education and Research (BMBF) within the BigGIS project (Ref. 01IS14012A).

[16] `http://spinrdf.org/`
[17] `http://openrefine.org/`
[18] `http://refine.deri.ie/rdfExport`

References

1. A Service Oriented Architecture for Linked Data Integration (2014)
2. Cherfi, S.S.S., Akoka, J., Comyn-Wattiau, I.: Federating information system quality frameworks using a common ontology. In: International Conference on Information Quality. pp. 160–173. Adelaide, New Zealand (2011)
3. Cruz, I.F., Ganesh, V.R., Caletti, C., Reddy, P.: Giva: a semantic framework for geospatial and temporal data integration, visualization, and analytics. In: 21st International Conference on Advances in Geographic Information Systems (SIGSPATIAL 2013). pp. 534–537. ACM (2013)
4. Dimou, A., Verborgh, R., Vander Sande, M., Mannens, E., Van de Walle, R.: Machine-interpretable dataset and service descriptions for heterogeneous data access and retrieval. In: Proceedings of the 11th International Conference on Semantic Systems (SEMANTICS 2015). pp. 145–152. ACM (2015)
5. Frank, M.: Integrating big spatio-temporal data using collaborative semantic data management. In: 16th International Conference on Web Engineering (ICWE 2016)
6. Gil, Y., Michel, F., Ratnakar, V., Hauder, M.: A semantic, task-centered collaborative framework for science. In: The Semantic Web: ESWC 2015 Satellite Events - ESWC 2015 Satellite Events Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers. Lecture Notes in Computer Science, vol. 9341, pp. 58–61. Springer (2015)
7. Harth, A., Knoblock, C.A., Stadtmüller, S., Studer, R., Szekely, P.A.: On-the-fly integration of static and dynamic sources data. In: Fourth International Workshop on Consuming Linked Data (COLD 2013). CEUR Workshop Proceedings, vol. 1034. CEUR-WS.org (2013)
8. Kämpgen, B., Stadtmüller, S., Harth, A.: Querying the global cube: Integration of multidimensional datasets from the web. In: Knowledge Engineering and Knowledge Management - 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings. Lecture Notes in Computer Science, vol. 8876, pp. 250–265. Springer (2014)
9. Lanthaler, M., Gütl, C.: Model your application domain, not your json structures. In: 22nd International World Wide Web Conference (WWW '13). pp. 1415–1420. International World Wide Web Conferences Steering Committee / ACM (2013)
10. Lefort, L., Bobruk, J., Haller, A., Taylor, K., Woolf, A.: A linked sensor data cube for a 100 year homogenised daily temperature dataset. In: International Workshop on Semantic Sensor Networks (SSN 2012). CEUR Workshop Proceedings, vol. 904, pp. 1–16. CEUR-WS.org (2012)
11. Maleshkova, M., Philipp, P., Sure-Vetter, Y., Studer, R.: Smart web services (smartws) - the future of services on the web. IPSI BgD Transactions on Advanced Research (TAR) 12(1), pp. 15–26 (2016)
12. Stolz, A., Hepp, M.: Currency conversion the linked data way. In: First Workshop on Services and Applications over Linked APIs and Data (ESWC 2013). CEUR Workshop Proceedings, vol. 1056, pp. 44–55. CEUR-WS.org (2013)