The GeoKnow Generator Workbench: An Integration Platform for Geospatial Data

Alejandra Garcia-Rojas[1], Daniel Hladky[1], Matthias Wauer[2], Robert Isele[3], Claus Stadler[4], and Jens Lehmann[4]

[1] Ontos AG, Nidau, Switzerland
[2] R & D, Unister GmbH, Germany
[3] Brox IT-Solutions GmbH, Germany
[4] AKSW Research Group, Leipzig, Germany

Abstract. Many of our information seeking tasks we perform every day have a geospatial dimension. However, limited effort has been done to integrate geospatial data in the web of data, mainly due to the complexity of this spatial dimension. The GeoKnow project aims to provide the necessary tools and methods to easily integrate geospatial data into the web of data. To this end, we present the GeoKnow Generator, a stack of components that covers several steps in the Linked Data Lifecycle. These components have been integrated in the Generator Workbench, and to demonstrate its usability we also present four different real world use cases where the Generator and Workbench are used.

Keywords: Linked Data, Geospatial, Semantic Web

1 Introduction

Geospatial data has high relevance in every day life, and is crucial for companies decision making. For instance, being able to make quick use of demographics and terrain data for a strategic business opening plan; or simply improving search engines to be able to answer questions like: find a good typical restaurant in Switzerland next to the lake of Zurich. These kind of questions can only be answered if geospatial data is combined with other kind of data, e.g. statistics or events.

The Linked Data (LD) Lifecycle[5] illustrates the process of extracting data form different types of sources, linking it with other datasets, enriching it further, assuring its quality as well as exploring and visualising it. The GeoKnow Generator, described in detail in the next section, is an interface to a set of software tools developed for different stages in the LD Lifecycle, to easily integrate geospatial data into the web of data. These components can also be used separately – preconfigured Debian packages are available in the LD-Stack (see above URL). The main contribution of this paper, presented in 3, is the GeoKnow Generator Workbench, which is a user interface that integrates all components and

[5] See `http://stack.linkeddata.org/`.

enables simple access and interaction with them. Section 4, presents different real world use cases that make use of the Generator and the Workbench.

2 Related Work

In a use case where the LD Lifecycle stages require to be implemented, one first has to have the knowledge of the Semantic Web concepts and then the knowledge of the existing tools that can accomplish the execution of the different stages. Using tools separately could be complex task and time consuming. The LOD2 Statistical Workbench [3], provided an integrated set of tools from the LD-Stack for the official statistical production process by government. This workbench support performing different operations such as extraction, transformations, querying, validation of RDF data cubes, interlinking, visualisation and publishing. This solution was suitable for a very specific use case, however more general configurable approach leave room for improvement. Unifiedviews[6] [4], is a Linked Data processing framework created under the EU project COSMODE[7]. Unifiedviews has being deployed also using components from the LD-Stack. This platform requires implementing Data Processing Units for each component in order to be integrated, which results in a tightly coupled architecture. Moreover, Unifiedviews doesn't provide support for authentication or authorisation features, but, it represents a relevant reference point for the GeoKnow Generator Workbench.

3 Generator Workbench

The LD Lifecycle describes the stages one may follow to integrate data in the web of data. The GeoKnow Generator Workbench aims to provide an integration of the different components that are used in the process. We have selected a set of more relevant tools for the use cases, and developed the required interfaces (APIs and GUIs). The components that are currently integrated in the Workbench are described in Table 1. An online demo of the Workbench is available at `http://generator.geoknow.eu`, together with some video tutorials.

The workbench allows users to triplify geospatial data, such as ESRI shapefiles and spatial tables hosted in major DBMSs using the GeoSPARQL, WGS84 or Virtuoso RDF vocabulary for geospatial representation of point features (TripleGeo). Non-geospatial data in RDF (local and online RDF files or SPARQL endpoints) or data from relational databases (via Sparqlify) can also be entered into the Generators triple store. Data from the Generator's triple store can be linked (using LIMES), enriched (using DEER), queried (using Virtuoso), visualised (using Facete2) and used to generate light-weight applications as JavaScript snippets (using Mappify) for specific geospatial applications. Most

[6] `http://www.unifiedviews.eu/`
[7] `http://www.comsode.eu/`

steps in the LD Life-Cycle have been integrated in the Generator as a graph-based workflow, which allows the user to easily manage new generated data. Moreover, the Workbench integrates a processing dashboard that provides feedback to the user about LD processes that he may have started, and allows him to execute them on demand. The actual architecture of the GeoKnow Generator is presented in Figure 1.

Tool	Category
Sparqlify[8]	RDB2RDF Mapping
TripleGeo [6]	Geo-spatial feature extraction
Deer[9]	Enrichment with with geo-spatial information
LIMES [5]	Linking Workbench
FAGI-gis [2]	Fusion of geospatial RDF data and metadata
Mappify[10]	Map view generator
Facete [7]	Faceted browser for spatial data
Virtuoso[11]	Hybrid RDBMS/Graph Column Store

Table 1. Overview on GeoKnow Linked Data Stack components.

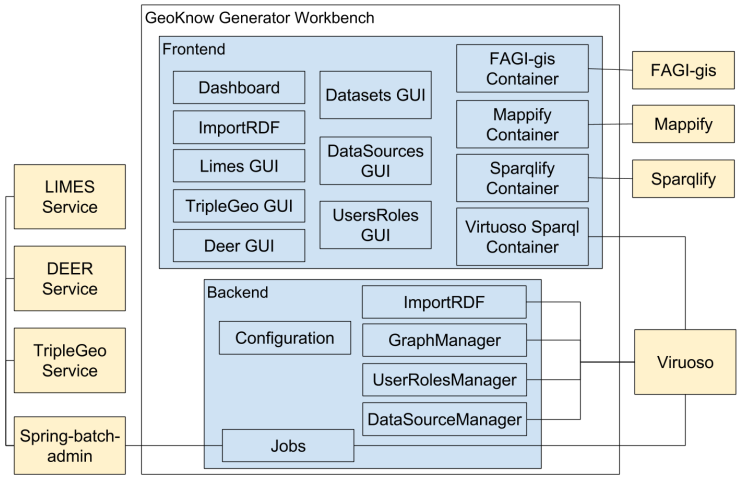

Fig. 1. GeoKnow Generator Workbench Architecture

The integration of such heterogeneous tools in a unified look and feel is a challenging task, especially when it comes to provide collaborative features

[8] http://aksw.org/Projects/Sparqlify.html
[9] http://aksw.org/Projects/DEER.html
[10] http://mappify.aksw.org/
[11] http://virtuoso.openlinksw.com/

and management of private data. The enterprise ready functionalities that the Generator Workbench offers to this end are:

Authentication and Role Management. The workbench provides a proxy-based security model for accessing the workbench and the creation of different roles that will restrict the access to the different components.

Authorisation. A graph-based security access control allows users to create and configure access control to datasets for the public and/or user specific.

Job Monitoring. For some of the software tools, which can have long runtime on large-scale input, the user can execute batch jobs that are configured and observable in a dashboard. This feature was implemented using Spring-batch[12].

Data Provenance. When working with several datasources and different processing stages, it is required to keep information about the provenance of certain triples. Within the workflow processes, the Workbench adds metadata about the tools used to process these data, timestamp, and authors.

4 Use Cases

The usability of the GeoKnow Generator and the Workbench have been tested in different use cases, described next.

4.1 Tourism e-Commerce

The Tourism e-Commerce use case by Unister[13], aims at improving the user experience when searching for a perfect holiday staying, trying to cover several specific user needs. This use case extends the existing internal data with public datasets in order to understand user needs, and translate them to pertinent vacation packages.

Geospatial information supports the ability of the search engine to understand descriptive features the user may request, such as mountains close to a given city, or a restaurant close to the river. The data has to support queries in order the engine is able to understand distances, areas and locations, and perform some logic around them. The GeoKnow Generator has been used to generate an interlinked dataset used in the implementation and evaluation of a motive-based search infrastructure, as depicted in Figure 2.

To integrate external datasets some transformation to internal data to RDF was performed using TripleGeo, and Sparqlify. The linking of external data and internal data was performed using LIMES. Besides integration of structured data, unstructured data such as hotel reviews can be processed using DEER. That way, we can identify related entities and integrate their attributes (such as locations), which can be used for further analysis of places and provides useful information for a search engine [1]. The resulting data was the capability of having related non-existing attributes (such locations) to some entities that could be used to analyse data and provide useful hints to the search engine.

[12] `https://github.com/spring-projects/spring-batch-admin`
[13] `http://www.unister.com/`

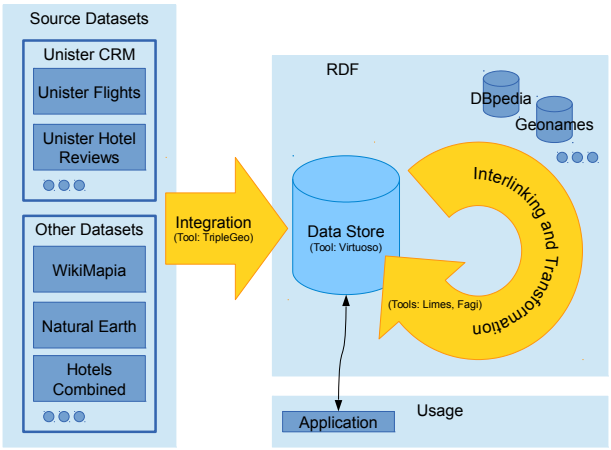

Fig. 2. Overview of the data preparation workflow for the Tourism e-Commerce dataset.

4.2 Supply Chain

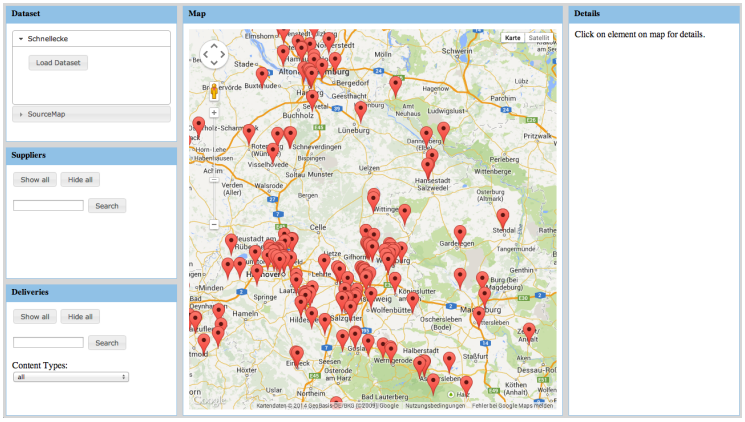

Fig. 3. Current screenshot of the the supply chain dashboard.

The Supply Chain Dashboard (see Figure 3), developed in the scope of the GeoKnow use cases, offers a unified spatial view on the logistics in the supply chain by connecting information from supply chain transactions to the Data Web. As a result, the flow of material and accompanying information can be observed in real-time, bottlenecks can be identified early, media breaks in the information flows are minimised. Companies can benefit from the Supply Chain

Dashboard by gaining a better picture of the current state of the supply chain and the spatial distribution of goods and products in the supply chain.

This scenario incorporates traffic, weather and transport information, and links it to the supply chain information. These data integration and linking were able thanks to the Sparqlify and LIMES components. This information allows a live visualisation of orders and shipments status in the Dashboard. Circulated messages and a supplier score card allows live analytics of the supply chain based on user-defined metrics. Experiences made with the supply chain infrastructure will generate business value for Brox[14] at the late stages of the project. It is certain that some results will flow into Brox's Xybermotive offer[15].

4.3 E-Governmet Services

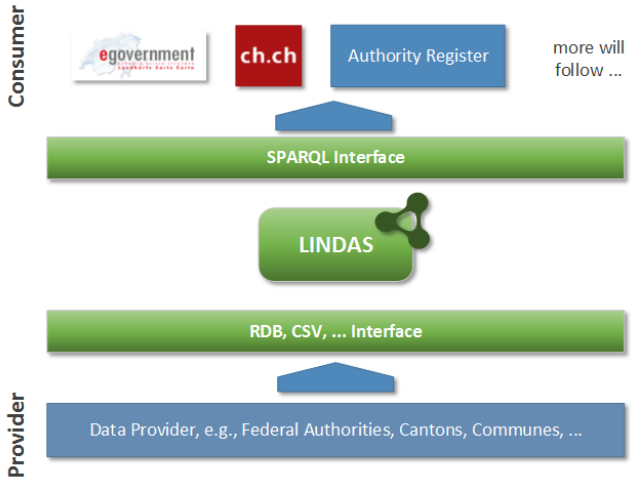

Fig. 4. Data provider LINDAS-plattform data user

Ontos AG[16] has developed the Linked Data Information Workbench (OntosLDIW), a generic, enterprise-ready workbench supporting the LD Lifecycle. The OntosLDIW, used as a foundation the GeoKnow Generator Workbench, it integrates more functionalities such as localisation, and is straightforward to integrate into existing applications. The OntosLDIW was applied to a real world e-government scenario for the State Secretariat for Economic Affairs (SECO) in Switzerland[17]. The Linked Data Service[18] (LINDAS) has as objective to provide

[14] http://brox.de/
[15] http://www.xybermotive.com/
[16] http://ontos.com/
[17] http://www.seco.admin.ch/?lang=en
[18] http://lindas-data.ch/

information about authorities, their services and software solutions are collected decentralised by the Swiss Confederation, the cantons or communes. The service gathers, homogenises, and publishes authority data using Semantic Web standard. The overview of LINDAS system is depicted in Figure 4.

Currently LINDAS integrates essential components from the Generator such RDF Import, CSV2RDF, and it has also integrated other components from the LD Stack, such as D2RQ. Within the next stage of development the governmental data will be linked to the linked open data world using LIMES.

4.4 Continental Automotive GmbH

Continental Automotive GmbH is one of the first commercial users of the authoring and visualisation software Facete, which they employ for data investigation for Continental products DropYa[19] and TruckYa[20]. DropYa is a geosocial network, where users can send and receive location-based messages sharing their experience and recommendations. TruckYa is a community-based tool for finding adequate parking spaces aimed at truck drivers bound by legal requirements regarding their resting periods.

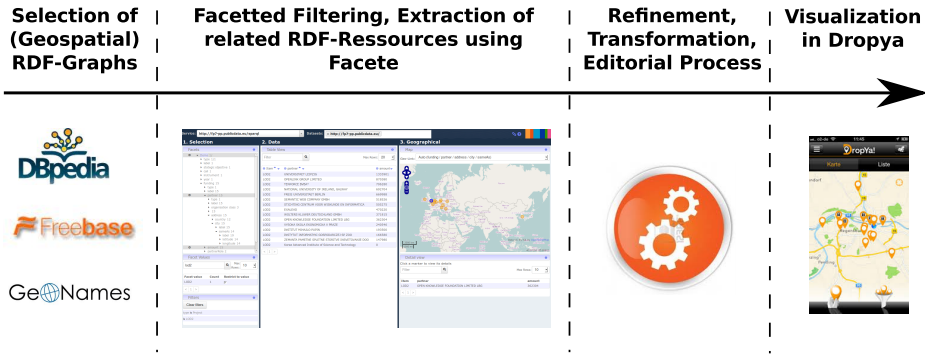

Fig. 5. Facete within the data generation process at Continental

As a basis for these tools, geospatial data is needed, for example motorway service areas including restaurants, showers, hostels. Of future interest are further touristic information such as museums and playgrounds which is readily contained in publicly available data sets like DBpedia, Geonames and Freebase. Continental is currently looking into how this Linked Open Data can be used to extend their existing proprietary data sets. In this investigation process, Facete provides the functionality to browse data on a map, view attributes of interest,

[19] https://dropya.net/

[20] http://www.continental-corporation.com/www/pressportal_com_en/themes/press_releases/3_automotive_group/cvam/press_releases/pr_2013_11_28_vdo_truckya_en.html

export relevant graph-parts and support the editorial process including enrichment and transformation.

5 Conclusions

We described the GeoKnow Generator, which integrates several components to facilitate the integration of geospatial data as Linked Data. A subset of the components of the GeoKnow Generator have been integrated in a Workbench, which is a enterprise ready solution for Linked Data processing. We have presented four real world use cases to demonstrate the usability of Generator components and the Workbench. The GeoKnow Generator Workbench features such as authentication, access control and provenance are essential to the presented use case.

Acknowledgments. This work is part of GeoKnow EU project and has been founded by the European Commission within the FP7 Information and Communication Technologies Work Programme (Grant Agreement No. 318159).

References

1. A. Both, A.-C. N. Ngonga, R. Usbeck, D. Lukovnikov, C. Lemke, and M. Speicher. A service-oriented search framework for full text, geospatial and semantic search. In *SEMANTiCS 2014*, 2014.
2. G. Giannopoulos, D. Skoutas, T. Maroulis, N. Karagiannakis, and S. Athanasiou. Fagi: A framework for fusing geospatial rdf data. In *On the Move to Meaningful Internet Systems: OTM 2014 Conferences*, volume 8841 of *Lecture Notes in Computer Science*, pages 553–561. Springer, 2014.
3. V. Janev, B. V. Nuffelen, V. Mijovi, K. Kremer, M. Martin, U. Miloševi, and S. Vrane. Supporting the linked data publication process with the lod2 statistical workbench. *Semantic Web Interoperability, Usability, Applicability*, 2014.
4. T. Knap, M. Kukhar, B. Machác, P. Skoda, J. Tomes, and J. Vojt. Unifiedviews: An ETL framework for sustainable RDF data processing. In *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, pages 379–383, 2014.
5. A.-C. Ngonga Ngomo and S. Auer. Limes - a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of IJCAI*, 2011.
6. K. Patroumpas, M. Alexakis, G. Giannopoulos, and S. Athanasiou. Triplegeo: an etl tool for transforming geospatial data into rdf triples. 2014.
7. C. Stadler, M. Martin, and S. Auer. Exploring the web of spatial data with facete. In *Proceedings of the companion publication of the 23rd international conference on World wide web companion*, pages 175–178, 2014.