The GeoKnow Generator Workbench – an Integrated Tool Supporting the Linked Data Lifecycle for Enterprise Usage

Andreas Both
R&D, Unister GmbH
Leipzig (Germany)
andreas.both@unister.de

Alejandra Garcia-Rojas
Ontos A.G.
Nidau (Switzerland)
alejandra.garciarojas@ontos.com

Matthias Wauer
R&D, Unister GmbH
Leipzig (Germany)
matthias.wauer@unister.de

Daniel Hladky
Ontos A.G.
Nidau (Switzerland)
daniel.hladky@ontos.com

Jens Lehmann
Universität Leipzig, AKSW
Leipzig (Germany)
lehmann@informatik.uni-leipzig.de

ABSTRACT

Linked Data promises to make data integration easier for academic and industrial use. However, performing such data integration tasks currently requires high investments because of several major challenges. Available tools are not connected to each other, access restrictions on private data and certain tools have to be enforced, processes have to be managable and easy to use, and, finally, data processing needs to be comprehensible in terms of provenance and traceability. The GeoKnow Generator Workbench solves these problems by providing an integrated Web interface on top of an extensible solution for easy access to tools dedicated to certain Linked Data lifecycle phases, also addressing major industrial requirements. While it focuses on geospatial aspects, it is generally applicable to Linked Data management tasks.

Categories and Subject Descriptors

H.3.5 [**Online Information Services**]: Web-based services; D.2.11 [**Software Architectures**]: Data abstraction, Domain-specific architectures

Keywords

linked data management, data processing, data publishing, data provenance, integrated workbench, geospatial information systems

1. INTRODUCTION

In the last decade, many open data sources have been published following the Linked Data (LD) principles [3]. Today, even some industrial applications are driven by LD (e.g., [7]). Many LD data sources include geospatial attributes. In general, geospatial data has a high relevance in everyday life and is crucial for decision making and search applications. For instance, being able to make use of demographics and terrain data for strategic business planning, or improving search engines for questions like "find a good typical restaurant in Vienna next to the Danube river". Finding answers to such questions depends on appropriate preprocessing of a combination of geospatial and related information.

The Linked Data lifecycle[1] (c.f., Figure 1) is a blueprint for extracting data from different types of sources, interlinking with other datasets, enrichment, quality assurance, as well as exploring and visualising it. Thus, it describes the processes needed to make LD useable. Based on the LD lifecycle, the GeoKnow Generator presented in this paper enables a seamless integrated workflow and comprehensive processing options, based on a variety of tools in a modular workbench Web application, meeting industrial requirements. Most of the integrated tools include specific functionality for working with geospatial data.

The paper is organized as follows. We discuss primary requirements in Section 2. The GeoKnow Generator Workbench is presented in Section 3. Section 4 outlines applications based on the proposed solution. Related work is discussed in Section 5. Finally, the paper closes with the conclusions and future work.

2. REQUIREMENTS

In this section, we describe the use cases and the related derived requirements which led to the creation of the GeoKnow Generator.

2.1 Use Cases

Tourism e-Commerce. In this use case by Unister[2] internal data have to be enriched with public geospatial data in order to improve online search applications. Thus, Unister can understand user's search motives and support queries beyond basic hotel features.

[1] See `http://stack.linkeddata.org/`.
[2] `http://www.unister.com/`

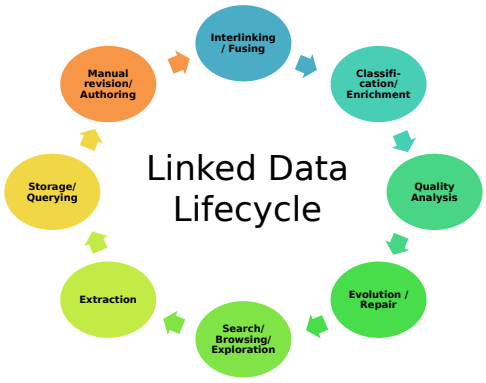


Figure 1: Linked Data Lifecycle

Supply Chain. In order to visualize key information of the logistics in a supply chain, information from supply chain transactions have to be connected to related LD. 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. This use case by a large automotive company incorporates traffic, weather, and transport information, which is linked to the supply chain information.

E-Government Services. The Linked Data Service[3] (LINDAS) has the objective to provide 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.

Automotive Data Investigation. Geosocial networks for sharing location-based messages, such as recommendations and notifications, benefit from providing context-related information. For services like community-based truck networks developed by Continental Automotive GmbH, relevant geospatial LD has to be filtered and selected, e.g., motorway service areas. Of future interest are further touristic information, such as museums and playgrounds, which are readily available in public data sets.

2.2 Requirements

Concerning functional requirements, the primary functionality of tools for LD lifecycle phases has to be extended towards geospatial data, e.g., by implementating geospatial distance metrics for interlinking and fusing datasets, and appropriate quality metrics. In addition, non-functional requirements include:

- Scalability for working with large data sets
- Authentication, Authorization and Role Management as a primary requirement in companies
- Data Provenance tracking for tracability of changes
- Job Monitoring and Robustness for applicability in production
- Modularity and Composability in order to provide flexibility w.r.t. integrating additional tools

[3] http://lindas-data.ch/

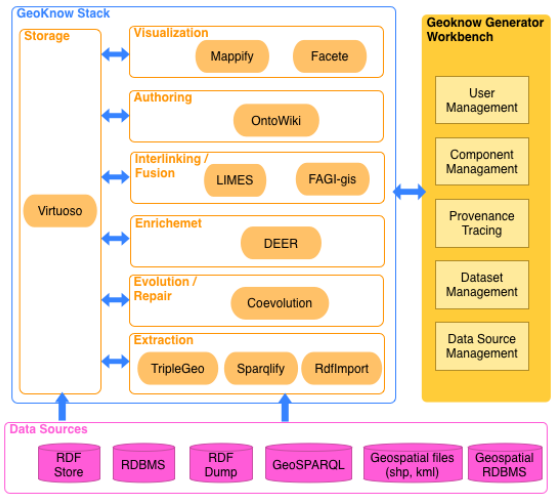


Figure 2: GeoKnow Generator Workbench

3. GEOKNOW GENERATOR

The GeoKnow Generator is a stack of tools for data preparation following the LD lifecycle. The GeoKnow Generator Workbench is the common entry point of all those tools. The actual architecture is presented in Figure 2. This diagram reflects the stack of tools integrated for each stage of the LD lifecycle. This architecture lays on following three pillars:

Software integration and deployment using the Debian packaging system. This infrastructure facilitates the packaging and integration as well as the maintenance of dependencies between the various components. Using the Debian system also enables the deployment on individual servers or cloud infrastructures.

Use of a central SPARQL endpoint and standardized vocabularies for knowledge base access and integration between the different tools. All components can access this central knowledge base repository and write their data back to it. In order for other tools to make sense out of the information it is important to define vocabularies for each of the stages of the LD lifecycle.

Integration of the user interfaces based on REST enabled Web applications. Currently the user interfaces of the various tools are technologically and methodologically heterogeneous. Thus, a common entry point for accessing the tools can forward a user to a specific UI component provided by a certain tool in order to complete a certain task. For tools that do not provide an interface, extra development effort is needed.

For integrating components, some JavaScript and basic RDF editing are required. Specifically, the AngularJS framework[4] is used for straight-forward creation of GUIs and application routing. A more detailed description of the GeoKnow Generator Workbench and how to integrate components can be found in the repository wiki[5].

[4] https://angularjs.org/
[5] https://github.com/GeoKnow/GeoKnowGeneratorUI/wiki

Tool	Description
Sparqlify [2]	SPARQL-to-SQL rewriter, enables to query RDBMS with SPARQL.
TripleGeo [9]	Geo-spatial feature extraction of ESRI shapefiles, GML, KML, INSPIRE-aligned, and several geospatially-enabled DBMSs
DEER [11]	Data enrichment with implicit geospatial information through dereferencing, interlinking and NLP.
LIMES	Link discovery framework, supports 13 similarity measures of which six are geospatial distance measures [8]
FAGI-gis [4]	Fusion of geospatial RDF data and metadata
Mappify [1]	Map view generator into in HTML/JavaScript snippets
Facete [12]	A web-based faceted browsing of RDF geospatial data
Coevolution	Service for managing dataset provenance and modifications
Virtuoso	Hybrid RDBMS/Graph Column Store cluster/cloud scalable.

Table 1: Integrated LD Stack components

Table 1 describes the actual software tools integrated in the GeoKnow Generator Workbench. Besides these integration work, the main benefits of the GeoKnow Generator Workbench are the following features: (1) Authentication and Role Management: Access to different components can be restricted via the Workbench using roles. (2) Authorisation: A graph-based security access control allows users to create and configure public and user-specific access control to datasets. For components accessing private graphs, Cross-origin resource sharing (CORS) and proxy-based model is provided. (3) 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 (Figure 3b). (4) Data Provenance: When working with several datasources and different processing stages, it is required to keep information about the provenance of certain triples. The Workbench adds metadata about the tools used to process these data, timestamp, and authors. (5) Scalability: Storage scalability is supported thanks to Virtuoso Cluster edition. Workbench and integrated tools can be easly scaled out to different nodes.

All software tools used in the GeoKnow Generator Workbench and the GeoKnow Generator Workbench itself are availeble in the LD Stack[6] repositories. The LD Stack is an independent project that aims to ease the distribution and installation and integration of LD tools developed in differnt research projects. GeoKnow project is an active contributor and supporter of the LD Stack.

4. APPLICATION IN USE CASES

In the Tourism e-Commerce use case, the GeoKnow Generator Workbench has been applied to generate an interlinked dataset used for a motive-based search infrastructure. External datasets have been transformed to RDF using TripleGeo and Sparqlify. The linking of external data and internal data was performed using LIMES with immense performance gains compared to a comparable custom approach. Besides integration of structured data, unstructured data such as hotel reviews can be processed using DEER. That way, related entities can be identified and integrated so their attributes (such as locations) can be used for further analysis of places, providing useful information for a search engine.

In the Supply Chain use case, a Dashboard (see Figure 3a) offers a unified spatial view on the logistics in the supply chain. 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. The required data integration and linking were enabled by the Sparqlify and LIMES components of the Workbench. The resulting information allows live visualisation of orders and shipments status in the Dashboard. Circulated messages and a supplier score card provide live analytics of the supply chain based on user-defined metrics.

Continental products DropYa and TruckYa use GeoKnow technology in the Automotive Data Investigation use case. 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. In the investigation process of assessing LD sources, Facete provides the functionality to browse data on a map, view attributes of interest, export relevant parts and support the editorial process.

The Ontos AG[7] Linked Data Information Workbench (OntosLDIW) is a generic, enterprise-ready workbench on the GeoKnow architecture supporting the LD Lifecycle. OntosLDIW was applied to a real world e-government scenario for the State Secretariat for Economic Affairs (SECO) in Switzerland[8]. The developed Linked Data Service (LINDAS) has centralized the tasks of the data scientist into one common workbench allowing to orchestrate, monitor and execute processes from one standardized UI. Thus, it reduces the efforts to learn various tools and front ends, improves efficiency, and reduces costs.

As generalized feedback from these use case applications, an integrated workbench brings the benefit of orchestrating the process from a single point of view. It reduces the time required for learning and switching between tools, and it reduces the interface and data exchange through a single point of access and common UI.

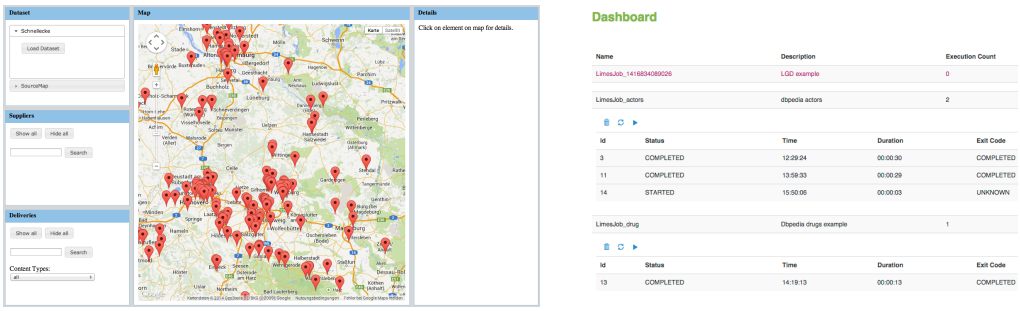
5. RELATED WORK

The LOD2 Statistical Workbench [5] provides an integrated set of tools from the LD Stack for official statistical production processes of governments. The workbench supports many different operations. This solution is suitable for a specific use case but lacks the general applicability of a more configurable approach. Unifiedviews[6] is a LD processing

[6] http://stack.linkeddata.org

[7] http://ontos.com/

[8] http://www.seco.admin.ch/?lang=en



(a) The supply chain dashboard.

(b) Task monitoring dashbord.

Figure 3: Use Cases and Applications of GeoKnow Generator Workbench

framework created under the EU project COSMODE[9] using components from the LD Stack. This platform requires implementing Data Processing Units for each component in order to be integrated. Moreover, Unifiedviews doesn't provide support for authentication or authorisation features. Still, it represents a relevant reference point for the GeoKnow Generator Workbench. [10] presents a workbench for publishing geospatial linked data. In contrast to our work, the data processing is highly specialized and does not provide solutions for all steps of the LD lifecycle.

6. CONCLUSIONS AND FUTURE WORK

The main contribution of this paper, presented in 3, is the GeoKnow Generator Workbench, which is a web-based user interface that integrates all components needed for processing data following the LD lifecycle. It enables simple access and interaction with the different components needed for different tasks. Moreover, it provides APIs for being integrated into other systems and to exchange the components currently available out-of-the-box. Future tools can be integrated easily. An online demo and video tutorials of the Workbench are available at `http://generator.geoknow.eu`.

We described the GeoKnow Generator and the main features enabling an enterprise use. All requirements, including those w.r.t managing geospatial data, are derived from real world use cases, which also demonstrate the usability of the Generator components and the Workbench in enterprise environments. In the future we will integrate additional tools and decouple the Workbench from Virtuoso.

Acknowledgments.

This work is part of the European Commission FP7 Project GeoKnow (GA No 318159).

7. REFERENCES

[1] Mappify: a tool to easily create interactive maps backed by semantic web technologies. `http://mappify.aksw.org/`.

[2] Sparqlify: a sparql-sql rewrite. `http://aksw.org/Projects/Sparqlify.html`.

[3] C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227, 2009.

[4] 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.

[5] 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 âĂŞ Interoperability, Usability, Applicability*, 2014.

[6] 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.

[7] G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R. Lee. Media meets semantic web–how the bbc uses dbpedia and linked data to make connections. In *The semantic web: research and applications*, pages 723–737. Springer, 2009.

[8] A.-C. N. Ngomo. Orchid - reduction-ratio-optimal computation of geo-spatial distances for link discovery. In *International Semantic Web Conference*, pages 395–410, 2013.

[9] K. Patroumpas, M. Alexakis, G. Giannopoulos, and S. Athanasiou. Triplegeo: an etl tool for transforming geospatial data into rdf triples. 2014.

[10] A. Shaon, A. Woolf, R. Boczek, W. Rogers, and M. Jackson. *An Open Source Linked Data Framework for Publishing Environmental Data under the UK Location Strategy*, volume 798. CEUR Workshop Proceedings, 2011.

[11] M. Sherif, A.-C. Ngonga Ngomo, and J. Lehmann. Automating RDF dataset transformation and enrichment. In *12th Extended Semantic Web Conference, Portoroz, Slovenia, 31st May - 4th June 2015*. Springer, 2015.

[12] 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.

[9] `http://www.cosmode.eu/`