The GeoKnow Generator: Managing Geospatial Data in the Linked Data Web

by Jon Jay Le Grange, Jens Lehmann, Spiros Athanasiou, Alejandra Garcia Rojas, Giorgos Giannopoulos, Daniel Hladky, Robert Isele, Axel-Cyrille Ngonga Ngomo, Mohamed Ahmed Sherif, Claus Stadler, Matthias Wauer

Within the GeoKnow project, various tools are developed and integrated which aim to simplify managing geospatial Linked Data on the web. In this article, we give a first presentation of the GeoKnow Generator, which is the platform providing a light-weight integration of those tools.

Introduction

In recent years, Semantic Web methodologies and technologies have strengthened their position in the areas of data and knowledge management. Standards for organizing and querying semantic information, such as RDF(S) and SPARQL have been adopted by large academic communities, while corporate vendors adopt semantic technologies to organize, expose, exchange and retrieve their data as Linked Data [1]. RDF stores have become robust enough to support volumes of billions of records (RDF triples), and also offer data management and querying functionalities very similar to those of traditional relational database systems. Currently, there are three major sources of open geospatial data in the Web: Spatial Data Infrastructures (SDI), open data catalogues, and crowdsourced initiatives. Crowdsourced geospatial data are emerging as a potentially valuable source of geospatial knowledge. Among various efforts we highlight OpenStreetMap, GeoNames, and Wikipedia as the most significant. Recently, GeoSPARQL [2] has emerged as a promising standard from W3C for geospatial RDF, with the aim of standardizing geospatial RDF data modelling and querying. Integrating Semantic Web with geospatial data management requires the scientific community to address two challenges: (i) the definition of proper standards and vocabularies that describe geospatial information according to RDF(S) and SPARQL protocols, that also conform to the principles of established geospatial standards, (eg OGC), (ii) the development of technologies for efficient storage, robust indexing, and native processing of semantically organized geospatial data.

The GeoKnow Project

Geoknow is an EU funded, three-year project that started in December 2012. While several research projects, such as LOD2[3], support the Linked Data LifeCycle, Geoknow addresses the key issues of integrating geographically related information on the Web, scalable integration over millions of geospatial entities within the Linked Data Web, as well as efficient browsing and exploration of geographic information. In particular, GeoKnow will apply the RDF model and the GeoSPARQL standard as the basis for representing and querying geospatial data.

GeoKnow Generator Architecture and Implementation

The GeoKnow Generator includes several different software tools for application users or application developers. The initial architecture is depicted in figure 1. The software tools that target expert users in DB administration or designers are essentially web applications and accessible through the Debian repository from the Linked Data Stack. The Generator Workbench is GeoKnow's main application that integrates preconfigured components from the Stack according to the Linked Data lifecycle as a workflow. It provides access to public data catalogues of the domain of knowledge and the option to add proprietary datasets. It also aims to provide a layer for user administration, authorisation and provenance. The components that are integrated in this Workbench communicate using HTTP, REST or SPARQL protocols.

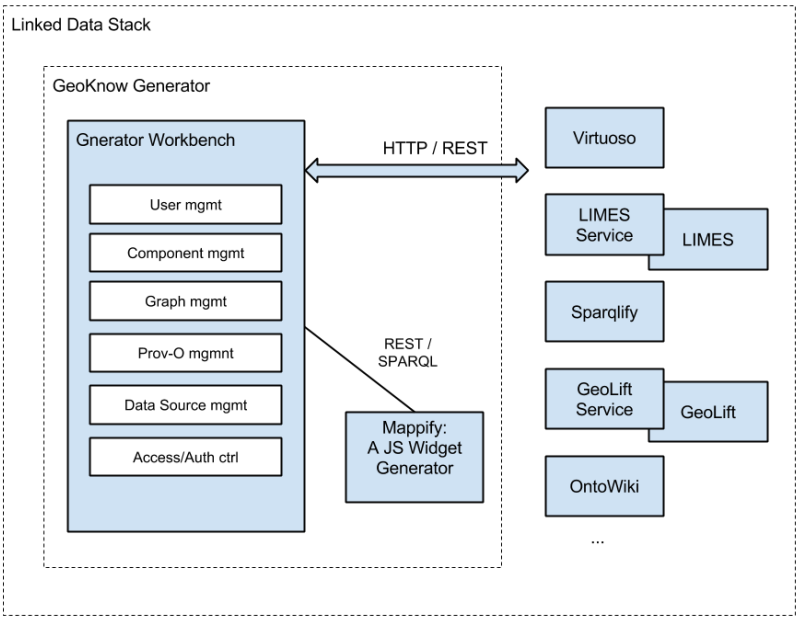

Figure 1: Current version of the GeoKnow Generator

A prototype of the GeoKnow Generator is already available at http://generator.geoknow.eu. It allows the user to triplify geospatial data, such as ESRI shapefiles and spatial tables hosted in major DBMSs using the GeoSPARQL, WGS84 or Virtuoso RDF vocabulary for point features geospatial representations (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 Generator's triple store. With these two sources of data it is possible to link (via LIMES), to enrich (via GeoLift), to query (via Virtuoso) , to visualize (via Facete) and to generate light-weight applications as JavaScript snippets (via Mappify) for specific geospatial applications. Most steps in the Linked Data lifecycle [1] have been integrated in the Generator as a graph-based workflow, which allows the user to easily manage new generated data. The current version of the GeoKnow Generator is presented in Figure 1. The components comprising it are available in the Linked Data Stack (http://stack.linkeddata.org)

GeoKnow Tools

Extraction and Loading

TripleGeo

TripleGeo is a utility developed by the Institute for the Management of Information Systems at Athena Research Center. This generic purpose, open-source tool can be used for integrating features from geospatial databases into RDF triples.

Sparqlify

Sparqlify is a SPARQL-to-SQL rewriter that enables one to define RDF views on relational databases and query them with SPARQL. It is currently in beta state and features basic support for spatial data types. Sparqlify powers the Linked Data interface and a SPARQL endpoint of the LinkedGeoData Server, where access to billions of virtual triples from the OpenStreetMap database is provided.

Querying and Exploration

Facete

Facete offers out-of-the-box faceted search over SPARQL end-points to ease the navigation of RDF data using advanced faceted search techniques. Spatial data is automatically detected and visualized on a map, even if the geometric information is only indirectly related to the resources specified by the faceted search.

Virtuoso

Virtuoso is a scalable high-performance RDF Quad Store available in open source and commercial forms, providing the core geo spatial knowledge storage for the GeoKnow generator. Virtuoso provides optimised distributed SPARQL and SQL query processing for heterogeneous data integration across data sources.

Mappify

Mappify is a tool to easily create map view snippets to be included in your web site. It builds on re-usable components of Facete and thus enables a facet based definition of points of interests based on a SPARQL-accessible dataset. Users are enabled to quickly style the map display by choosing marker icons and defining templates for the content to show when clicking the markers.

Authoring

OntoWiki

OntoWiki facilitates the visual presentation of a knowledge base as an information map, with different views on instance data. It enables intuitive authoring of semantic content, with an inline editing mode for editing RDF content, similar to WYSIWIG for text documents.

Linking

LIMES

LIMES is a link discovery framework for the Web of Data which addresses both the time complexity of link discovery and the complexity of discovering accurate link specifications [6]. To address the first

problem, LIMES implements a hybrid approach which is uses set semantics to combine the output of manifold measure-specific algorithms. The second problem is addresses by novel unsupervised as well as supervised (batch and active) approaches to learning link specifications.

Enriching and Data Cleaning

Geolift

GeoLift is a spatial mapping component which aims to enrich RDF datasets with geo-spatial information. To achieve this goal, GeoLift relies on three atomic modules based on dereferencing, linking and Named Entity Recognition and Disambiguation. The dereferencing module enriches the geo-spatial datasets by finding the URI objects in the source dataset and dereferencing the URIs in order to add related geographical information, e.g. geo:lat if it exists. The linking module adds valuable geo-spatial information to the datasets through linking the dataset that is targeted to be enriched with another dataset. This produces geo-spatial relationships that are associated to the linked URIs in the enriched dataset. The Named Entity Recognition and Disambiguation is based on the FOX[1] and AGDISTIS[2] frameworks, which find person, location and organization mentions in text and ground them in DBpedia.

Conclusion and Future Work

Geoknow is concluding its first year and has already achieved important advancements. The first step was to perform a thorough evaluation of the current standards and technologies for managing geospatial RDF data and identify major shortcomings and challenges [4]. The next step has already produced significant output in the form of ready-for-use tools comprising the GeoKnow Generator. These components are being further enhanced and enriched. For example, Virtuoso RDF store is being extended in order to fully support OGC geometries and the GeoSPARQL standard and FAGI is being developed to support fusion of thematic and geospatial metadata of resources, either manually or automatically. Also, within 2014 the consortium will start testing the use cases and evaluating the performance and scalability of the GeoKnow Generator. Finally, future activities include, among others, the enhancement of the already developed frameworks, as well as the development of sophisticated tools for (a) aggregation of crowdsourced geospatial information and (b) exploration and visualization of spatio-temporal RDF data.

Acknowledgement: The research leading to these results has received funding under the European Commission's Seventh Framework Programme from ICT grant agreement (no. 318159) for GeoKnow. The consortium consists of the following partners: Institute of Applied Computer Science / University of Leipzig (Germany), Institute for the Management of Information Systems/Athena Research and Innovation Center (Greece), OpenLink Software Ltd (United Kingdom), Unister GmbH (Germany), Brox (Germany), Ontos AG (Switzerland), and Institute Mihailo Pupin (Serbia).

Links:

[1] http://fox.aksw.org
[2] http://aksw.org/projects/agdistis

GeoKnow project: http://geoknow.eu
LOD2 project: http://lod2.eu
Linked Data Stack: http://stack.linkeddata.org
GeoKnow Github: https://github.com/GeoKnow
Generator Demo: http://generator.geoknow.eu
Open Geospatial Consortium: http://www.opengeospatial.org/

References:

[1] S. Auer, J. Lehmann: "Making the web a data washing machine - creating knowledge out of interlinked data", Semantic Web Journal, volume 1, number 12, p. 97-104, IOS Press, 2010, http://www.semantic-web-journal.net/sites/default/files/swj24_0.pdf
[2] M. Perry, J. Herring (eds): "OGC GeoSPARQL standard - A geographic query language for RDF data", Open Geospatial Consortium Inc, v.1.0, 27/04/2012, https://portal.opengeospatial.org/files/?artifact_id=47664
[3] S. Auer, L. Bühmann, J. Lehmann, M. Hausenblas, S. Tramp, B. van Nuffelen, P. Mendes, C. Dirschl, Robert Isele, Hugh Williams: "Managing the lifecycleof Linked Data with the LOD2 Stack", In: Proceedings of the 11st International Semantic Web Conference (ISWC), Springer, 2012.
[4] K. Patroumpas et al.: "Market and Research Overview", GeoKnow EU/FP7 project deliverable 2.1.1, 2013, http://svn.aksw.org/projects/GeoKnow/Public/D2.1.1_Market_and_Research_Overview.pdf
[5] A. G. Rojas, S. Athanasiou, J. Lehmann, D. Hladky: "GeoKnow: Leveraging Geospatial Data in the Web of Data", In: Open Data Workshop, W3C, London, 2013.
[6] A.-C. Ngonga Ngomo: "On Link Discovery using a Hybrid Approach". In: Journal on Data Semantics 1(4): 203-212, 2012