1. Publishable Summary

The GeoKnow project (http://geoknow.eu) aims to make geospatial data more accessible and re-usable on the Web of Data and contributes to the transformation of the Web into a place where geospatial data can be published, queried, reasoned, and interlinked according to Linked Data principles. The partners involved in GeoKnow are (alphabetically): Athena, BROX, InfAI (coordinator), IMP, Ontos, OpenLink and Unister.

Motivation

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 are adopted by large academic communities, while corporate vendors adopt semantic technologies to organize, expose, exchange and retrieve their data as Linked Data. 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 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.

Results of the First Year

The Linked Data Stack: The underlying vision and structure of GeoKnow is to realize a Linked Data Lifecycle for geospatial knowledge. The lifecycle contains different phases such as data extraction, authoring, querying, visualisation etc. Each of the lifecycle phases can be supported by components, which are developed and extended within the consortium. To allow a lightweight integration between those tools, we build on the LOD2 Stack infrastructure of the LOD2 project, one of the largest Linked Data research projects in Europe. Within GeoKnow, we extended this and developed the Linked Data Stack as a project independent infrastructure and the jointly developed successor of the LOD2 Stack. The Linked Data Stack is available on http://stack.linkeddata.org and contains a repository of Debian packages to manage the installation and dependencies between Linked Data Stack components. The first official release is planned in the first half of 2014.

The GeoKnow Generator: Based on a comprehensive requirements analysis using a survey, we started the development of the GeoKnow Generator (GKG) interface. An early prototype

is already available at http://generator.geoknow.eu. The GKG is the interface, which provides an integrated view over the Linked Data Stack components developed in GeoKnow. 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.

Benchmarking and Standard Compliance: Within the first year, we performed a comprehensive survey on whether current geospatial triples stores support the GeoSPARQL standard. Furthermore, we developed a benchmarking laboratory to assess query performance, performed optimisations of triples stores and evaluated them.

Data Integration: Geospatial data integration is one of the core goals of GeoKnow. In the first year, we developed the ORCHID algorithm, which is now the fastest algorithm for geospatial RDF data integration according to our evaluation on large scale datasets.

Prototype Development: Within GeoKnow many software prototypes were developed and made available under open source licenses. Many of those are available at https://github.com/GeoKnow, which has been very active throughout the project. Newly developed software components are:

- GeoLift – annotates existing RDF datasets with geospatial information
- TripleGeo – converts shapefiles and other geospatial structures to triples
- Mappify – allows to create simple geospatial web applications
- Jassa – a JavaScript library which allows developing powerful applications on top of SPARQL endpoints
- FAGI – a tool for fusion and aggregation of geospatial information

The following prototypes were improved throughout the first year:

- Sparqlify – a SPARQL-2-SQL rewriter with an enhanced architecture providing better performance and geospatial support
- LinkedGeoData – the RDF version of OpenStreetMap, which is now automatically published monthly and integrated with GADM (Global Administrative Boundaries)
- Facete – a generic, facet-based RDF browser
- LIMES – a framework for RDF data integration, which now contains the ORCHID algorithm to interlink geospatial data efficiently
- Virtuoso – triple store and middleware, which now contains geospatial query performance improvements

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. 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.

The GeoKnow Consortium

End-user communities: Unister and BROX will assure that the project is driven by real-world problems. BROX's mission is to provide the viewpoint and requirements of logistic companies. Unister adds the corporate viewpoint and a comprehensive enterprise technology to the consortium. OpenStreetMap is integrated via the LinkedGeoData project. Open Street Map(OSM) represents the widest possible public with OSM's web services being one of the most visited GIS applications on the Internet.

High-tech companies: Ontos, OpenLink and IMP. Both Ontos and OpenLink are W3C members. OpenLink is an expert in distributed data management and integration. It brings the Open Source Virtuoso suite and other software products and know-how to the consortium. Ontos contributes its expertise in technically coordinating the architecture of research projects, its research on semantic data integration and its experience with large repositories of enterprise business objects. IMP is an applied research institute which in many aspects works more like a company. IMP has 20 years long experience in the design, realization, and implementation of intelligent software systems and other customized IT solutions. IMP has extensive experience in using Linked Data tools and technologies and developing Linked Data applications.

University research groups: InfAI and Athena. InfAI is a world-leading research group in social semantic collaboration technologies and brings the OntoWiki open-source framework and Web information integration technologies (such as LIMES, Triplify, DBpedia and LinkedGeoData) to the project. Athena is a leading research centre for geo-spatial data management, with a particular focus on spatial data modelling, geo-spatial web services and evaluating GIS system performance. The research institutions (InfAI, Athena) have outstanding experience in the dissemination of research results via all kinds of channels, such as conferences, workshops, journal papers.