# GeoKnow: Leveraging Geospatial Data in the Web of Data

Alejandra Garcia-Rojas[1], Spiros Athanasiou[2], Jens Lehmann[3] and Daniel Hladky[1]

[1]Ontos AG
[2]Institute for the Management of Information Systems, R.C. "Athena"
[3]Institute of Applied Computer Science, University of Leipzig

## 1 Motivation

Producing and updating geospatial data is expensive and resource intensive. Hence, it becomes crucial to be able to integrate, repurpose and extract added value from geospatial data to support decision making and management of local, national and global resources. Spatial Data Infrastructures (SDIs) and the standardisation efforts from the Open Geospatial Consortium (OGC) serve this goal, enabling geospatial data sharing, integration and reuse among Geographic Information Systems (GIS). Geospatial data are now, more than ever, truly syntactically interoperable. However, they remain largely isolated in the GIS realm and thus absent from the Web of Data. Linked data technologies enabling semantic interoperability, interlinking, querying, reasoning, aggregation, fusion, and visualisation of geospatial data are only slowly emerging. The vision of GeoKnow is to leverage geospatial data as first-class citizens in the Web of Data, in proportion to their significance for the data economy.

## 2 Open Geospatial Data on the Web

Currently, there are three major sources of open geospatial data in the Web: Spatial Data Infrastructures, open data catalogues, and crowdsourced initiatives.

**Spatial Data Infrastructures** (SDIs) were created to promote the discovery, acquisition, exploitation and sharing of geographic information. They include technological and organisational structures, policies and standards that enable efficient discovery, transfer and use of geospatial data using the web [1]. Research and development in this field is closely tied to standardisation activities led by international bodies, namely the ISO/TC 211[1], OGC[2] and W3C[3]. In Europe, the INSPIRE Directive follows the OGC open standards, and has

[1]ISO /TC 211 Geographic Information/Geomatics, http://www.isotc211.org/
[2]Open Geospatial Consortium, http://www.ogc.org
[3]W3C, http://www.w3.org/

defined common data models for a number of application domains, such as hydrography, protected sites and administrative units, to enhance interoperability of spatial data sets of the different European countries[4]. It provides the legal and technical foundations to ensure member state SDIs are compatible and usable on a transboundary context.

The major open standard Web services regarding discovery and querying of geospatial data in SDIs are OGCs Catalogue Service and Web Feature Service respectively. The first allows the discovery of geospatial data based on their metadata (e.g. scale, coverage) and the second enables querying of geospatial data. Additional standards provide access to maps and tiles (Web Map Service, Web Tile Service) and enable developers to programmatically invoke and compose complex geospatial analysis services (Web Processing Service). Currently practically all GIS and geospatial databases are fully compatible with these standards; GIS users can consume geospatial data from SDIs and publish geospatial data to SDIs with a few clicks. On a practical level, it is clear that SDIs must be considered as diachronic and stable data infrastructures. They represent a significant investment from the public and private worldwide and are the basis for interoperability among significant scientific domains. Further, they constitute the most prominent source for high-quality open geospatial data. Thus, any contribution and advancement must either be directly involved in standardization efforts, or be based solely on existing standards, without directly affecting their applications.

**Open data catalogues** provide open geospatial data by a) encapsulating existing SDIs and/or b) ad hoc publishing available geospatial data as files. In the latter case, geospatial data are published as regular open data. The only difference regards the use of file formats of the geospatial domain (e.g. shp, kml) and availability of data for specific coordinate reference systems (typically national CRS). In the former case, an available national/regional SDI is exploited as a source for harvesting its geospatial data. The Catalogue Service is used to discover available data, and their metadata are added in the open data catalogue for homogenised data discovery. The actual data are available as exported file snapshots in common geospatial formats as before, or through the query services provided by the SDI. Consequently, open data catalogues typically offer geospatial data as files and at best expose any available SDI services for data access.

**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. GeoNames[5] provides some basic geographical data such as latitude, longitude, elevation, population, administrative subdivision and postal codes. This data is available as text files and also accessible through a variety of web services such as free text search, find nearby or even elevation data services. Providing a larger variety of data, OpenStreetMap[6] (OSM) has become an important platform for mapping, browsing and visualising spatial data on the Web. OSM data is available in

[4]EC. Inspire directive, 2009. http://inspire.jrc.ec.europa.eu/

[5]http://www.geonames.org/

[6]http://www.openstreetmap.org/

different formats[7] which can be imported into a database for its usage; it also provides web services to do search by name and inverse geocoding functionality.

## 3 Semantic Web Technologies for Geospatial Data

The benefits of semantic technology for spatial data management are explored in a number of topics. For example, ontologies have been used in the form of taxonomies on thematic web portals (e.g. habitat or species taxonomies, categories of environmentally sensitive areas, or hierarchical land use classifications). The role of these ontologies is however limited. They provide background knowledge, but only in some experimental prototypes they are used for constructing search requests or for grouping of search results into meaningful categories. Further, in experimental settings, there are examples of using OWL for bridging differences in conceptual schemas, e.g. [2]. The role of ontologies and knowledge engineering in these prototypes is basically to provide methodologies for integration and querying [3][4]. Ontologies have played an important role in structuring data of geospatial domains [5][6]. However, semantic technology has not influenced spatial data management yet, and mainstream GIS tools are not yet extended with semantic integration functionality.

Early work included the Basic Geo Vocabulary[8] by the W3C, which enabled the representation of points in WGS84, and GeoRSS [7] , which provided support for more geospatial objects (lines, rectangles, polygons). In addition, GeoOWL[9] was developed to provide a more flexible model for geospatial concepts. Furthermore, topological modelling of geometric shapes in RDF can be done with the NeoGeo Geometry Ontology[10]. However, all these ontologies only supported WSGS84, and currently offer limited support for geospatial operations required in real world GIS workloads.

GeoSPARQL has emerged as a promising standard from W3C for geospatial RDF, with the aim of standardising geospatial RDF data insertion and query. GeoSPARQL provides various conformance classes concerning its implementation of advanced reasoning capabilities (e.g. quantitative reasoning), as well as several sets of terminology for topological relationships between geometries. Therefore, different implementations of the GeoSPARQL specification are possible, depending on the respective domain/application. In addition, GeoSPARQL closely follows existing standards from OGC for geospatial data, to facilitate spatial indexing from relational databases.

## 4 Enter the Geospatial Data Web

GeoKnow is a recently established EU research project, motivated by our previous work in the LinkedGeoData [8] project (LGD), which makes OpenStreetMap data available as an RDF knowledge base. As a result, OSM data were introduced in the LOD cloud and interlinked with GeoNames, DBpedia [9], and multiple other data sources. LGD intended to simplify information creation and aggregation related to spatial features. During this exercise, several research challenges were found such as scalability with spatial data, query performance,

[7] $http://wiki.openstreetmap.org/wiki/OSM_file_formats$
[8] http://www.w3.org/2003/01/geo/
[9] GeoOWL. http://www.w3.org/2005/Incubator/geo/XGR-geo-20071023/
[10] NeoGeo Geometry Ontology. http://geovocab.org/geometry.html

spatial data modelling, flexible transformation of special data, as well data operations such routing data. It was realised that geospatial data, specially scientific data, available on the web can open new opportunities to improve management and decision making applications.

Our vision is to make geospatial data accessible on the web of data and turn the Web in a place where geospatial data can be published, queried, reasoned, and interlinked, according to the Linked Data principles (see e.g. [10] for a description of the data lifecycle). This will move geospatial data beyond syntactic interoperability to actual semantic interoperability, and to services that can geospatially reason on the Web. Linked data will not only be extended with spatial data to be able to improve information retrieval based on geospatial data [11], or to answer questions that were not possible with isolated geospatial data, but also represents a step towards the discoverability of data that share geospatial features (i.e. supported by querying and reasoning), and a boosting for the geospatial data integration through geospatial data merging and fusing tools.

Our work will repurpose SDI standards, enabling the existing vast body of geospatial knowledge to be introduced in the Data Web. Further, we will apply the RDF model and the GeoSPARQL standard as the basis for representing and querying geospatial data. In particular, GeoKnow contributions will be in the following areas:

**Efficient geospatial RDF querying.** Existing RDF stores lack performance and geospatial analysis capabilities compared to geospatially-enabled relational DBMS. We will focus on introducing query optimisation techniques for accelerating geospatial querying at least an order of magnitude.

**Fusion and aggregation of geospatial RDF data.** Given a number of different RDF geospatial data for a given region containing similar knowledge (e.g. OSM, PSI and closed data[11]) we will devise automatic fusion and aggregation techniques in order to consolidate them and provide a data set of increased value and quantitative quality metrics of this new data resource

**Visualisation and authoring.** We will develop reusable mapping components, enabling the integration of geospatial RDF data as an addition data resource in web map publishing. Further, we will support expert and community-based authoring of RDF geospatial data within interactive maps, fully embracing crowdsourcing.

**Public-private geospatial data.** To support value added services on top of open geospatial data, we will develop enterprise RDF data synchronisation workflows that can integrate open geospatial RDF with closed, proprietary data. We will focus on the supply chain and e-commerce use cases.

**GeoKnow Generator.** This will consist of a full suite of tools supporting the complete lifecycle of geospatial linked open data. The GeoKnow Generator will enable publishers to triplify geospatial data, interlink them with other

[11] $http://ec.europa.eu/information_{s}ociety/policy/psi/index_{e}n.htm$

geospatial and non-geospatial Linked Data sources, fuse and aggregate linked geospatial data to provide new data of increased quality, visualise and author linked geospatial data in the Web.

Having the web of data enriched with spatial data implies the consideration of standards such as SDI in the creation of a Geospatial Semantic Web. Working with spatial data is challenged by the scalability requirements due to the size of datasets, the integration of data considering the different models, and the transformation of data in specialised domains. GeoKnow aims to provide easy to use tools for non experts in cartography nor Linked Data to exploit the data and create web based geospatial enabled applications.

Work on GeoKnow is founded by the European Commission within the FP7 Information and Communication Technologies Work Programme (Grant Agreement No. 318159).

## References

[1] D. Nebert. Developing spatial data infrastructures: The sdi cookbook. technical report, global spatial data infrastructure, 2004.

[2] Catherine Dolbear and Glen Hart. Ontological bridge building - using ontologies to merge spatial datasets. In *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*, pages 15–20. AAAI, 2008.

[3] Tian Zhao, Chuanrong Zhang, Mingzhen Wei, and Zhong-Ren Peng. Ontology-based geospatial data query and integration. In *GIScience*, volume 5266 of *Lecture Notes in Computer Science*, pages 370–392. Springer, 2008.

[4] Agustina Buccella, Alejandra Cechich, and Pablo R. Fillottrani. Ontology-driven geographic information integration: A survey of current approaches. *Computers and Geosciences*, 35(4):710–723, 2009.

[5] Albrecht, Jochen, Derman, Brandon, Ramasubramanian, and Laxmi. Geo-ontology tools: The missing link. *Transactions in GIS*, 12(4):409–424, 2008.

[6] Eva Klien and Florian Probst. Requirements for geospatial ontology engineering. In *8th Conference on Geographic Information Science (AGILE 2005)*, pages 251–260. Citeseer, 2005.

[7] Open Geospatial Consortium Inc. An introduction to georss: A standards based approach for geo-enabling rss feeds. White paper, OGC, 2006.

[8] Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web Journal*, 3(4):333–354, 2012.

[9] Mohamed Morsey, Jens Lehmann, Sören Auer, Claus Stadler, and Sebastian Hellmann. Dbpedia and the live extraction of structured data from wikipedia. *Program: electronic library and information systems*, 46:27, 2012.

[10] Sören Auer and Jens Lehmann. Making the web a data washing machine - creating knowledge out of interlinked data. *Semantic Web Journal*, 2010.

[11] Krzysztof Janowicz, Marc Wilkes, and Michael Lutz. Similarity-based information retrieval and its role within spatial data infrastructures. In Thomas J. Cova, Harvey J. Miller, Kate Beard, Andrew U. Frank, and Michael F. Goodchild, editors, *GIScience*, volume 5266 of *Lecture Notes in Computer Science*, pages 151–167. Springer, 2008.