Collaborative Project

GeoKnow - Making the Web an Exploratory for Geospatial Knowledge

Project Number: 318159  Start Date of Project: 2012/12/01  Duration: 36 months

Deliverable 7.1.7

Final Report

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>Public</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due Date of Deliverable</td>
<td>Month 36, 30/11/2015</td>
</tr>
<tr>
<td>Actual Submission Date</td>
<td>Month 36, 30/11/2015</td>
</tr>
<tr>
<td>Work Package</td>
<td>WP7</td>
</tr>
<tr>
<td>Task</td>
<td>T7.1</td>
</tr>
<tr>
<td>Type</td>
<td>Report</td>
</tr>
<tr>
<td>Approval Status</td>
<td>Approved</td>
</tr>
<tr>
<td>Version</td>
<td>1.0</td>
</tr>
<tr>
<td>Number of Pages</td>
<td>27</td>
</tr>
<tr>
<td>Filename</td>
<td>D7.1.7 Final Report.pdf</td>
</tr>
</tbody>
</table>

Abstract: This document summarises all research activities performed and software components developed in GeoKnow. The usage of GeoKnow components is demonstrated by presenting final results of Use Cases. Finally, several sustainability plans for the different software components and a summary of dissemination activities are described.

The information in this document reflects the author’s views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/her sole risk and liability.

Project funded by the European Commission within the Seventh Framework Programme (2007 - 2013)
History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Reason</th>
<th>Revised by</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>20/11/2015</td>
<td>Initial draft of this document</td>
<td>Alejandra Garcia Rojas M.</td>
</tr>
<tr>
<td>0.2</td>
<td>22/11/2015</td>
<td>Internal contribution</td>
<td>All partners</td>
</tr>
<tr>
<td>0.3</td>
<td>22/11/2015</td>
<td>Use Cases update</td>
<td>Mattias Wauer and Rene Pietzsch</td>
</tr>
<tr>
<td>0.5</td>
<td>26/11/2015</td>
<td>Internal and peer Review</td>
<td>Vuk Mijovic (IMP)</td>
</tr>
<tr>
<td>0.5</td>
<td>28/11/2015</td>
<td>Address peer review comments</td>
<td>Alejandra Garcia Rojas M.</td>
</tr>
<tr>
<td>1.0</td>
<td>30/11/2015</td>
<td>Approval</td>
<td>Daniel Hladky</td>
</tr>
</tbody>
</table>

Author List

<table>
<thead>
<tr>
<th>Organization</th>
<th>Name</th>
<th>Contact Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ontos</td>
<td>Alejandra Garcia Rojas M.</td>
<td><a href="mailto:alejandra.garciarojas@ontos.com">alejandra.garciarojas@ontos.com</a></td>
</tr>
<tr>
<td>Unister</td>
<td>Mattias Wauer</td>
<td><a href="mailto:matthias.wauer@unister-gmbh.de">matthias.wauer@unister-gmbh.de</a></td>
</tr>
<tr>
<td>Brox</td>
<td>Rene Pietzsch</td>
<td><a href="mailto:rpietzsch@brox.de">rpietzsch@brox.de</a></td>
</tr>
</tbody>
</table>
Executive Summary

This report aims to summarise the achievements of GeoKnow project. This document summarises the research achievements related to:

- the extraction and management of geospatial data within the RDF standard,
- the spatial data aggregation and fusion,
- the quality assessment of geospatial data,
- the visualisation and curation of geospatial data,
- and software integration efforts for working with Linked Data tools.

It also describes GeoKnow Use Cases achievements, the sustainability plans for GeoKnow exploitable results which includes software tools and benchmarking tools, and finally a summary of dissemination activities.
Table of Contents

1 Introduction 5

2 Project Outcomes 7
   2.1 Geospatial information extraction and management 7
   2.2 Spatial knowledge aggregation and fusing 8
   2.3 Quality Assessment 8
   2.4 Visualisation and Data Curation 9
   2.5 Complementary Tools 10
   2.6 Benchmarking Systems 10
   2.7 GeoKnow Generator Workbench 10

3 Use Cases Achievements 13
   3.1 e-Commerce 13
   3.2 Supply Chain 17

4 Sustainability Plans 21
   4.1 Linked Data Stack 21
   4.2 Individual Software Tools 21

5 GeoKnow Dissemination 26

6 Conclusions 27
List of Figures

1  Linked Data Life Cycle and GeoKnow contribution  .................................................. 5
2  GeoKnow Generator and Workbench  ........................................................................ 11
3  e-Commerce data preparation flow  ........................................................................ 15
4  e-Commerce Motive-based search demonstration  .................................................... 16
5  Supply Chain data preparation flow  ........................................................................ 18
6  Supply Chain dashboard and Xybermotive widgets  .................................................. 20

List of Tables

1  Integrated dataset for e-Commerce use case  ................................................................. 14
2  Integrated dataset for Supply Chain use case  .............................................................. 17
3  Sustainability Plans for GeoKnow tools  ..................................................................... 22
4  GeoKnow dissemination activities summary  ............................................................... 26
1 Introduction

GeoKnow project has now arrived to the end providing several tools for bringing and working with geospatial data in the Web of data. The research activities carried out these last three years are summarized in this document, and include:

- the extraction and transformation tools for extracting geospatial data from Geospatial Information Systems (GSI),
- the improvement of the capabilities for storing and managing this data in RDF stores, including benchmarking tools for evaluating these capabilities,
- the methodologies for an efficient linking and fusion of geospatial data, including benchmarking tools for evaluation,
- the development of tools for visualising and exploring datasets, and
- the tools for assessing the quality of the data
- the creation of a unified interface for working with the different components supporting the LD lifecycle

These research activities produced software components that fulfil one or more stages in the Linked Data (LD) lifecycle. As depicted in figure 1, GeoKnow components cover all the different stages in the lifecycle. To facilitate working with such heterogeneous components, GeoKnow developed the GeoKnow Generator Workbench, which is an interface that unifies and orchestrates the use of GeoKnow tools. The use of the
GeoKnow Generator components and Workbench was demonstrated during the project with two use case demonstrators by Unister¹ and Brox². This document presents the data preparation for the use cases. However, a more extensive list of use cases was presented in D7.1.5 GeoKnow Show Case³.

Most of the components developed in GeoKnow were distributed in the Linked Data Stack⁴ platform, which has been maintained by GeoKnow. With the end of this project we need to make sure that the efforts for maintaining this platform will continue and the software created will be also reused. Thus, this report presents several sustainability plans provided by the different GeoKnow partners.

This document is organised as follows. Section 2 summarises the research achievements and the contribution to the state of the art. Section 3 describes how GeoKnow components were applied to the two GeoKnow use cases and the achieved results. Section 4 we describe the sustainability plans for GeoKnow exploitable results which includes software tools and benchmarking tools. Finally in section 5 a summary of dissemination activities for the three years is presented.

¹ http://www.unister.com/
² http://www.brox.de/
⁴ http://stack.linkeddata.org/
2 Project Outcomes

This project has produced 14 open source tools for supporting the integration of existing geospatial data into the web of data. This project has offered several different ways of integrating geospatial and non-geospatial data from different sources (RDBMS, file system, Endpoints).

2.1 Geospatial information extraction and management

One of the main contributions of GeoKnow was to take geospatial data out of GIS for making it accessible on the web. This will allow the access to non experts, and to have a self-explanatory spatial information structures accessible via standard Web protocols and support for ad-hoc definable, flexible information structures. GeoKnow developed and improved Sparqlify and TripleGeo⁵. These tools allow for transforming geospatial data from several conventional formats, into RDF triples, compliant with several standards (GeoSPARQL, Virtuoso vocabulary, etc). Sparqlify has been tested for mapping OpenStreetMap (OSM) database to RDF, and TripleGeo supports several different geospatial databases (PostgreSQL, Oracle, MySQL, IBM DB2) , and also file formats (ESRI shapes, GML, KML).

As the need for location data within Linked Data applications has increased it has accordingly been a requirement for RDF Triple stores to support multiple geometries at Web scale. At the beginning of the GeoKnow project the Virtuoso RDF QUAD store only supported the “point” Geometry type and during the course of the project it has been enhanced to support some 14 additional Geometry types (Pointlist, Ring, Box, Linestring, Multilinestring, Polygon, Multipolygon, Collection, Curve, Closedcurve, Curvepolygon and Multicurve) and associated functions, with near full compliance with the GeoSPARQL / OGC standards now in place. Support for the GEOS (Geometry Engine - Open Source) Library has been implemented in Virtuoso, enhancing its Geospatial capabilities further.

The Virtuoso query optimiser has been enhanced to improve geospatial query performance including parallelisation of their execution⁶. In addition improvements have been made to the RDF storage optimisation reorganising physical data placement according to geospatial properties implementing a structured-aware RDF using Characteristic Sets⁷. The benefits of geospatial query optimisation and geospatial clustering for cluster deployments make use of the structure aware RDF using Characteristic Sets that was implemented and demonstrated⁸. Over the course of the project annual benchmarking of the Virtuoso QUAD store have been performed to demonstrate the improvements in the state of the art in Geospatial querying by using the GeoBench tool⁹. In the very last report¹⁰ Virtuoso was benchmarked using the newer OSM dataset. In the Description of Work, a scalability up to 25 billion triples and query times below 0.5s was envisioned, but presented results used more than 50 billion triples, and average query execution times of 0.46s for a power run, thus, results have exceeded expectations.

⁵ http://svn.aksw.org/projects/GeoKnow/Public/D2.2.1_Integration_of_Geospatial_Databases.pdf
⁷ http://svn.aksw.org/projects/GeoKnow/Public/D2.4.1_Geospatial_Clustering.pdf
⁸ http://svn.aksw.org/projects/GeoKnow/Public/D2.5.1_Distributed_Geospatial_Querying.pdf
¹⁰ http://svn.aksw.org/projects/GeoKnow/Public/D1.3.3_Continuous_Report_on_Performance_Evaluation.pdf
2.2 Spatial knowledge aggregation and fusing

GeoKnow project also aims at enriching the web of data with the geospatial dimension, so it has contributed with the development of interlinking and fusing methods adopted to spatial information. Two of the first tools achieving this are Data Extraction and Enrichment Framework (DEER), formerly known as GeoLift, and LIMES¹¹. DEER adds the spatial dimension in a dataset describing locations found in the links or in unstructured data (using a NLP component). LIMES was extended to enable linking within complex resources in geo-spatial data sets (e.g., polygons, line-strings, etc.). Furthermore, the improvement of these geo-linking tools was extended to scale using map-reduce algorithms in a cloud-based architecture¹². For evaluating these developments, the corresponding benchmarks were created (see 2.6). This experimental benchmark consisted of linking cities from DBpedia and LinkedGeoData. Initial results suggested that by using a geospatial dimension and a mean distance when linking datasets, a perfect linking accuracy could be achieved¹³. The result of this research was accredited with the Best Research Paper award at ESWC 2013. For working directly with geometries, FAGI framework¹⁴ was created for fusing different RDF representations of geometries into a consistent map. This tool receives as input two datasets and a set of links that interlink entities between the datasets and produces a new dataset where each pair of linked entities is fused into a single entity. The fusion is performed for each pair of matched properties between two linked entities, according to a selected fusion action, and considers both spatial and non-spatial properties (metadata). Fusing geospatial data may lead to a very time consuming process. Thus, improvements were proposed and implemented for optimising several processes (focusing on the minimisation of data transfer and the exploitation of graph-joining functionality) and a benchmark was designed to evaluate those improvements¹⁵. FAGI was also extended with additional functionality that support exploration, manual authoring, several options for batch fusion actions and link discovery and learning mechanisms for recommending fusion actions and annotation classes for fused entities¹⁶.

2.3 Quality Assessment

GeoKnow also worked on providing tools to improve the quality of existing datasets. The OSM community constantly contributes with enrichment and enhancement of OSM maps, and provided the needed tools for doing so. GeoKnow contributed in improving the quality of the annotations providing by the users by generating classification and clustering models in order to recommend categories for new entities inserted into OSM¹⁷. OSMRec, the tool developed for this aim, can be used for recommending OSM categories for newly created geospatial entities, based on already existing annotated entities in OSM¹⁸. Other data quality assessment on geospatial data was investigated, first by identifying the metrics that can be used to assess the data pertaining to various aspects such as coverage, surface area and structuredness. These metrics were used to evaluate community-generated datasets²⁰. The metrics outcome was used to create two software tools for assessing the quality of the datasets. CROCUS produces statistics about the data, it generates three types of Data Cubes, where the first Data Cube refers to the accuracy, second and third DataCube addresses the completeness and

¹²http://svn.aksw.org/projects/GeoKnow/Public/GeoKnow_D_3_1_2_final.pdf
¹⁶http://svn.aksw.org/projects/GeoKnow/Public/D3.2.3_Fusing_of_geospatial_relations.pdf
¹⁷http://svn.aksw.org/projects/GeoKnow/Public/D3.3.1_Prototype_for_spatial_knowledge_aggregation.pdf
¹⁸http://svn.aksw.org/projects/GeoKnow/Public/D3.3.2_Context_sensitive_spatial_knowledge_aggregation.pdf
¹⁹http://svn.aksw.org/projects/GeoKnow/Public/D3.4.1._Metrics_for_linked_geospatial_information.pdf
²⁰http://svn.aksw.org/projects/GeoKnow/Public/D3.4.2_Comparison_with_other_data_sets_Assessment.pdf
consistency of spatial data\textsuperscript{21}. And, the GeoKnow Quality Evaluator (GQE), reuses CROCUS and implements a set of geospatial data quality metrics (e.g., dataset coverage, coherence, avg. polygons/per class, etc) to compare different datasets across these metrics\textsuperscript{22}. These tools were used to evaluate three different datasets: LinkedGeoData\textsuperscript{23}, NUTS\textsuperscript{24}, GeoLinkedData\textsuperscript{25}. The results from this evaluation helped to understand the overall structure of the datasets and the variety of the data. Another data assessment tool created in GeoKnow was the RDF Data Validation Tool, which is based on integrity constraints defined by the Integrity constraints defined by the RDF Data Cube vocabulary, and is focused on statistical data\textsuperscript{26}.

2.4 Visualisation and Data Curation

The exploration and visualisation of data is a crucial task for final users. GeoKnow aimed at creating maps that are dynamically enriched and adopted to the needs of special user communities. Thus, modern software frameworks were explored to support the creation of such interfaces. GeoKnow developed reusable JavaScript libraries\textsuperscript{27} for interfacing with SPARQL endpoints\textsuperscript{28}. These libraries were used for instance in Mappify, which is a tool for easily generating and sharing maps as widgets, and Facete, which is a faceted browser for RDF spatial and non-spatial data enhanced with editing support\textsuperscript{29}. The editing capabilities consist basically in the definition of the interaction between an endpoint and the UI (Facete). The RDF Edit eXtension (REX) tool\textsuperscript{30} interface was implemented to support two kinds of data editing, one dealing with geospatial data on a map, and other for editing triples. Furthermore, Lodtenant was developed to support curating RDF data by means of workflows realised as batch processes\textsuperscript{31}. After data curation process, one may require the possibility of saving changes for using them later or propagating them to the other datasets. One of the Unister requirements consisted in the capability of managing and synchronising changes between different versions of private and public interlinked datasets\textsuperscript{32}. This requirement derived the deployment of the Co-Evolution Service component, which is a web application with a REST interface that allow managing dataset changes\textsuperscript{33}. Another visualization component was developed for visualising spatio-temporal data. Exploratory Spatio-Temporal Analysis of Linked Data ESTA-LD is a tool for spatiotemporal analysis of linked statistical data that appear at different levels of granularity. Finally, Mobile-based visualisation was also covered in GeoKnow. The GEM application allows to perform faceted browsing fully exploits the Linked Open Data paradigm\textsuperscript{34}. This tool allows browsing any number of SPARQL endpoints and filtering resources based on their type and constraints on properties, as well as leverage GPS positioning to deliver semantic routing.

\textsuperscript{21} http://svn.aksw.org/projects/GeoKnow/Public/D3.5.1_Initial_Report_On_Spatial_Data_Quality_Assessment.pdf
\textsuperscript{22} http://svn.aksw.org/projects/GeoKnow/Public/D3.5.2_Final_Report_On_Spatial_Data_Quality.pdf
\textsuperscript{23} http://linkedgeodata.org/
\textsuperscript{24} http://datahub.io/dataset/linked-nuts
\textsuperscript{25} http://geo.linkeddata.es/
\textsuperscript{26} http://svn.aksw.org/projects/GeoKnow/Public/D4.6.1_Quality_assessment_services_for_ESTA-LD.pdf
\textsuperscript{27} http://js.geoknow.eu/
\textsuperscript{28} http://svn.aksw.org/projects/GeoKnow/Public/D4.1.1_Final_release_spatial_semantic_exploration_component_public.pdf
\textsuperscript{29} http://svn.aksw.org/projects/GeoKnow/Public/D4.1.2_Final_release_spatial_semantic_exploration_component.pdf
\textsuperscript{30} http://svn.aksw.org/projects/GeoKnow/Public/D4.2.1_Spatial_authoring_widget_Set.pdf
\textsuperscript{31} http://svn.aksw.org/projects/GeoKnow/Public/D4.2.2_Spatial_curation_interface.pdf
\textsuperscript{32} http://svn.aksw.org/projects/GeoKnow/Public/D4.3.1_Concept_for_public-private_Co-Evolution.pdf
\textsuperscript{33} http://svn.aksw.org/projects/GeoKnow/Public/D4.3.2_Public-private_Co-Evolution.pdf
\textsuperscript{34} http://svn.aksw.org/projects/GeoKnow/Public/D4.5.1_Initial_release_of_the_mobile_spatial_semantic_visualization_exploration_and_authoring_tool.pdf and http://svn.aksw.org/projects/GeoKnow/Public/D4.5.2_GEM_Finak_release.pdf
2.5 Complementary Tools

Complementary activities related when working with geospatial data but not exclusively, are the collaborative tools. In GeoKnow we explored first the use of a subscription and notification services for being informed about specific data updates\(^3\), and also the integration of Social Media data as a relevant source of information for the GeoKnow use cases\(^4\).

2.6 Benchmarking Systems

The GeoBenchLab\(^5\) is a collection of benchmarking systems designed within GeoKnow project. These tools are described next.

**FacetBench** is mainly used for evaluating Virtuoso with regard to scalability, performance and other quantitative indicators can be timely accessed and demonstrated. The metrics used are query mixes per hour and average execution times for the different types of queries contained in the mix\(^6\). This benchmark was enhanced to be able to compare performance of relational spatial and RDF data management systems\(^7\).

**DataCubeValidationBench** consists of a set of SPARQL queries that are used to validate RDF Data Cubes. The purpose of this benchmark is mainly to evaluate Virtuoso with respect to performance and scalability. Performance improvements will be tracked by measuring execution times of these queries.

**DeerBench** consists of a set of DEER RDF specifications. Throughout the GeoKnow project, the execution time of executing those link specifications will be measured to track the effect of performance improvements.

**FAGI-gisBench** was created to evaluate the scalability FAGI-gis fusing tool. The benchmark was used to compare two different versions of the software with respect to the execution time of several subprocesses that take place throughout the whole fusion process.

**JassaBench** is small framework for measuring the time needed to carry out sequences of actions with the Jassa API. These actions include panning/zooming a map view and changing facet configurations.

**LinkBench** consists of a set of LIMES link specifications. Throughout the GeoKnow project, the execution time of executing those link specifications will be measured to track the effect of performance improvements.

2.7 GeoKnow Generator Workbench

The GeoKnow Generator Workbench provides an unified interface and data access to most of the tools described earlier in this section. It enables simple access and interaction with the different components needed in the LD Lifecycle. This Workbench was designed under the requirements specification of the GeoKnow use cases\(^8\).

In general, these requirements include:

- Scalability for working with large data sets

---

\(^3\) [http://svn.aksw.org/projects/GeoKnow/Public/D4.4.1%20Subscription%20and%20notification%20service.pdf](http://svn.aksw.org/projects/GeoKnow/Public/D4.4.1%20Subscription%20and%20notification%20service.pdf)


\(^5\) [https://github.com/GeoKnow/GeoBenchLab](https://github.com/GeoKnow/GeoBenchLab)


\(^7\) [http://svn.aksw.org/projects/GeoKnow/Public/D1.3.2_Continuous_Report_on_Performance_Evaluation.pdf](http://svn.aksw.org/projects/GeoKnow/Public/D1.3.2_Continuous_Report_on_Performance_Evaluation.pdf)

\(^8\) [http://svn.aksw.org/projects/GeoKnow/Confidential/D1.1.1_Initial_Common_Requirements_Specification.pdf](http://svn.aksw.org/projects/GeoKnow/Confidential/D1.1.1_Initial_Common_Requirements_Specification.pdf)
- Authentication, Authorisation and Role Management as a primary requirement in companies
- Data Provenance tracking for traceability of changes
- Job Monitoring and Robustness for applicability in production
- Modularity and Composability in order to provide flexibility w.r.t. integrating additional tools

The actual architecture of the GeoKnow Generator is presented in Figure 2. This diagram reflects the stack of tools integrated for each stage of the LD lifecycle and the Workbench. 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 standardised 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.

The Workbench is available online to test at http://generator.geoknow.eu and it was used in the GeoKnow use cases described in the following section.
3 Use Cases Achievements

This section describe the achievements concerning the two GeoKnow use cases. The e-Commerce use case lead by Unister was presented in the first and second review meeting with the EU. This use case has evolved during the third year and this report reflects the last year achievements. The Supply Chain use case is lead by Brox and this use case will be presented in the third year review meeting with the EU.

3.1 e-Commerce

In the Tourism e-Commerce use case, the initial step was to prepare internal datasets where potential interlinks can be used for answering business questions\(^4\). Therefore, the GeoKnow Generator was applied to generate an interlinked dataset used for a motive-based search infrastructure\(^5\). The generated dataset was then used to answer the following use case questions:

**How can we provide a search targeting the user’s motives? (Use Case UC-U1)** The first version of the motive-based search prototype was based on search motives manually defined on textual and factual features of a hotel\(^6\), we have extended the search infrastructure so it is much more flexible and comprehensive than what was shown in the evaluation\(^7\), with a basic prototype using drop-down boxes to select features and motives. In other words, while we were able to identify hotels suitable for children or wellness last year, this year we focused on geospatial aspects, i.e., whether hotels are in the mountains or in culturally interesting areas. Thus, we developed a semi-automatic approach for identifying geospatial motives and generating respective geometries that resemble this motive, using Voronoi tessellation (see Deliverable D6.3.3 for details). The primary advantage of this solution is its simplicity and robustness: the approach only requires a set of resources with geo-coordinates, although more complex geometries such as polygons can also be used as input. If some points in the source dataset are missing, the quality of the approach remains high. Using this approach, we are able to compute geospatial motives of a certain entity, such as a hotel, in less than 10 milliseconds. In addition to that, we implemented two custom index extension plugins for Virtuoso. The bitset index extension enables faster join processing, e.g., for searching hotels with many different features. The ElasticSearch index extension supports scalable and flexible document indexing and more comprehensive text search facilities, so far with comparable performance. The geospatial motive-based search facilities, along with a much improved user interface providing suggestions and auto-completion, have been integrated into a prototype portal BlueKiwi.de for improving hotel search.

**How can we personalise suggestions in newsletters and similar media? (Use Case UC-U2)** We have seen no progress in this use case, mainly due to providing and using personal data is difficult due to data security and privacy compliance regulations.

**How can we support strategic decisions using the interlinked data? (Use Case UC-U3)** We are currently working on an analytics application based on the motive-based search concept. First results are discussed in Deliverable D6.3.4.

**How can we improve the visualisation of our offers to customers? (Use Case UC-U4)** A potential Unister customer would like to find a holiday without having anything specific in mind. Unister would like to present him with an appealing visualisation of holiday regions in connection with Unister products,
geographic data (points of interests, mountains...), factual data and semantic facets of the selected area. Based on preprocessed data and the precomputed motives, we can apply other research on respective user interfaces in order to address this use case. This is also documented in Deliverable D6.3.4.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>OpenStreetMap / LinkedGeoData</td>
<td>General geospatial information with high coverage</td>
<td>20B</td>
</tr>
<tr>
<td>DBpedia</td>
<td>General geospatial information, typically with less coverage</td>
<td>327M (English) + 20M (German)</td>
</tr>
<tr>
<td>Geonames</td>
<td>General structured geospatial information</td>
<td>125M</td>
</tr>
</tbody>
</table>

Table 1: Integrated dataset for e-Commerce use case

The external datasets used for answering business cases are described in table 1. Figure 3 present the data preparation for the Unister use case, described as follow:

1. Extraction of Unister hotels, reviews and regions originally in a RDBMS using Sparqlify
2. Link Unister regional data with DBpedia using LIMES
3. Enrich Unister regions with DEER
4. Extraction of Unister region geometries (polygons) from RDBMS with TripleGeo
5. Link Unister regions with polygons using LIMES
6. Preprocess geospatial motives based on DBpedia, OSM and hotel information using custom process⁴⁵.
7. Search and explore information using Unister applications
8. With Facete browse motives concepts
9. Motive-based search was evaluated using Unister App (BlueKiwi)

Using this process, we were able to provide geospatial motive-based search on a Web application prototype [http://bluekiwi.de](http://bluekiwi.de). As shown in figure 6b, the user is supported entering his search query by suggesting relevant regions, features and motives. Next, he can see the results as shown in figure 6c. This enables users to find hotels by entering their ideas and desires, creating a user experience much closer to that of a traditional travel agency.

⁴⁵[http://svn.aksw.org/projects/GeoKnow/Confidential/D6.3.3_Final_Prototype_Motive_based_Search.pdf](http://svn.aksw.org/projects/GeoKnow/Confidential/D6.3.3_Final_Prototype_Motive_based_Search.pdf)
Figure 3: e-Commerce data preparation flow
(a) Motive-based search using Blue Kiwi UI

(b) Search Results

Figure 4: e-Commerce Motive-based search demonstration
3.2 Supply Chain

In order to visualise key information of the logistics in a supply chain, information from supply chain transactions have to be connected to related data. 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 suppliers, weather and news information, which is linked to the supply chain information. Specifically this use case aims to answer the following questions:

**How can we Increase transparency in the supply chain?** The main goal is to be able to have a live view on orders and shipments in the supply chain network, as well as bottlenecks in the supply chain network. For this objective, the representation in RDF of the Supply Chain concepts was developed\(^4^6\). This representation was implemented in a Supply chain Dashboard\(^4^7\) where real-time visualisation of the supply chain is possible, and it also allows the simulation of it.

**How to show / evaluate influences based on external knowledge?** This question requires the integration workflow for external structured and unstructured knowledge. Weather and news data was integrated for demonstrating this use case, and a Supply Chain Metrics (SCOR) was defined to assess performance of each supplier based on various metrics.\(^4^8\). This concept was later extended to structured and unstructured data\(^4^9\).

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Triples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weather data(^5^0)</td>
<td>NCDC worldwide daily observations of 2014</td>
<td>37 million</td>
</tr>
<tr>
<td>Weather Stations(^5^1)</td>
<td>NCDC Stations worldwide ( 100.000 Weather Stations)</td>
<td>686</td>
</tr>
<tr>
<td>(Unister) News</td>
<td>German News Article from 2009 till 2011</td>
<td>9.7 million</td>
</tr>
<tr>
<td>GADM dataset</td>
<td>Different levels of administrative areas worldwide</td>
<td>4 million</td>
</tr>
</tbody>
</table>

Table 2: Integrated dataset for Supply Chain use case

For answering questions and to perform further analytics three external datasets were integrated (see table 2. Reader may notice the use of a News dataset provided by Unister. The reason for using this dataset is because the use case required to have historical data on news, thus this data collected by Unister was useful to make the proof of concept in the Supply Chain use case. The figure 5, presents the data preparation followed to solve the use case data requirements for answering the business questions.

The data preparation for the Supply Chain use case follows next steps:

1. Import of Supplier and product data from RDF files. This data is already available in RDF form.
2. Extract Weather Stations and Observations. This weather database contains a world wide historical climate data, and can be downloaded as CSV files. To use the data:

\(^{4^7}\) [http://svn.aksw.org/projects/GeoKnow/Public/D5_3_1_Release_of_the_supply_chain_dashboard.pdf](http://svn.aksw.org/projects/GeoKnow/Public/D5_3_1_Release_of_the_supply_chain_dashboard.pdf)  
\(^{4^8}\) [http://svn.aksw.org/projects/GeoKnow/Public/D5_4_1.pdf](http://svn.aksw.org/projects/GeoKnow/Public/D5_4_1.pdf)  
Figure 5: Supply Chain data preparation flow

(a) Load the data in a database. Due to the data size it was easier to use a database than working directly with the files.

(b) Create a Sparqlify mapping and expose the data to a Sparqlify endpoint

3. Use TripleGeo to convert the GADM regions from SHP to RDF representation

4. Link the weather stations near Suppliers using LIMES. This links will be used in the metrics generation and in the Supply Chain Dashboard for showing the nearest Weather Station information.

5. Supply Chain Data simulator generates synthetic data about orders and shipments.

6. Metrics generator: calculates metrics data cubes over for the generated simulation data. During metrics calculation a link for each observation to the respective GADM administrative region is added in order to allow geospatial analytics.

7. Visualise metrics using ESTA-LD utilising the links to the administrative regions to show aggregated data for selected geospatial regions

8. RapidMiner, a Data Mining and Analytics Application, demonstrates capabilities beyond the scope of ESTA-LD like influence analysis, multi line plots, etc.
9. News extractor extracts article texts from the Unister news archive and send the article texts to FOX for NER analysis and annotation.

10. Visualise the Supply Chain Dashboard on mobile device. It enables on the go management view to the details of your supply chain.

11. Visualise the Supply Chain Dashboard. This offers an unified spatial view on the logistics in the supply chain, allows live visualisation of orders and shipments status.

Figure 6a shows a screenshot of the realised Supply Chain Dashboard, the tool for the supply chain manager that answers the business questions. It depicts the suppliers participating in the supply chain on a map widget, and it renders background news as well as metrics for a selected supplier. The complete solution for this use case is demonstrated with the integration of the Supply Chain Dashboard and Xybermotive\(^{52}\) (a commercial Web-EDI system). This integration allows:

- to extend the supply chain network visualisation with information about the specific orders and shipments coming in and going out of the connected Xybermotive systems.
- to support the analysis of the product flow, considering the senders origin and related background information like weather, disaster and political (news) conditions.
- solve customer inquiries about delivery on short notice should be answered more easily by providing insight in capacity information (available to promise - ATP) from connected Xybermotive systems in your supply chain. This helps to make automated delivery promises upon customer request.
- to enable browsing for alternate suppliers in order to remedy shortages or out of stock situations or to circumvent bad weather or disasters

\(^{52}\)http://www.xybermotive.com/
Figure 6: Supply Chain dashboard and Xybermotive widgets
4  **Sustainability Plans**

This section describes the future plans from the different partners concerning the software components developed within GeoKnow and the distribution platform, the Linked Data Stack, which has being maintained with GeoKnow resources.

4.1  **Linked Data Stack**

One of the main contributions of GeoKnow was the continuation of the Linked Data Stack (LDStack) repository. The main objective of this repository is to allow people to rapidly setup and use GeoKnow software tools.

The Linked Data Stack, called LOD2 Stack, was an outcome from LOD2 EU project, which ended in October 2013. GeoKnow team, specifically Ontos AG, took over the LD2 Stack to continue supporting this initiative. We created a website to communicate efforts and provide documentation on how to use and contribute to the stack[^3].

For making sure that the LDStack continues to function after GeoKnow we will split maintenance activities reported in D1.4.3 between InfAI and Ontos. InfAI is the biggest contributor to the stack and has experience with Debian packages. Therefore, InfAI will take over all technical maintenance of the stack and Ontos will continue working on the communication tasks. A couple of telco meetings already took place and an official hand over is planned beginning of January 2016.

4.2  **Individual Software Tools**

Some components in GeoKnow are company products such Virtuoso. The Generator Workbench has become a Ontos product too. These software tools will continue their development within the corresponding company. The rest of the software developed in GeoKnow will continue improving thanks to other research projects where some partners are involved. Some of the known projects that will for sure allow the continuation of research tools are:

- **GEISER** is a project funded by the German Ministry for Economic Affairs and Energy (BmWi) with a budget above 3 million Euro and a planned start in March 2016. GEISER aims to build intelligent data-driven geo services and runs until February 2019.

- **Big Data Europe** (BDE) is a Horizon 2020 Big Data support action, which runs until the end of 2017. It aims to build a Big Data platform for the main societal challenges in Europe as identified in the Horizon 2020 programme.

- **LEDS** is a project funded by the Federal Ministry of Education and Research (BmbF) which runs from July 2015 until June 2018.

- **ALIGNED** is a Horizon 2020 EU project on the alignment of data and software engineering which runs from February 2015 until January 2018.

- **SAKE** is a project funded by the German Ministry for Economic Affairs and Energy (BmWi) which runs from December 2014 until November 2017.

- **QAMEL** is an EuroStars project on question answering which runs from November 2015 to October 2018.

**DIESEL** is an EuroStars project for distributed Search among large enterprise data is an EuroStars project started in October 2015 to September 2018.

**HOBBIT** is a Horizon 2020 Big Data project with a budget above 4 million Euro on benchmarking large scale Linked Data, which is currently planned to run from January 2016 until the end of 2019. It aims to build a Big Linked Data platform for benchmarking various parts of Big Data value chain (as defined by the European Commission).

Table 3 list the sustainability plan for each GeoKnow component.

<table>
<thead>
<tr>
<th>Name</th>
<th>Maintainer</th>
<th>Sustainability Plan</th>
<th>Until</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Standalone applications</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Facete</td>
<td>InfAI</td>
<td>Developed and evaluated further in the project GEISER by the organisations InfAI and to a lesser degree YellowMap and Metaphacts.</td>
<td>2019</td>
</tr>
<tr>
<td>Mappify</td>
<td>InfAI</td>
<td>Developed and evaluated further in the project GEISER by the organisations InfAI and to a lesser degree YellowMap and Metaphacts.</td>
<td>2019</td>
</tr>
<tr>
<td>CubeViz</td>
<td>InfAI</td>
<td>The project will be further enhanced as core part of the LEDS funded project.</td>
<td>2018</td>
</tr>
<tr>
<td>LinkedGeoData</td>
<td>InfAI</td>
<td>Developed further by InfAI for ERTICO in the context of the Big Data Europe project. Developed further by IAIS in the BDE project. The dataset will be used for benchmarks in the HOBBIT project.</td>
<td>2017</td>
</tr>
<tr>
<td>Sparqlify</td>
<td>InfAI</td>
<td>Developed further by InfAI for ERTICO in the context of the Big Data Europe project. The component will be benchmarked and evaluated in the HOBBIT project and reused in DIESEL.</td>
<td>2017</td>
</tr>
<tr>
<td>LimesService</td>
<td>Ontos</td>
<td>This service is used by the LDIW of Ontos, as far as this product continues evolving and used by customers, this service will be also maintained.</td>
<td>Indefinitely</td>
</tr>
<tr>
<td>SAIM</td>
<td>InfAI</td>
<td>Developed further in the project LEDS as an integral part of LIMES. The branding / project name may be subject to change.</td>
<td>2018</td>
</tr>
<tr>
<td>Deer-Service</td>
<td>Ontos</td>
<td>This service is used by the LDIW of Ontos, as far as this component is required by customers this service will be maintained by Ontos.</td>
<td>Indefinitely</td>
</tr>
<tr>
<td>Virtuoso</td>
<td>OpenLink</td>
<td>Development, maintenance and support will continue for open source and commercial use as was the case before the project.</td>
<td>Indefinitely</td>
</tr>
<tr>
<td>Service</td>
<td>Organization</td>
<td>Details</td>
<td>Maintenance Period</td>
</tr>
<tr>
<td>----------------------------------</td>
<td>--------------</td>
<td>--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>TripleGeo-Service</td>
<td>Ontos</td>
<td>This service is used by the LDIW of Ontos, as far as this component is required by customers this service will be maintained by Ontos.</td>
<td>Indefinitely</td>
</tr>
<tr>
<td>FAGI-gis</td>
<td>Athena</td>
<td>FAGI-gis is an open source project, thus, the Semantic Web community is able to further develop and extend it. In parallel, the development of the tool will be continued within other projects of Athena.</td>
<td>2017</td>
</tr>
<tr>
<td>autosparql-lite</td>
<td>InfAI</td>
<td>The question answering (QA) components will be further developed in the QAMEL (the particular branding as “AutoSPARQL” may be subject to change) and benchmarked and evaluated in the HOBBIT project.</td>
<td>2018</td>
</tr>
<tr>
<td>RDFUnit</td>
<td>InfAI</td>
<td>RDFUnit is a cornerstone of the ALIGNED EU project in which major extensions of it are planned and have been done.</td>
<td>2018</td>
</tr>
<tr>
<td>rdf-data-cube-validation</td>
<td>IMP</td>
<td>IMP plans to continue cooperating with the Serbian Statistical Office and the Serbian Business Registers Agency in order to publish their statistical datasets as Linked Data. Furthermore, IMP plans to use this tool in future projects, which will result in further developments when needed and maintenance</td>
<td>2017</td>
</tr>
<tr>
<td>ESTA-LD</td>
<td>IMP</td>
<td>ESTA-LD enables exploration and analysis of linked geospatial statistical data. In cooperation with the Serbian Statistical Office, the tool will be further developed and extended with additional features for analyzing data. It is also planned to be a part of IMP’s future R&amp;D projects.</td>
<td>2017</td>
</tr>
<tr>
<td>GEM</td>
<td>IMP</td>
<td>GEM (Geospatial Exploration on the Move) is a general purpose tool for browsing linked geospatial data and navigating, and the first mobile application of its kind. Therefore, IMP plans to continue maintaining the tool and include it in future projects.</td>
<td>2017</td>
</tr>
<tr>
<td>Coevolution Service</td>
<td>Unister</td>
<td>Coevolution Service implementation will likely be continued based on requirements by Unister’s internal data integration processes. If requested by project partners, updated binaries may be provided at GitHub.</td>
<td>Indefinitely</td>
</tr>
<tr>
<td>GeoKnowGeneratorUI</td>
<td>Ontos</td>
<td>The Generator Workbench itself won’t be maintained after GeoKno, How</td>
<td>Indirect outcome</td>
</tr>
<tr>
<td>spring-batch-admin-geoknow</td>
<td>Ontos</td>
<td>This component is also used in the Ontos LDIW, so will be maintained by Ontos</td>
<td>Indefinitely</td>
</tr>
<tr>
<td>Library/Tool</td>
<td>Organization</td>
<td>Description</td>
<td>Year</td>
</tr>
<tr>
<td>------------------------------</td>
<td>--------------</td>
<td>------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>OSMRec</td>
<td>Athema</td>
<td>OSMRec is a fully functional JOSM plugin. Thus, the development and, mainly, the support for the tool will be continued within other projects of Athena.</td>
<td>2017</td>
</tr>
<tr>
<td>GeoKnow Quality Evaluator (GQE)</td>
<td>InfAI</td>
<td>GQE will be continued in ALIGNED EU project</td>
<td>2018</td>
</tr>
<tr>
<td>CROCUS</td>
<td>InfAI</td>
<td>CROCUS will be continued in ALIGNED EU project</td>
<td>2018</td>
</tr>
</tbody>
</table>

**Libraries**

<table>
<thead>
<tr>
<th>Library/Tool</th>
<th>Organization</th>
<th>Description</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEER</td>
<td>InfAI</td>
<td>Developed and evaluated further in the project GEISER by the InfAI.</td>
<td>2019</td>
</tr>
<tr>
<td>LIMES</td>
<td>InfAI</td>
<td>Developed and evaluated further in the projects Sake and GEISER by InfAI. Developed further in LEDS w.r.t. to scalability. The component will be benchmarked and evaluated in the HOBBIT project.</td>
<td>2019</td>
</tr>
<tr>
<td>JASSA</td>
<td>InfAI</td>
<td>The Jassa components are essential for the Facete application. Thus, improvements to Facete are likely to lead to updates of this library. The sustainability plan of Facete applies.</td>
<td>2019</td>
</tr>
</tbody>
</table>

**Benchmarks**

<table>
<thead>
<tr>
<th>Benchmark</th>
<th>Organization</th>
<th>Description</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>FacetBench</td>
<td>OpenLink</td>
<td>Intend to continue using and maintaining this tool as a geospatial benchmark tool for Virtuoso and other RDF Stores. May also consider making it available as git project for community engagement</td>
<td>Indefinitely</td>
</tr>
<tr>
<td>DataCubeValidation-Bench</td>
<td>IMP</td>
<td>Data Cube Validation will continue to be used and maintained by IMP. In this respect, IMP will continue to maintain the benchmark, and possibly extend it in order to compare performance of different RDF stores.</td>
<td>Indefinitely</td>
</tr>
<tr>
<td>DeerBench</td>
<td>InfAI</td>
<td>This benchmarking system will be part of the benchmarking systems that will be enhanced in HOBBIT, a Horizon 2020 Big Data project</td>
<td>2019</td>
</tr>
<tr>
<td>FAGI-gisBench</td>
<td>Athena</td>
<td>FAGI-gisBench comprises a self contained and ready to use mechanism for evaluating the performance of FAGI-gis in terms of runtimes. In case there occurs the need to update it (e.g. with different datasets) the benchmark will be enriched and extended.</td>
<td>2017</td>
</tr>
<tr>
<td>JassaBench</td>
<td>InfAI</td>
<td>JassaBench is a small tool for measuring the time needed to execute sequences of actions against the Jassa API. As part of Facete’s sustainability, JassaBench may be extended as part of addressing performance issues, such as to detect future regressions.</td>
<td>2019</td>
</tr>
<tr>
<td>LinkBench</td>
<td>InfAI</td>
<td>This benchmarking system will be part of the benchmarking systems that will be enhanced in HOBBIT, a Horizon 2020 Big Data project</td>
<td>2019</td>
</tr>
<tr>
<td>-----------</td>
<td>-----------</td>
<td>--------------------------------------------------------------------------------------------------------------------------</td>
<td>------</td>
</tr>
</tbody>
</table>

5 GeoKnow Dissemination

All the research activities and software tools have been disseminated through different channels. GeoKnow has reported every year these activities\(^5\). A summary of these is presented in table 4.

<table>
<thead>
<tr>
<th>Activity</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Released deliverables in website</td>
<td>20</td>
<td>22</td>
<td>29</td>
</tr>
<tr>
<td>Open source projects in GitHub</td>
<td>8</td>
<td>25</td>
<td>52</td>
</tr>
<tr>
<td>LDStack releases</td>
<td>8</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>Published scientific papers</td>
<td>28</td>
<td>33</td>
<td>26</td>
</tr>
<tr>
<td>Sponsored events</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Organised workshops</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Project Presentations</td>
<td>5</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Paper/Poster/Demo presentations</td>
<td>7</td>
<td>6</td>
<td>9</td>
</tr>
<tr>
<td>Blogpost</td>
<td>16</td>
<td>13</td>
<td>15</td>
</tr>
<tr>
<td>Website visitors</td>
<td>2504</td>
<td>3087</td>
<td>3902</td>
</tr>
<tr>
<td>Twitter followers</td>
<td>125</td>
<td>241</td>
<td>332</td>
</tr>
<tr>
<td>Facebook likes</td>
<td>66</td>
<td>98</td>
<td>171</td>
</tr>
</tbody>
</table>

Table 4: GeoKnow dissemination activities summary

6 Conclusions

This document summarised GeoKnow research activities together with the outcomes: software tools and benchmarks tools. These tools were used by e-Commerce and Supply Chain use cases, and their respective achievements were presented. E-Commerce use case demonstrated a motive-based search creating a different way of vacation planning approach. The Supply Chain use case demonstrated to support supply chain transparency through the Dashboard visualisation, and the management decision making support.

In this report we have also presented different sustainability plans for GeoKnow outcomes. The Linked Data Stack will be maintained by Ontos and InfAI. Some components like Virtuoso and the GeoKnow Generator Workbench will be continued as OpenLing and Ontos products respectively. The rest of the software tools and benchmarks will be continued under the umbrella of different European projects.