The Semantic Evolution of Observation Data: Visualizing spatiotemporal water data

Sylvain Bouveret*, Philippe Genoud*
Kim Hobus**, Danielle Ziebelin*

*Université Grenoble-Alpes CNRS LIG
<prenom.nom>@imag.fr
http://www.liglab.fr/fr/presentation/equipes/steamer
**Yukon Water Board
Kim.Hobus@gov.yk.ca
http://www.yukonwaterboard.ca

Résumé. The Coordinated Online Information Network (COIN) incorporates semantic web technology that integrates, publishes and visualizes time series water data allowing users to access a multitude of datasets in order to compare the data and draw conclusions. It employs an efficient and systematic data storage and retrieval system and can convert the data to a standard format for visualization. COIN utilizes a number of standards from OGC (Open Geospatial Consortium) and W3C (RDF, OWL, SPARQL and GeoSPARL) and benefits from generic ontologies transforming semantics to data, enriching data sets and making it available and interoperable via WFS and WMS standards. These principles facilitate publication and exchange of data across the web, increasing transparency and interpretability. Through modernized data submission and retrieval we hope to break down the silos of data, allowing users to visualize time series water quality and hydrometric data from multiple sources to increase knowledge in relationship to impacts on Yukon water.

1. Introduction

The study of the evolution of observation data is a central task in environmental management; it is essential for researchers to review and compare other data. Environmental management has become a multidisciplinary field in which stakeholders have to consult a multitude of different sources (water/soil chemical analyses, agricultural elements, land use etc). However, the heterogeneity of models, data, metadata and formats and their change over time, remains a major difficulty in integrating different sources. Currently, after the tremendous growth of Web 2.0, we are witnessing an evolution of the World Wide Web to what the W3C refers to as web data: a model for simple, flexible and powerful data. The Resource Description Framework (RDF) (Cyganiak et al, 2014), is based on web infrastructure and facilitates publication and exchange of data across the web. The representation models and the ontologies, expressed in RDFS (Brickley and Guha 2004) and OWL (Hitzler et al., 2012), give a semantics to data. In this paper, we present how data sources could be integrated in an RDF graph, visually presented in a time series format, and processed by associating a semantic model from ontologies, enriched by linking to other data sets.

COIN was developed in order to help Yukon Water Board officials in their evaluation of water license applications. Officials need to synthesize multiple information concerning a particular area and its immediate surrounding area. COIN allows us to understand the various steps necessary for such an approach and utilizes different technologies necessary for its implementation and according to the OGC and W3C web services description and orchestration recommendations, COIN use: XML, HTTP, WFS, WCS, WMS, and CSW. For observations and measurements: SensorML, O&M ISO 19156 were studied. For the exchange of hydrological time-series, as we did in COIN, WaterML 2.0 is the standard exchange format that we used. Our case study focused on the Yukon ecosystems: biological, chemical and physical data come either from water monitoring stations or sampling results. The purpose of these data sets is to help researchers to answer questions about water regulation, water quality, water management, and identify changes in the Yukon environment. Over the years a number of datasets have been collected, sometimes different types and forms, stored in various file formats. For easy operation, we propose a semantic web architecture that follows the Semantic Web stack [W3C 2005], implements processing functionalities in our application with queries, ontologies, inference rules and analysis services based on an RDF format and then links them to other external data sources. See Fig.1 for a visualization of this process.

2. Architecture of COIN

The basic idea of open and shared data is to establish a way to create an open and extendable infrastructure, which gives free access to information, its use and re-use. That means nonproprietary formats with low barriers and an open data license to ensure the possible re-use by anyone. It is crucial to link information and data with their context as it creates new knowledge. Tim Berners-Lee in 2010 presented his 5* model to explain the cost and the benefit of such architecture:

‘*’ Information is available on the Web (any format) under an open license

‘**’ Information is available as structured data (e.g. Excel instead of an image scan of a table)

‘***’ Non-proprietary formats are used (e.g. CSV instead of Excel)

****’ URI identification is used so that people can point at individual data

‘*****’ Data is linked to other data to provide context

W3C recommend by the W3C linked Data Cookbook the following steps to publish data:

To meet these objectives, we chose to build a triple store web application, as a web of data approach. According to the W3C, web services are self-contained and self-described application components using XML and HTTP for communication with other applications by

using open protocols. To be OGC compliant the web services need to be developed in the following way:

- for spatial data WFS for vector and discrete data,
- WMS for image and maps,
- CSW for catalogue service and metadata
- WCS for continuous and raster data access

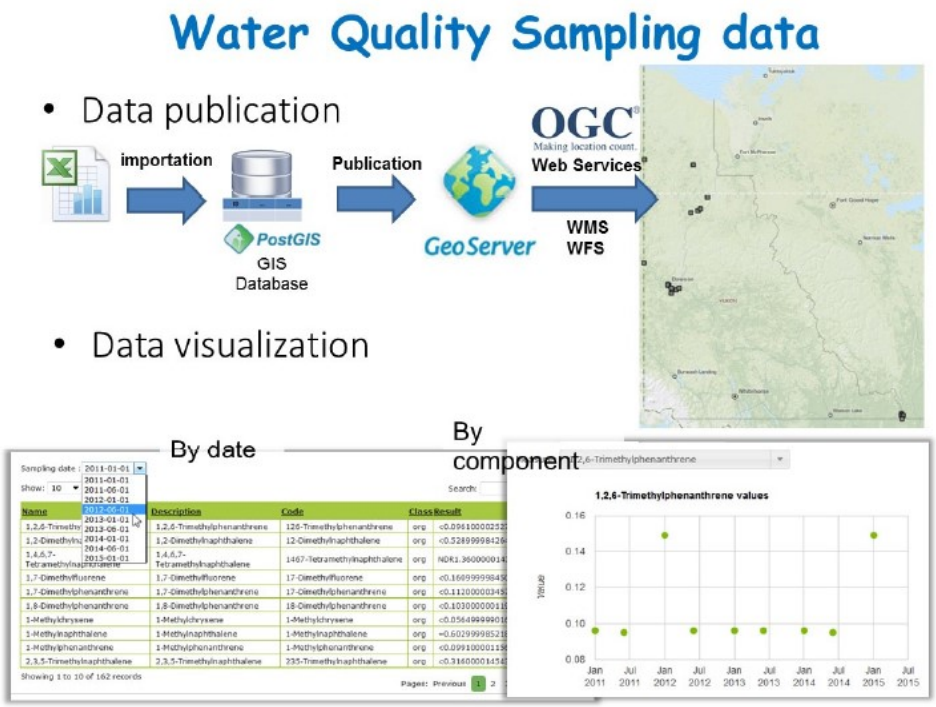


FIG. 1 - *The objective of COIN is to help the public, government, First Nation governments, and licensing officers to answer questions about water quality, water management and water allocations. To reach this objective we created an online information network with an ontology-based interface to visualize data and information source content.*

Web services are essential in the orchestration of internet-based workflows. Currently, two main architectural styles are most commonly used: SOA and REST. SOA services use remote procedure calls to invoke functions on remote systems. REST, or Representational State Transfer, is an alternative architectural model where each resource has a URI.

We have developed the following features:

- Interactive map of Yukon
- Display hydrometric data in a graph as well as allow for downloading data
- Display water quality data in a table or graph
- Search water quality data by chemical as well as by geographic area
- Set data submission standards for water quality data
- Filter map layers to anything within a geographic area

- Display and search Yukon Water Board (YWB) water license information
- Display water flow direction
- Link to Wikipedia for information regarding chemistry
- Tool to determine water allocated within a watershed.

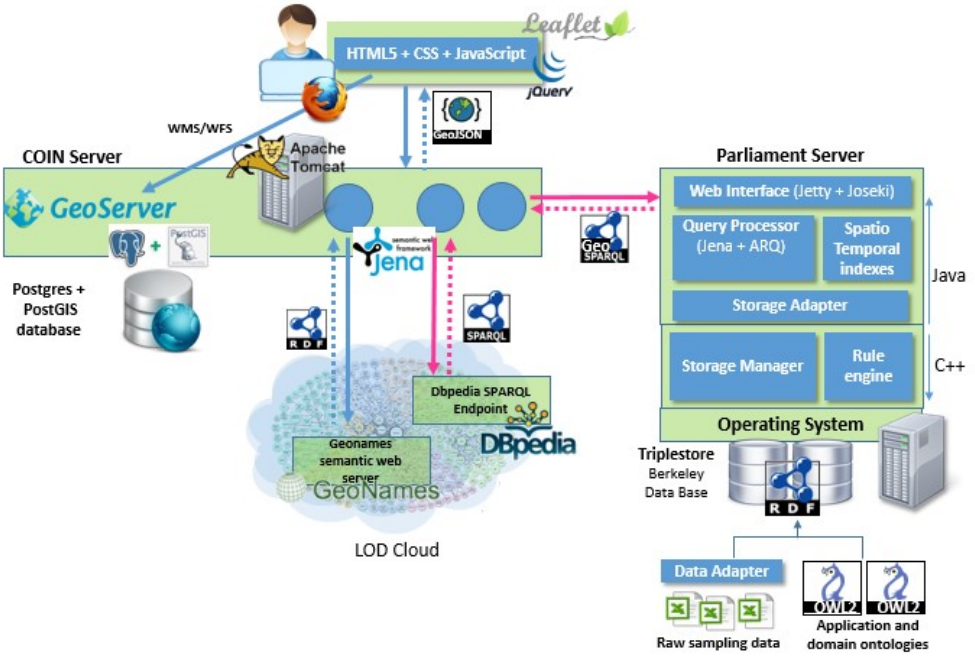


FIG. 2 – *The architecture shows how the different elements are organized. The RDF triple store for storing and integrating spatiotemporal datasets and the use of the web services provide access to the map and visualization interfaces though GIS layers. The RDF representation with the ontology is connected though data on the web.*

3. COIN application data

Numerous datasets are available from different sources in different formats ie. database, excel, and pdf. Figure 1 presents an example of sampling stations: a set of measures at different dates, in different locations and with different technology. The COIN application can then make this data accessible to users in an easy way with a graphical web user interface using maps. According to the OGC and W3C web services description and orchestration recommendations, COIN uses: XML, HTTP, WFS, WCS, WMS, and CSW as mentioned above. For observations and measurements: SWE and SOS, SensorML, O&M ISO 19156 and for exchange hydrological time-series, WaterML 2.0 is the standard exchange format. GeoSciML is a standard of data format. These are the principles that we applied to environmental data on water quality in northern Canada and which were used to study the impact on water in the Yukon Territory (Lim al., 2008).

The objective is to store different and heterogeneous types of observations (hydrometric stations, water licenses, water quality analysis etc). These data structures which store selected fields, values and tables are the key elements of the data model.

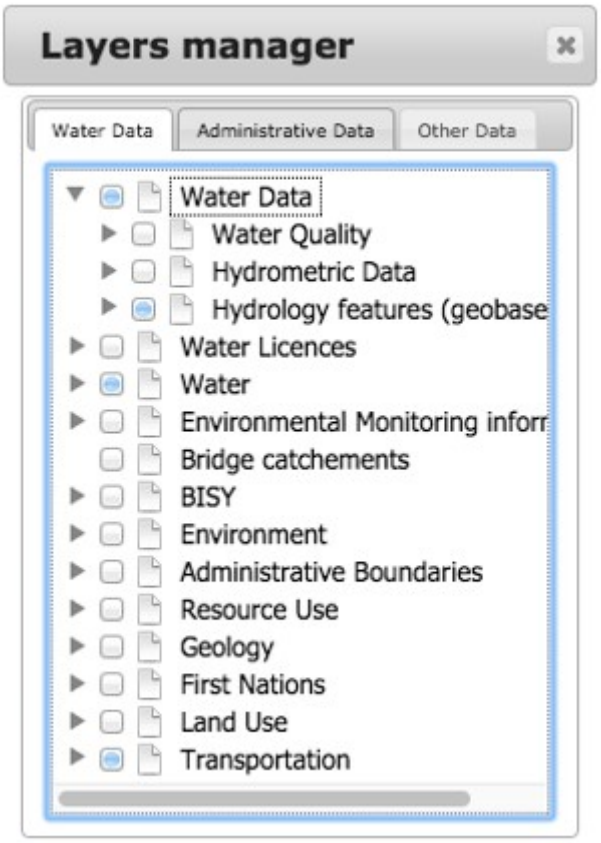


FIG. 3- *The data collected at sampling and collecting sites include: geographical description (administrative and geographic description), of station, of monitoring sites, type and size of the sampling area (ponds, lakes, springs, rivers, wetlands, watersheds), date of sampling.*

To build our model we propose a data dictionary that describes all tables and fields. The data series identify which variables have been measured at which locations and for what time stamps or periods. The data series are sets of observation values with specific properties (continuous or discontinuous measurements, units). These data series illustrate the dataflow and derivations and estimations are inferred associated to these data values. An appropriate method allows to obtain computed data values, provide logical grouping of data and make explicit the relationship between different monitoring and observation sites. It is critical that the different data sources are carefully documented and annotated with an adequate metadata structure. This structure will be used by different actors to collect data sets, and will enable easy access to the information

The aim of open and linked data is to make accessible the datatsets on a website but also to connect them to other scientific, economic, social and possibly political data sources in relationships with these elements. COIN's datasets provide different representation of sites : raw data, water quality, water microbiological information, environmental status of the site, information about agricultural and mining waste etc from different land and territory status. In order to ensure that these data sets are available to be reused and correlated to other data sets, they must be available as open and linked data. The initiative of open and linked data (Linked Open Data) (Bizer et al., 2009) follows this line, whose principles were set out by Tim Berners-Lee (Berners-Lee 2006):

(1) use of URIs (Uniform Resource Identifiers) to name (identify) things,
(2) use of HTTP URIs to consult these addresses,
(3) when a URI is accessed, provision of useful information using open standards (RDF, SPARQL, ...), and
(4) inclusion of links to other URIs in order to discover more linked data

4. Ontologies and models

To model data in the form of an RDF graph, an OWL ontology has been specifically defined for this application. This ontology includes a number of classes and properties for representing observational data. To represent spatial information (coordinates of sampling sites, geometry regions) we relied on the GeoSPARQL standard proposed by OGC (OGC, 2012) (Battle Kolas, 2012). Figure 5 shows in the form of a UML diagram the various classes and relationships (owl: ObjectProperties) defined for our application.

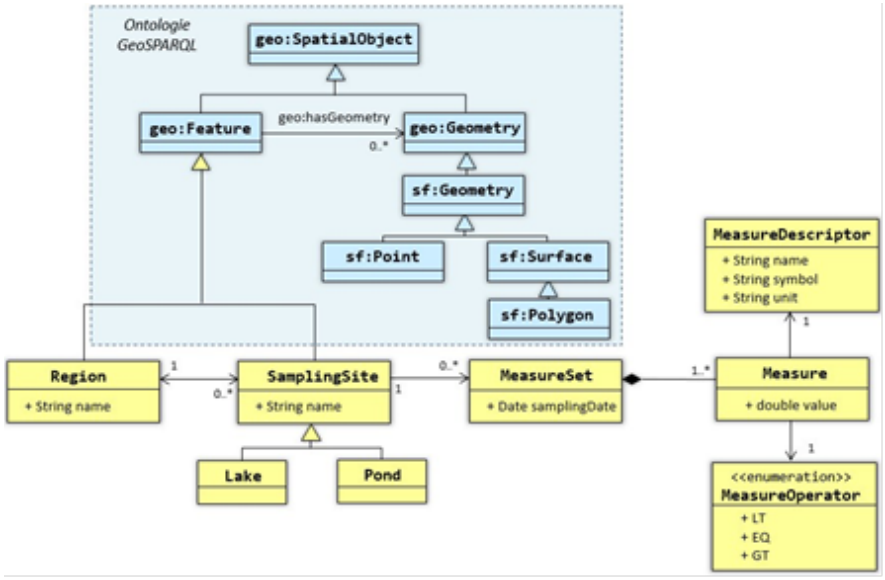


FIG. 4: *The ontology of the application and its links with GeoSPARQL ontology*

We have extended our application ontology with a general ontology from hydrology used by the Consortium of Universities for the Advancement of Hydrologic Science (CUAHSI) (Couch et al., 2014).

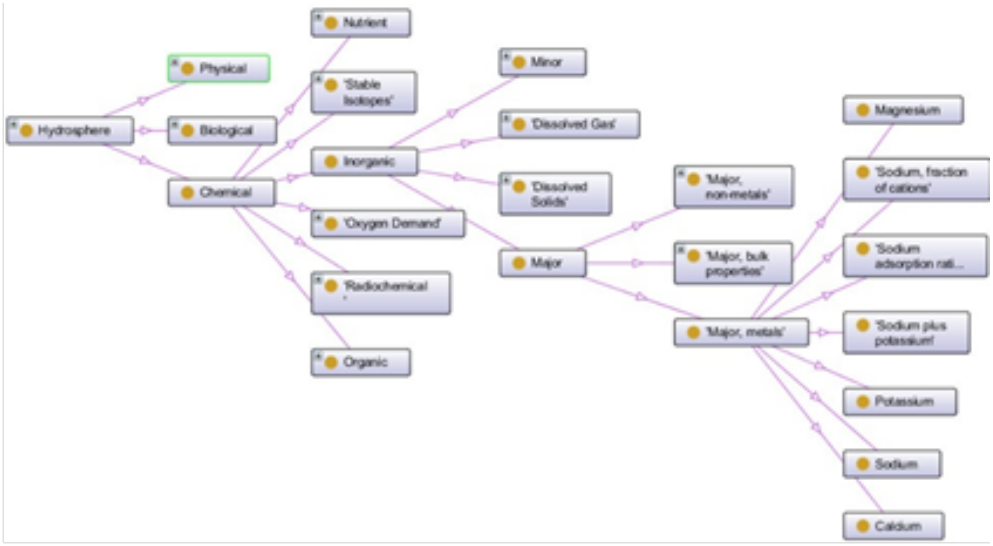


FIG. 5: *An extract of the ontology of the CUAHSI hydrosphere. Only the classes corresponding to the major metal concepts are developed in this hierarchy.*

This ontology defines a taxonomy that can hierarchically structure more than 4,000 words describing physical, chemical and biological measures related to water. It is used by the System Information CUAHSI (CUAHSI-HIS Hydrologic Information System) and consists of a set of servers and databases connected to client applications such as web services to facilitate the discovery of time series data collected at a given point. We have taken this ontology, defined in tabular form, and translated it into an OWL class hierarchy (Figure 5). The use of this ontology in our model is made by combining corresponding measures in the terminology of CUAHSI with the descriptor measures identified during sampling.

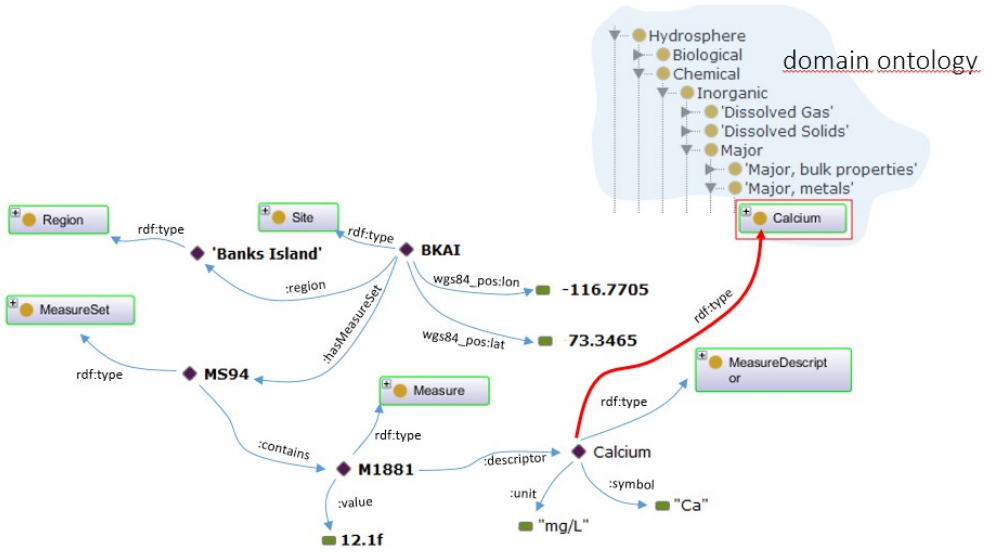


FIG. 6: *On the BKAI site at Banks Island a set of MS94 measures was taken. The M1881 12.10 measurement value is associated with a descriptor indicating that this is a measure of Calcium and its unit is mg / l. This descriptor is connected to the domain ontology. CUAHSI by a relationship rdf: type.*

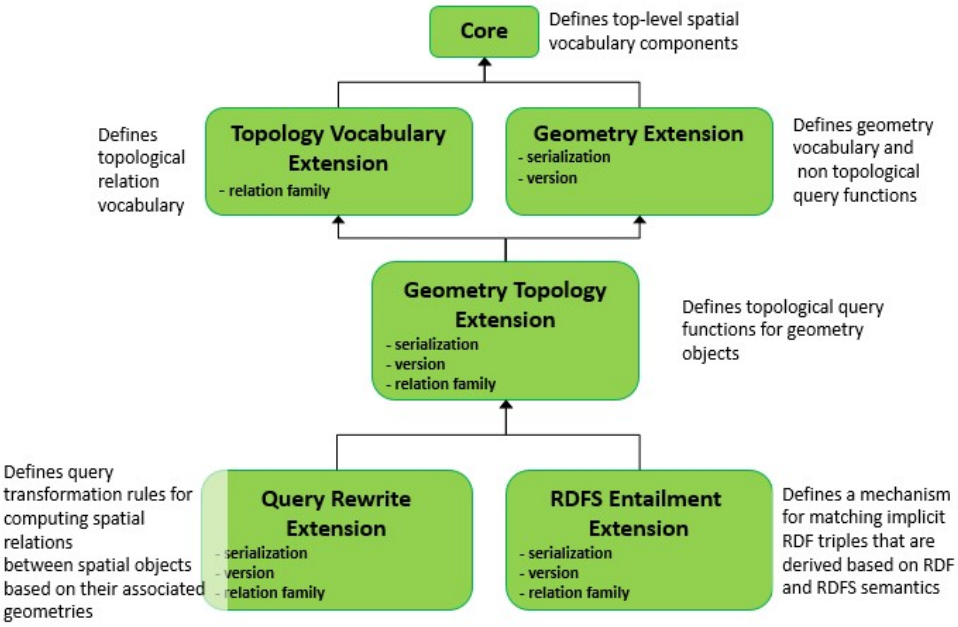
5. Exploring and analysing data

To explore and analyze the observational data thus represented, we used a web mapping interface for data queries through SPARQL and its spatial extension GeoSPARQL.

Our sampling results, stored in excel, were converted to RDF and stored in a specially designed database for storage and data recovery, called "triplestore". Although published in 2012, few triplestores currently support the standards GeoSPARQL (Athanasiou et al, 2013.) Strabo, (Kyzirakos et al, 2012.) USeekM and Parliament. Our choice fell on Parliament, which has a relatively good balance between ease of installation and use, supports GeoSPARQL and has acceptable performance (although far from the performance offered by the spatial databases (Patroumpas et al.,2014)). On the Parliament server, an RDF graph is created in which are loaded:

– Ontologies used: the domain ontology (prefix cuahsionto :), the application ontology (prefix ccionto :);

– Observational data from Excel files and converted into RDF using vocabularies defined by previous ontology and GeoSPARQL vocabulary for their spatial dimension.



When loading, Parliament automatically performs a number of inferences: RDFS inference over a number of OWL inferences (equivalent classes or properties; inverse properties: symmetrical, transitive or functional). Once loaded, the data can be queried via GeoSPARQL requests transmitted (via http) to the access point of the Parliament server (Jetty server + Joseki).

The interface of the COIN application allows the user to visualize the different sampling sites on an interactive map. The selection of a sampling site on the map provides access to the various comments made at each site. In the dialog box that appears the user has three tabs (FIG. 7) that allow the user:

- To filter the measures to display. Filtering is achieved through the hierarchy of concepts of the domain ontology. The RDFS inferences made in loading data help to automatically add a link ("rdfs: subClassOf") between each superclass of the Cuashi ontology to the descriptors measures. Only comments with a measurement concept as a descriptor are displayed, with a subclass descriptor of the selected concepts.
- To view all the comments for a given date.
- To view all the comments for a given measurement (time series).

A second enrichment of the initial data that allows their representation in the form of an RDF graph allows the ability to link with other external data sets published, respecting the principles of open and linked data. To demonstrate the potential of such enrichment we linked our data represented using our domain ontology with data from DBpedia (Lehmann et al., 2015). With the links to DBpedia , the user can access additional information that the application will look for dynamically in the web data (FIG 7). The DBpedia link allows for a given observation in Wikipedia to provide a description of the item measured.

FIG. 7: *Filtering observations using the domain concepts of hierarchy are displayed as comments regarding Major metals and pesticides*

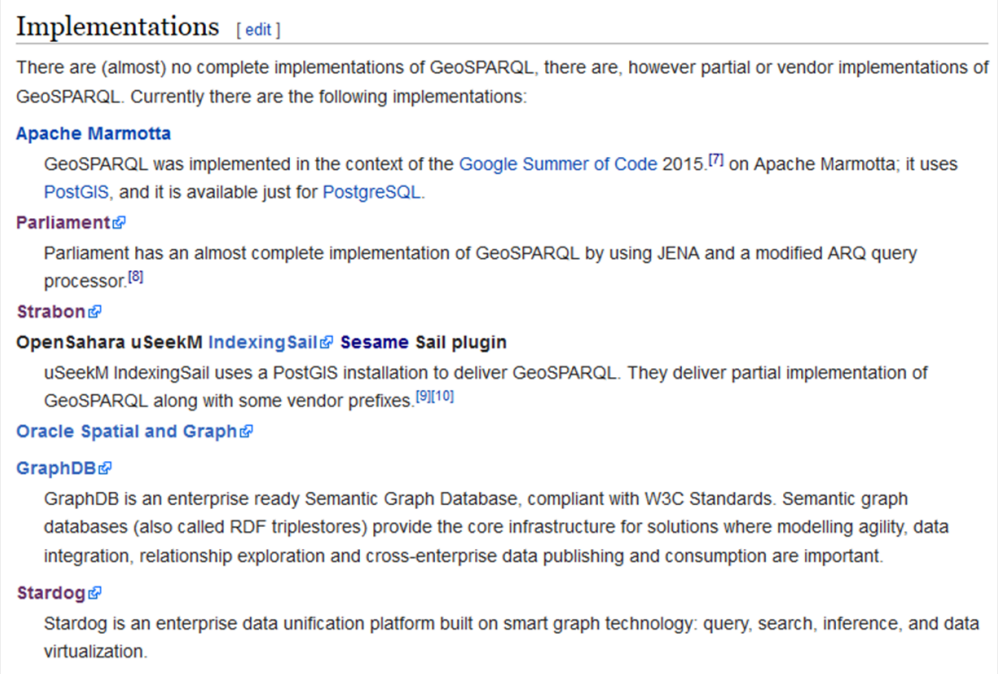
Furthermore, the use of GeoSPARQL allows for queries combining both a spatial component and a semantic component. For example, the following query "Find all the pond sites located in an area for which there is a case for a heavy metal whose value is greater than 15.2 mg / l" can be expressed using a query GeoSPARQL:

- Using the RDF Schema type inferences to select only the "Pond" type sites (Pond) (RDF triplet pattern Site rdf: type ccionto: Pond?) and have an observation corresponding to a heavy metal (triplet pattern? md rdf: type cuahsionto: C2268 where md is a measure descriptor and cuahsionto: C2268 URI concept "Major Metal" in the domain ontology (ontology CUAHSI)). The ontology includes synonyms such as “heavy metals” and “major metals”.

- GeoSPARQL uses quantitative spatial processing capabilities by selecting only sites whose geometry is inside a selected area: triplets patterns geo Site:

hasGeometry siteGeom?. and geo siteGeom: asWKT siteWKT?. possible to recover the geometry of a site in the variable WKT siteWKT; the spatial filter FILTER

(geof: sfWithin (siteWKT, "<http://www.opengis.net/def/crs/OGC/1.3/CRS84> Polygon ((- 132.35 69.74, 75.26 -132.35, -132.35? 69.74, 69.74 -132.35)) "^^ geo: wktLiteral)) allows to select only sites whose geometry is within the selected area.



Implementations [edit]

There are (almost) no complete implementations of GeoSPARQL, there are, however partial or vendor implementations of GeoSPARQL. Currently there are the following implementations:

Apache Marmotta

GeoSPARQL was implemented in the context of the Google Summer of Code 2015.[7] on Apache Marmotta; it uses PostGIS, and it is available just for PostgreSQL.

Parliament

Parliament has an almost complete implementation of GeoSPARQL by using JENA and a modified ARQ query processor.[8]

Strabon

OpenSahara uSeekM IndexingSail Sesame Sail plugin

uSeekM IndexingSail uses a PostGIS installation to deliver GeoSPARQL. They deliver partial implementation of GeoSPARQL along with some vendor prefixes.[9][10]

Oracle Spatial and Graph

GraphDB

GraphDB is an enterprise ready Semantic Graph Database, compliant with W3C Standards. Semantic graph databases (also called RDF triplestores) provide the core infrastructure for solutions where modelling agility, data integration, relationship exploration and cross-enterprise data publishing and consumption are important.

Stardog

Stardog is an enterprise data unification platform built on smart graph technology: query, search, inference, and data virtualization.

FIG. 8 : *GeoSPARQL implementation from wikipedia.*

6. Conclusion

The application described in this article addresses the hydrologic and environmental observation data to analyze and publish these data. The results of COIN's analysis enable us to understand better how the water constituents move through the watersheds. COIN software shows how semantic heterogeneity in heterogeneous data can be easily accessed and interpreted by using the web of data techniques. Beyond the analyses and the computing of data time series the system allows the user's questions to be answered by using ontologies and spatial relations. Additional linked data extend COIN's analysis to quantify human modification and land use effects on hydrologic and hydrochemical water information. The additional information included by linking data techniques increases the ability of the system to answer these types of question and greatly facilitates extensive analysis.

7. Références

Athanasiou et al, 2013 : Athanasiou S., Bezati L., Giannopoulos G., Patroumpas K., Skoutas D. (2013). Market and Reaserach Overview. GeoKnow Deliverable 2.1.1.

Berners-Lee 2010 5* model : http://5stardata.info/en/

Brickley and Guha 2004: Brickley D. Guha R. V. (2004). RDF vocabulary description language 1.0: RDF Schema. W3C Recommendation 10 February 2004. http://www.w3.org/tr/rdf-schema/.

Couch et al., 2014 : Couch A., Hooper R., Pollak J., Martin M., Seul M. (2014), Enabling Water Science at the CUAHSI Water Data Center, 7th Int'l Congress on Env. Modelling and Software, San Diego, California, USA.

Cyganiak et al, 2014 : Cyganiak B., Wood D., Lanthaler M. (2014). RDF 1.1 Concepts and Abstract Syntax. W3C Recommendation 25 February 2014. http://www.w3.org/tr/rdf11-concepts/.

Hitzler et al., 2012 : Hitzler P., Krötzsch M., Parsia B., Patel-Schneider P.F., Rudolph. S. (2012). OWL 2 Web Ontology Language Primer (Second Edition). W3C Recommendation 11 December 2012. http://www.w3.org/tr/owl2-primer

Kyzirakos et al, 2012 : K. Kyzirakos, M. Karpathiotakis, M. Koubarakis : Strabon : A Semantic Geospatial DBMS In: Proceedings of the 11th International Semantic Web Conference (2012).

Lim al., 2008 : Lim D., Smol J., Douglas M. (2008). Recent environmental changes on Banks Island (N.W.T., Canadian Arctic) quantified using fossil diatom assemblages. Journal of Paleolimnology, vol. 40, n° 1, p. 385-398.

O&M ISO 19156 : http://www.opengeospatial.org/standards/om

Patroumpas et al.,2014 : Patroumpas K., Giannopoulos G., Athanasiou S. (2014). Towards GeoSpatial semantic data management: strengths, weaknesses, and challenges a head. Proceedings of the 22nd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Dallas, Texas, USA.

SensorML : http://www.opengeospatial.org/standards/sensorml

WaterML 2.0 : http://www.opengeospatial.org/standards/waterml

W3C 2005, Semantic Web stack : https://www.w3.org/Consortium/techstack-desc.html