Numerical Groundwater Model Results As Linked Open Data

Xavier Almolda Cardona, Francisco Batlle Pifarré
Hydromodel Host, S.L.
Barcelona, Spain
xavi.almolda@hydromodelhost.com, francisco.batlle@hydromodelhost.com

***Abstract*— A service has been developed that facilitates the management of groundwater bodies (aquifers) using linked open data. Numerical models simulating the behavior of each aquifer are hosted in Cloud Services, both for data model as for the model execution itself in a flexible virtual dedicated computer. Data are periodically and automatically obtained from open databases (mainly climatic conditions data) and public agencies (water level observations and detractions) in order to update the model. An ontology has been designed to describe groundwater data coming either from measuring sites or model results. This ontology has been applied to an area of interest with an associated numerical model and transformed its results to Linked Open Data (RDF) files. This set of Linked Open Data files has been stored in a RDF store (Strabon) located in a cloud platform. Finally, a set of Web Services has been designed to query the above mentioned database as a public interface. This methodology has been applied to the Delta del Llobregat Aquifer located south of Barcelona (Spain).**

Keywords-Linked Open Data; cloud computing; groundwater; numerical models; ontology; Web Service.

I. INTRODUCTION

There is an increasing environmental awareness and quantitatively based management of natural resources, but specifically on the Water Framework Directive (WFD), which demands not only quantitative assessment of the status of water bodies, but also their expected evolution. Moreover, it requires active engagement of the public, which requires providing access to model results. The goal is to facilitate frequent updating of the model by means of integration of the large volume of observation data being collected by Water Agencies, and communication of model results. Numerical groundwater models allow assessing the current status of those bodies, their evolution under natural conditions and their sensitivity to human actions.

Many projects and initiatives are trying to maximize the exploitation of linked data. The W3C Linking Open Data Project [1] aims at making data freely available to everyone by publishing open data sets on the Web and by linking items from different data sources, and the Geospatial Semantic Web [2] goal is to use the Web as a platform for geospatial knowledge integration.

In the context of the EU Project MELODIES (Maximizing the Exploitation of Linked Open Data In Enterprise and Science) [3] a service has been developed that facilitates the management of groundwater bodies by means of numerical models using linked open data.

In this paper, a methodology is described in which the numerical models are hosted in Cloud Services that provide space and computing capabilities for executing model simulations [4]. This project is also using Strabon [5], an open-source system for storing and analyzing time-evolving geospatial Linked Data, like dynamic meshes of values that change over time.

Numerical models typically discretize spatial geometry using grids or meshes. Although some ontologies for groundwater data exist [6], an ontology from the model results has been derived since no ontologies have been found related to groundwater data derived from numerical models; and it has been complemented with an ontology for times series observations. The linked open data that results from applying this ontology has been stored in the above mentioned Strabon storage and made available via web services.

II. GROUNDWATER MODEL RESULTS DATA

The groundwater model data that will be published as Linked Open Data consist of time series, either of point values or mesh values. A mesh value is a collection of values over a space discretization, such a as a finite element mesh.

A. *Time series of mesh values*

Time series of mesh values consist of a list of mesh values for a given variable, each one with its date and time property. Each mesh value consists of a space discretization and a set of values for each node or element of this space discretization. The space discretization is defined by a set of points in space called nodes, and a set of elements that connect these nodes forming polygons. For the time series of mesh values, each new value has been stored as a mesh with its geometry and values for nodes and/or elements.

Meshes have been stored as Resource Description Framework (RDF) triples in an RDF store [7].

B. *Time series of point values*

Time series of point values consists of a list of time and value pairs, for a given location in space and a specific variable. The locations of these time series are sites where the measurements are made, like wells or meteorological stations.

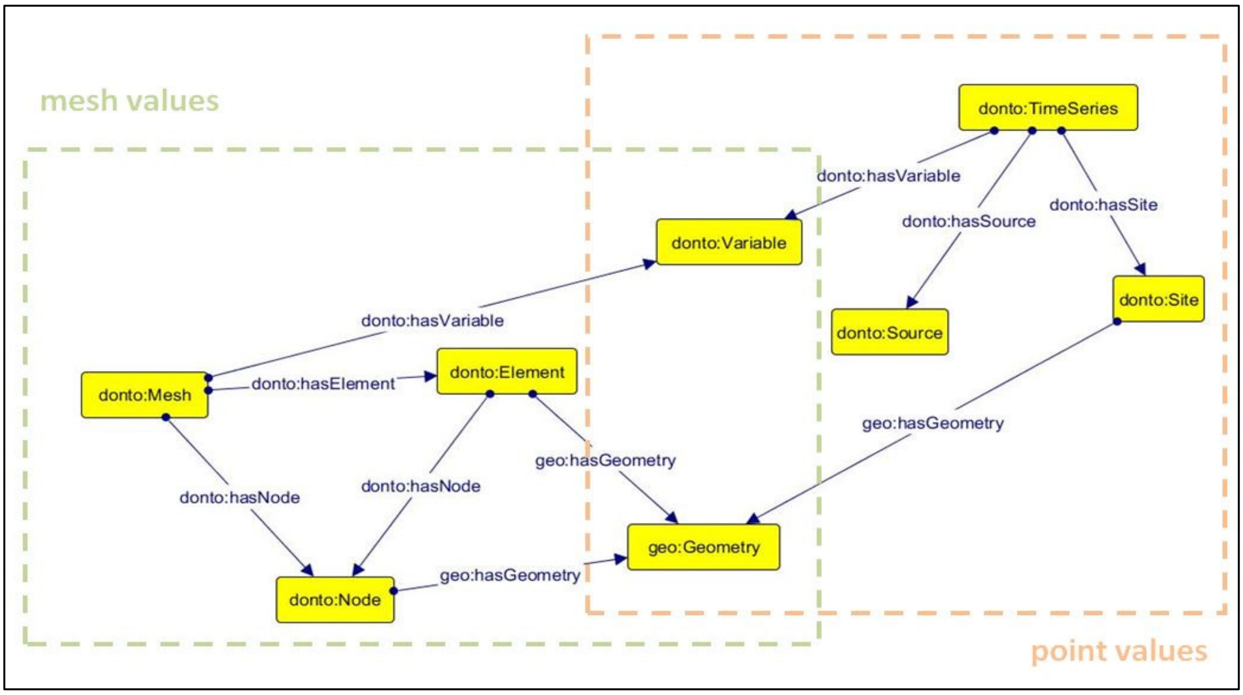

Figure 1. Groundwater results ontology diagram

Due to the large amount of time series data of point values, the individual values have been stored in an Observations Data Model (ODM) database [8], and Representational State Transfer (REST) web services have been provided to query and retrieve this data as WaterML [9].

For the purpose of linking these time series of point values with other Open Data datasets and to facilitate the discovery of new time series, a catalog of available time series for the model has been stored as RDF data in an RDF store, with start time, end time and location among other properties.

III. MODEL ONTOLOGY

In order to represent the data from the model as RDF triples, an ontology [10] has been created to describe the objects that contain this information.

Time series of values defined over a spatial discretization are defined as a list of meshes, with values on its elements or nodes, and with a *hasTimestamp* property that indicates the date and time of the simulated values.

In order to expose the time series of point values as Open Linked Data, the ontology describes the elements necessary to represent the catalog of time series available for the model. This catalog allows the user to retrieve the time series information for a given location, variable or source and use this data to retrieve the values from web services.

A diagram of the main classes of the ontology is shown in Figure 1. The elements shown have two different prefixes, the prefix *geo* for the elements of the geoSPARQL namespace, and the prefix *donto* for the elements of our ontology.

A. Mesh

A mesh is an unstructured grid that consists of nodes and elements. The elements are sets of connected nodes. The property *hasVariable* links this mesh to a Variable object that indicates the type of values for the nodes and/or elements. The data property *hasTimestamp* represents the date and time of the simulated values for this mesh.

B. Element

An element is a geometric element connecting a list of nodes. Each element has a geometry property. This geometry property is redundant because it can be obtained from the nodes of the element, but it can be useful for spatial queries involving elements.

C. Node

Each node of the unstructured mesh has a geometry property with its location of space and another property with its calculated value.

D. Variable

A variable is the type of property calculated for the nodes or elements, such as water level or chloride concentration.

The data properties for a Variable are:

- *hasName*: Name of the variable.
- *hasCode*: Code used to identify the variable.
- *hasGeneralCategory*: Category of the variable.
- *hasValueType*: An indication of whether the value represents an actual measurement, a calculated value, or is the result of a model simulation.
- *hasDataType*: An indication of the kind of quantity being measured, such as (according to WaterML

controlled vocabulary): Average, Best Easy Systematic Estimator, Categorical, Constant Over Interval, Continuous, Cumulative, Incremental, Maximum, Median, Minimum, Sporadic, Standard Deviation, Variance and Unknown.

E. TimeSeries

Each Time Series is identified by three characteristics:

- The *variable* measured.
- The *site* where the measurements were taken.
- The *source* that collected the measurements.

Each Time Series has the following data properties:

- *hasBeginDateTime*: Date and time of the first value of the series.
- *hasBeginDateTimeUTC*: Date and time of the first value of the series in UTC time.
- *hasEndDateTime*: Date and time of the last value of the series.
- *hasEndDateTimeUTC*: Date and time of the last value of the series in UTC time.
- *hasVariableUnitsName*: Name of the variable observed.
- *hasTimeUnitsName*: Name of the time units of the variable.
- *hasQCLCode*: Code that indicates the Quality Control Level of the measurements.
- *hasMethodDescription*: Description of the method used for obtaining the values.
- *hasValueCount*: Number of values in this series.
- *hasValues*: link to the web service with the values for this series.

F. Site

A site is the location in space where measurements are made. It has a geometry property indicating its geographic location, that is, a point.

It also has the following data properties:

- *hasName*: Name of the site.
- *hasCode*: Unique code that identifies the site.

G. Source

A source element contains the information on the original source of the observation.

It has the following properties:

- *hasContactName*: Name of contact, or title of organization.
- *hasDescription*: Full text description of the source of the data.
- *hasOrganization*: Name of the organization that collected the data.

IV. LINKED OPEN DATA

In order to make the groundwater model data available as linked data, it has been transformed to RDF triples and these triples have been stored in an RDF store.

The data from the time series of mesh values have been encoded as RDF triples and stored in Strabon, a RDF store with spatial and temporal capabilities that allows the use of stSPARQL [5] to query this data.

Data can be retrieved by accessing the web page of the Strabon endpoint or by making REST web service calls to the Strabon endpoint. The result can be obtained in a variety of formats, such as RDF/XML [11] or KML [12]. An example of a stSPARQL query on the Strabon endpoint web page can be seen in Figure 2.

The main purpose of linking time series data is not necessary to access each of the individual value of the time series of point data as RDF but the whole series, hence only the catalog of the available time series has been encoded as RDF data.

For each time series available, the catalog provides, among other information, the location in space (point), the variable measured, the first and last date of the values, the original source of the data, and a link to a web service with the necessary parameters to retrieve the set of values of this time series as WaterML.

A REST web service has been implemented, and is part of a family of web services that provide access to the time

```
SELECT DISTINCT ?elemWKT ?tstamp
WHERE {
?mesh rdf:type donto:Mesh;
     donto:hasElement ?elem;
     donto:hasTimestamp ?tstamp.
?node rdf:type donto:Node;
     donto:hasValue ?nodeValue.
?elem rdf:type donto:Element;
     geo:hasGeometry ?eg;
     donto:hasNode ?node.
?eg rdf:type geo:Geometry;
   geo:asWKT ?elemWKT.
FILTER ( ?nodeValue > 2000 )
}
```

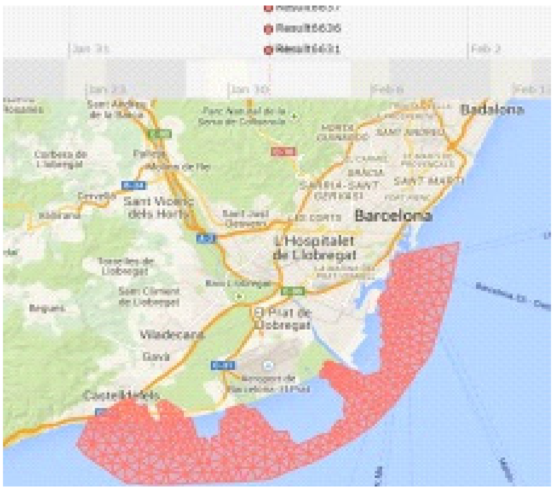

Figure 2. stSPARQL query and results (mesh elements filtered by a certain condition) on the Strabon endpoint

 ISBN: 978-1-61208-445-9

series data as WaterML. The main methods implemented are the following:

- *GetSites*(): Returns a list of sites with their basic information.
- *GetSiteInfo*(): It returns the information of one site.
- *GetVariableInfo*(): Returns information about a time series variable.
- *GetValues*(): It returns a time series for a given variable, at a given location and for a given time interval.

The diagram in Figure 3 shows the architecture of the linked data service.

V. APPLICATION

This technology has been tested and applied to the groundwater model of the Llobregat Delta Aquifer [13].

This model has a two-layered mesh of 2920 nodes and 5538 elements and it calculates flow and transport (chloride) data. The first 11 monthly values of the model simulation have been converted to RDF and uploaded to Strabon as a time series of mesh values. A catalog of more than 7000 time series of point data has also been uploaded as RDF to Strabon. The result is a dataset with more than 800,000 RDF triples for both model simulation and a time series catalog. A web portal with public access has also been created to test data [14].

VI. CONCLUSIONS

An ontology has been defined to describe the elements of groundwater model results and the time series of observations used by the model. This ontology describes how to convert this information into RDF. The use of RDF allows other services or applications to cross their data with the model results.

By choosing a spatiotemporal RDF store like Strabon, topological or temporal relations can be established to link the results from the numerical model with other existing datasets.

By publishing the catalog of time series and the links to obtain the values as Linked Open Data, the time series can be linked with other spatial or temporal information without having to store each of the individual values in the RDF store.

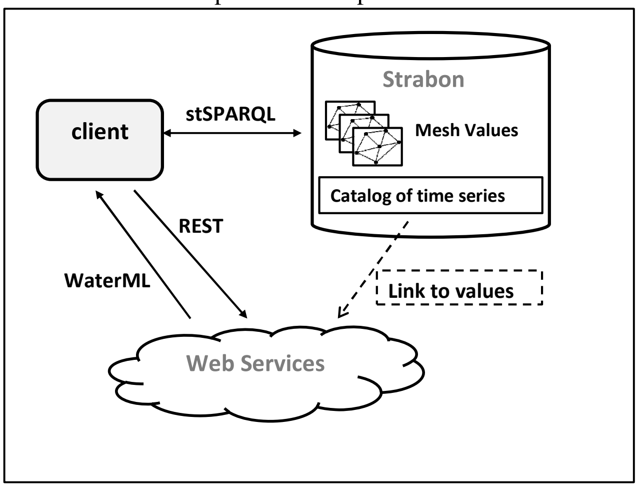

Figure 3. Linked Open Data Service architecture schema

Finally, this methodology has been tested and applied to a real model located near Barcelona (Spain).

ACKNOWLEDGMENT

This work has been funded by the FP7 project MELODIES under grant agreement number 603525.

REFERENCES

[1] C. Bizer, T. Heath, and T. Berners-Lee, “Linked Data - The Story So Far”, IJSWIS, Vol. 5, Issue 3, pp. 1-22, 2009

[2] “Geospatial Semantic Web Community Group”, Retrieved from https://www.w3.org/community/geosemweb/ [retrieved: March, 2015]

[3] J. D. Blower, D. Clifford, P. Gonçalves, and M. Koubarakis, “The MELODIES Project: Integrating diverse data using Linked Data and Cloud Computing” in ESA Big Data From Space Conference, 2014, pp. 244-247.

[4] J. Jódar, X. Almolda, F. Batlle, and J. Carrera, “Model Hosting for continuous updating and transparent Water Resources Management”, Geophysical Research Abstracts (15), EGU2013-13009-1, 2013.

[5] K. Kyzirakos, M. Karpathiotakis, and M. Koubarakis, “Strabon: A semantic geospatial DBMS”, in International Semantic Web Conference (1), 2012, pp. 295-311.

[6] B. Brodaric and T. Hahmann. “Towards a Foundational Hydro Ontology for Water Data Interoperability”, Proc. of the 11th Int. Conference on Hydroinformatics (HIC-2014).

[7] R. Cyganiak, D. Wood, and M. Lanthaler, “RDF 1.1 Concepts and Abstract Syntax”, W3C Recommendation, 25 February 2014. URL: http://www.w3.org/TR/2014/REC-rdf11-concepts-20140225/. The latest edition is available at http://www.w3.org/TR/rdf11-concepts. [retrieved: March, 2015]

[8] D. G. Tarboton, J. S. Horsburgh, and D. R. Maidment, “CUAHSI Community Observations Data Model (ODM)”, May 2008. Available from: http://his.cuahsi.org/documents/odm1.1designspecifications.pdf. [retrieved: March, 2015]

[9] D. Valentine and I. Zaslavsky, “CUAHSI WaterML 1.1. Specification. Introduciotn to WaterML Schema”, June 2009. Available from: http://his.cuahsi.org/documents/WaterML_1_1_part1_v2.docx. [retrieved: March, 2015]

[10] N. Shadbolt, W. Hall, and T. Berners-Lee, “The Semantic Web Revisited”, IEEE Intelligent Systems Journal, May/June 2006, pp. 96-101.

[11] F. Gandon and G. Schreiber, “RDF 1.1 XML Syntax”, W3C Recommendation, 25 February 2014. URL: http://www.w3.org/TR/2014/REC-rdf-syntax-grammar-20140225/. The latest published version is available at http://www.w3.org/TR/rdf-syntax-grammar. [retrieved: March, 2015]

[12] T. Wilson, “OGC KML 2.2.0”, Document #07-147r2, Open Geospatial Consortium. April 2008. Available from: http://www.opengeospatial.org/standards/kml. [retrieved: March, 2015]

[13] E. Abarca, E. Vázquez-Suñé, J. Carrera, B. Capino, D. Gámez, and F. Batlle, “Optimal design of measures to correct seawater intrusion”, Water Resources Research, Vol.42, W09415, doi:10.1029/2005WR004524, 2006.

[14] “Hydromodel Host Open Data Portal”, Available from: http://h2-lod.cloudapp.net. [retrieved: March, 2015]