Assessing Health Effects of Water Pollution Using a Semantic Water Quality Portal

Evan W. Patton[1], Ping Wang[1], Jin Guang Zheng[1], Linyun Fu[1], Timothy Lebo[1], Li Ding[1], Qing Liu[2], Joanne S. Luciano[1], and Deborah L. McGuinness[1]

[1] Tetherless World Constellation, Rensselaer Polytechnic Institute, Troy, NY USA 12180
{pattoe, wangp5, zhengj3, ful2, lebot, jluciano, dlm}@rpi.edu
2 Tasmanian ICT Centre, CSIRO, Australia
Q.Liu@csiro.au

Abstract. We demonstrate a semantically enabled approach for environmental monitoring as embodied in our semantic water quality portal. The portal assesses water quality utilizing two data sources, the United States Environmental Protection Agency (EPA) and the United States Geological Survey (USGS), by the user's choice from a number of regulations, e.g. federal level regulations established by the EPA as well as state departments of environmental protection. The portal identifies pollution events using an OWL-based reasoning system and provides browsing facets generated from provenance data encoded using the Proof Markup Language (PML). We show how exposing these measurements and their provenance as semantic data enables them to be combined with additional external data sources to look for correlations between pollution levels and health effects seen in nearby populations. This submission highlights the interactive demonstration aspects of the portal and augments the more detailed technical description of the semantic infrastructure, reasoning, and benefits of the approach that has been accepted for presentation in the Semantic Web In Use track [1].

Keywords: water quality, health effects, semantic web, data integration, scientific data visualization

1 Introduction

Environmental data are collected by a number of governmental agencies, including the Environmental Protection Agency and the United States Geological Survey at the federal level and various agencies at the state level. In addition, each organization may have the ability to establish regulations within its respective jurisdiction. We have designed a general approach to environmental monitoring using a semantic infrastructure. The Tetherless World Constellation Semantic Water Quality Portal [1] uses this approach to help a broad range of users interpret water quality data, e.g. enabling users to quickly assess what measurement sites in their communities are polluted as a function of a set of regulatory information. By default, the portal uses data from the EPA to classify bodies of water and facilities accountable under the

Clear Water Act to identify polluting facilities and water sources. Our initial demo[1] includes four regulatory agencies, viz., California, New York, Massachusetts, and Rhode Island, to highlight how different state regulations classify local water sources. This is accomplished by importing different regulatory ontologies defined using OWL 2 and classifying the instances in the datasets using whichever ontologies are active, a task that would be time consuming without the aid of semantic technologies.

2 Overview of the Semantic Water Quality Portal

The primary display for the portal is shown in Fig. 1. Top center is a field where users enter a US postal (zip) code for a particular region. To the right of the map are facets generated from PML provenance information captured during data ingest, including source agency, available regulatory information, types of water sources to display, and filters for selecting water sources based on certain characteristics or potential health effects. When a user enters a postal code, the system performs a lookup against Geonames[2] to identify the county and state containing the postal code and loads the data from a SPARQL endpoint for that county. Once the data are returned from the SPARQL endpoint, they are loaded into Jena with Pellet to perform the OWL 2 classification based on the selected facets. The details of the facets are as follows:

Data Source. Uses data points from the specified data sources, based on the user's needs or trust in an organization. Currently, the portal supports data from USGS and EPA, but is extensible to import data collected by state agencies and non-governmental organizations.

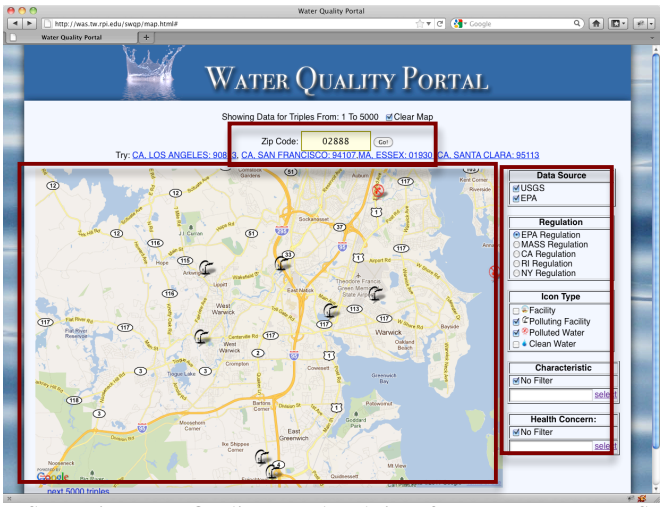

Figure 1. The Semantic Water Quality Portal web interface. Users enter a US postal code and can customize the data categorization using the facets on the right side of the display.

[1] The demonstration can be accessed at http://was.tw.rpi.edu/swqp/map.html and a video is available at http://was.tw.rpi.edu/swqp/iswc2011.mov

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

Regulation. Applies the specified regulation ontology to the data. Converters that parse regulation information off agency websites generate these ontologies, making the process more scalable compared to human entry of these values.

Icon Type. Allows the user to customize what types of sites are displayed on the map. The current four choices are Facility, Polluting Facility (i.e. a facility not in regulatory compliance), Clean Water, and Polluted Water. This list is easily extensible thus allowing new categories of pollution events to be identified and mapped.

Characteristic. Allows the user to select sites that have measurements for a certain characteristic. For example, if the user is interested in facilities with excessive concentrations of *E. Coli* bacteria, he simply checks the *E. Coli* box in the displayed form and refreshes the map using the "Go" button to obtain the new set of locations.

Health Concern. The ontology contains statements that relate water characteristics to various health effects, e.g. high levels of arsenic may cause circulatory problems. If the user would like to find locations that express high levels of contaminants related to a health issue, he selects the effect in the dialog and re-executes his query.

A primary use of the portal is viewing data relative to different regulations to compare regulations from different agencies. This gives users a powerful tool for assessing how polluted water sources are in their area relative to differing regulations. To do this, they select the source regulation using the Regulations facet and proceed to load the data. Once the data have loaded, they can switch regulations to an alternative agency and reload. This can result in previously "clean" water sources being marked as polluted, allowing users to pinpoint what areas need improvement to meet regulatory compliance. Such a tool could be useful, for example, for environmental activists looking to make an argument for stricter water standards.

3 Time series Analysis and Mashups

The semantic approach to data management enables the capability to use queried data to build visualizations. The portal includes time series visualizations to demonstrate the potential to disseminate information via graphs. When a user selects a facility, he can graph a contaminant against its regulatory limits (see Fig. 2). Color-coded data points show whether measurements violate the regulation with the limits drawn in blue. These data, along with the PML for their history, are queried via SPARQL and then rendered using Protovis[3]. The data and the ontologies can be combined with third party data. For example, we are investigating combining this time series data with county data obtained from the Center for Disease Control's Behavior Risk Factor Surveillance System Survey (BRFSS) [2], which surveys households across the United States, to identify relationships between water pollution and health conditions.

[3] http://mbostock.github.com/protovis/

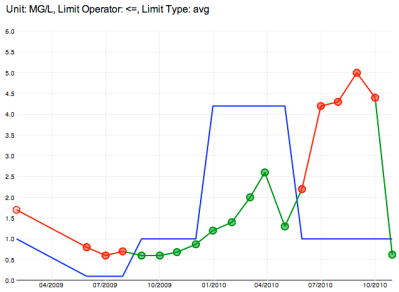

Figure 2. Example time-series view of pollution events. Limits indicated by blue line.

4 Conclusions and Future Work

We demonstrated a semantic-technology-based approach to environmental monitoring in the Tetherless World Constellation Semantic Water Quality Portal. We described the overall design and highlighted some ways that it benefits from utilizing semantic technologies, including application of multiple, interchangeable regulatory ontologies and provenance information generated during data aggregation. The portal demonstrates the benefits and potential of applying semantic web technologies to environmental information systems.

A number of extensions to this portal are ongoing. First, we intend to encode regulations for all states that differ from federal regulations. Second, data from other sources, e.g. the BRFSS mentioned herein, or weather data, may yield new ways of identifying pollution events and their relationships to public health. For example, a contaminant control strategy may fail if heavy rainfall causes flooding, carrying contaminants outside of the prescribed area. It would be possible with real-time sensor data to observe how these weather events impact the potability of water sources in the immediate area and whether this affected the population based on a rise in illness. Lastly, we would like to apply this architecture to other applications, e.g. the Clean Air Status and Trends demo[4], by enabling these applications to expose regulation and sample data using our tiered ontology.

References

1. Wang, P., Zheng, J. G., Fu, L. Y., Patton, E. W., Lebo, T., Liu, Q., Luciano, J. S., McGuinness, D. L.: TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring. In: The 10th International Semantic Web Conference, accepted (2011)
2. Centers for Disease Control and Prevention (CDC). *Behavior Risk Factor Surveillance System Survey Data*. Atlanta, Georgia: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention, 2009 http://www.cdc.gov/brfss/smart/2009.htm

[4]http://logd.tw.rpi.edu/demo/epa_aqs