A Semantic Portal for Next Generation Monitoring Systems

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

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

Abstract. We present a semantic technology-based approach to emerging monitoring systems based on our linked data approach in the Tetherless World Constellation Semantic Ecology and Environment Portal (SemantEco). Our integration scheme uses an upper level monitoring ontology and mid-level monitoring-relevant domain ontologies. The initial domain ontologies focus on water and air quality. We then integrate domain data from different authoritative sources and multiple regulation ontologies (capturing federal as well as state guidelines) to enable pollution detection and monitoring. An OWL-based reasoning scheme identifies pollution events relative to user chosen regulations. Our approach captures and leverages provenance to enable transparency. In addition, SemantEco features provenance-based facet generation, query answering, and validation over the integrated data via SPARQL. We introduce the general SemantEco approach, describe the implementation which has been built out substantially in the water domain creating the SemantAqua portal, and highlight some of the potential impacts for the future of semantically-enabled monitoring systems.

Keywords: Environmental Portal, Provenance-Aware Search, Water Quality Monitoring, Pragmatic Considerations for Semantic Environmental Monitoring

1 Introduction

Concerns over ecological and environmental issues such as biodiversity loss [1], water problems [14], atmospheric pollution [8], and sustainable development [10] have highlighted the need for reliable information systems to support monitoring of ecological and environmental trends, support scientific research and inform citizens. In particular, semantic technologies have been used in environmental monitoring information systems to facilitate domain knowledge integration across multiple sources and support collaborative scientific workflows [17]. Meanwhile, demand has increased for direct and transparent access to ecological and environmental information. For example, after a recent water quality episode in Bristol County, Rhode Island where *E. coli* was reported in the water, residents requested information

concerning when the contamination began, how it happened, and what measures were being taken to monitor and prevent future occurrences.[1]

In this paper, we describe a semantic technology-based approach to ecological and environmental monitoring. We deployed the approach in the Tetherless World Constellation's Semantic Ecology and Environment Portal (SemantEco). SemantEco is an exemplar next generation monitoring portal that provides investigation support for lay people as well as experts while also providing a real world ecological and environmental evaluation testbed for our linked data approach. The portal integrates environmental monitoring and regulation data from multiple sources following Linked Data principles, captures the semantics of domain knowledge using a family of modular simple OWL2 [7] ontologies, preserves provenance metadata using the Proof Markup Language (PML) [11], and infers environment pollution events using OWL2 inference. The web portal delivers environmental information and reasoning results to citizens via a faceted browsing map interface[2].

The contributions of this work are multi-faceted. The overall design provides an operational specification model that may be used for creating ecological and environmental monitoring portals. It includes a simple upper ontology and initial domain ontologies for water and air. We have used this design to develop a water quality portal (SemantAqua) that allows anyone, including those lacking in-depth knowledge of water pollution regulations or water data sources, to explore and monitor water quality in the United States. It is being tested by being used to do a redesign of our air quality portal[3]. It also exposes potential directions for monitoring systems as they may empower citizen scientists and enable dialogue between concerned citizens and professionals. These systems, for example, may be used to integrate data generated by citizen scientists as potential indicators that professional collection and evaluation may be needed in particular areas. Additionally subject matter professionals can use this system to conduct provenance-aware analysis, such as explaining the cause of a water problem and cross-validating water quality data from different data sources with similar contextual provenance parameters (e.g. time and location).

In this paper, section 2 reviews selected challenges in the implementation of the SemantEco design in the SemantAqua portal on real-world data. Section 3 elaborates how semantic web technologies have been used in the portal, including ontology-based domain knowledge modeling, real-world water quality data integration, and provenance-aware computing. Section 4 describes implementation details and section 5 discusses impacts and several highlights. Related work is reviewed in section 6 and section 7 describes future directions.

[1] Morgan, T. J. 2009. "Bristol, Warren, Barrington residents told to boil water" Providence Journal, September 8, 2009. http://newsblog.projo.com/2009/09/residents-of-3.html

[2] http://was.tw.rpi.edu/swqp/map.html

[3] http://logd.tw.rpi.edu/demo/clean_air_status_and_trends_-_ozone

2 Ecological and Environmental Information Systems Challenges

SemantEco provides an extensible upper ontology for monitoring with an initial focus on supporting environmental pollution monitoring with connections to health impacts. Our initial domain area for an in depth dive was water quality. The resulting portal is a publicly accessible semantically-enabled water information system that facilitates discovery of polluted water, polluting facilities, specific contaminants, and health impacts. We are in the process of extending it to include air quality data as well as industrial connections to the operating entities of polluting facilities. We faced a number of challenges during implementation, which we will now discuss.

2.1 Modeling Domain Knowledge for Environmental Monitoring

Environmental monitoring systems must model at least three types of domain knowledge: background environmental knowledge (e.g., water-relevant contaminants, bodies of water), observational data items (e.g., the amount of arsenic in water) collected by sensors and humans, and (preferably authoritative) environmental regulations (e.g., safe drinking water levels for known contaminants). An interoperable model is needed to represent the diverse collection of regulations, observational data, and environmental knowledge from various sources.

Observational data include measurements of environmental characteristics together with corresponding metadata, e.g. the type and unit of the data item, as well as provenance metadata such as sensor locations, observation times, and optionally test methods and devices used to generate the observation. A light-weight extensible domain ontology is ideal to enable reasoning on observational data while limiting ontology development and understanding costs.

A number of ontologies have been developed for modeling environmental domains. Raskin et al. [13] propose the SWEET ontology family for Earth system science. Chen et al. [5] models relationships among water quality datasets. Chau et al. [4] models a specific aspect of water quality. While these ontologies provide support to encode the first two types of domain knowledge, they do not support modeling environmental regulations.

Environmental regulations describe contaminants and their allowable thresholds, e.g. "the Maximum Contaminant Level (MCL) for Arsenic is 0.01 mg/L" according to the National Primary Drinking Water Regulations (NPDWRs)[4] stipulated by the US Environmental Protection Agency (EPA). Water regulations are established both at the federal level and by different state agencies. For instance, the threshold for Antimony is 0.0056 mg/L according to the Rhode Island Department of Environmental Management's Water Quality Regulations[5] while the threshold for Antimony is 0.006 mg/L according to the Drinking Water Protection Program[6] from the New York Department of Health. To capture the diversity of the water regulations,

[4] http://water.epa.gov/drink/contaminants/index.cfm

[5] http://www.dem.ri.gov/pubs/regs/regs/water/h20q09.pdf

[6] http://www.health.ny.gov/environmental/water/drinking/part5/tables.htm

we generated a comparison table[7] (including provenance) of different contaminant thresholds at federal and state levels.

2.2 Collecting Environmental Data

Environmental information systems need to integrate data from distributed data sources to enrich the source data and provide data validation. For water quality monitoring, two major U.S. government agencies publish water quality data: the Environmental Protection Agency (EPA) and US Geological Survey (USGS). Both release observational data based on their own independent water quality monitoring systems. Permit compliance and enforcement status of facilities is regulated by the National Pollutant Discharge Elimination System (NPDES[8]) under the Clean Water Act (CWA). The NPDES datasets contain descriptions of the facilities (e.g. name, permit number, and location) and measurements of water contaminants discharged by the facilities for up to five test types per contaminant. USGS publishes data about water monitoring sites and measurements from water samples through the National Water Information System (NWIS)[9].

Although environmental datasets are often organized as data tables, it is not easy to integrate them due to syntactic and semantic differences. In particular, we observe multiple needs for linking data: (i) the same concept may be named differently, e.g., the notion "name of contaminant" is represented by "CharacteristicName" in USGS datasets and "Name" in EPA datasets, (ii) some popular concepts, e.g. name of chemical, may be used in domains other than water quality monitoring, so it would be useful to link to other accepted models, e.g. the ChemML chemical element descriptions and (iii) most observational data are complex data objects. For example, Table 1 shows a fragment from EPA's measurement dataset, where four table cells in the first two columns together yield a complex data object: "C1" refers to one type of water contamination test, "C1_VALUE" and "C1_UNIT" indicate two different attributes for interpreting the cells under them respectively, and the data object reads "the measured concentration of fecal coliform is 34.07 MPN/100mL under test option C1". Effective mechanisms are needed to allow connection of relevant data objects (e.g., the density observations of fecal coliform observed in EPA and USGS datasets) to enable cross-dataset comparisons.

Table 1. For the facility with permit RI0100005, the 469th row for Coliform_fecal_general measurements on 09/30/2010 contains 2 tests.

C1_VALUE	C1_UNIT	C2_VALUE	C2_UNIT
34.07	MPN/100ML	53.83	MPN/100ML

[7] http://tw.rpi.edu/web/project/TWC-SWQP/compare_five_regulation

[8] http://www.epa-echo.gov/echo/compliance_report_water_icp.htm

[9] http://waterdata.usgs.gov/nwis

2.3 Provenance Tracking and Provenance-Aware Computing

In order to enable transparency and encourage community participation, a public information system should track provenance metadata during data processing and leverage provenance metadata in its computational services. Similarly, an environmental monitoring system that combines data from different sources should maintain and expose data sources on demand. This enables data curators to get credit for their contributions and also allows users to choose data from trusted sources. The data sources are automatically refreshed from the corresponding provenance metadata when the system ingests new data.

Provenance metadata can maintain context information (e.g. when and where an observation was collected), which can be used to determine whether two data objects are comparable. For example, when pH measurements from EPA and USGS are validated, the measurement provenance should be checked: the latitude and longitude of the EPA and USGS sites where the pH values are measured should be very close, the measurement time should be in the same year and month, etc.

3 Semantic Web Approach

We believe that a semantic web approach is well suited to the general problem of monitoring, and explore this approach with a water quality monitoring portal at scale.

3.1 Domain Knowledge Modeling and Reasoning

We use an ontology-based approach to model domain knowledge in environmental information systems. An upper ontology[10] defines the basic terms for environmental monitoring. Domain ontologies extend the upper ontology to model domain specific terms. We also develop regulation ontologies[11] that include terms required for describing compliance and pollution levels. These ontologies leverage OWL inference to reason about the compliance of observations with regulations.

Upper Ontology Design

Existing ontologies do not completely cover all the necessary domain concepts as mentioned in section 2.1. We provide an upper ontology that reuses and is complementary to existing ontologies (e.g. SWEET, FOAF). The ontology models domain objects (e.g. polluted sites) as classes and their relationships (e.g. hasMeasurement, hasCharacteristic[12]) as properties. A subset of the ontology is illustrated in Figure 1. A polluted site is modeled as something that is both a

[10] http://escience.rpi.edu/ontology/semanteco/2/0/pollution.owl#

[11] e.g., http://purl.org/twc/ontology/swqp/region/ny; others are listed at http://purl.org/twc/ontology/swqp/region/

[12] Our ontology uses characteristic instead of contaminant based on the consideration that some characteristics measured like pH and temperature are not contaminants.

measurement site and polluted thing, which is something that has at least one measurement that violates a regulation.

This ontology can be extended to different domains by adding domain-specific classes. For example, water measurement is a subclass of measurement, and water site is the intersection of body of water, measurement site and something that has at least one water measurement[13]. Our water quality extension is also shown in Figure 1.

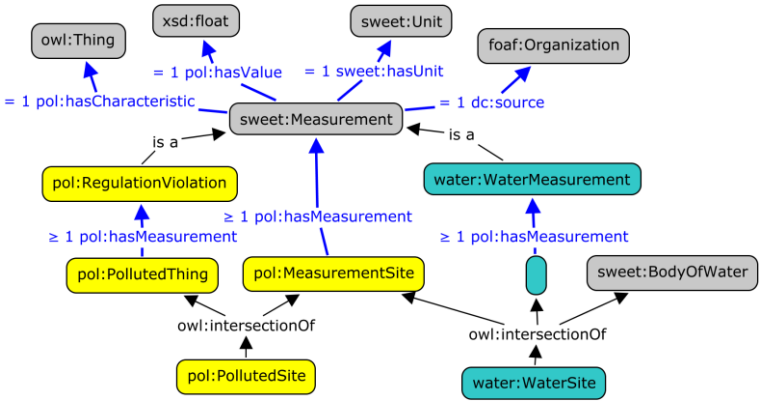

Fig. 1. Portion of the TWC Environment Monitoring Ontology.

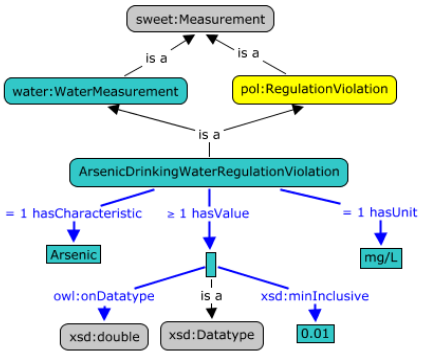

Fig. 2. Portion of the EPA Regulation Ontology.

Regulation Ontology Design

Each domain must define its own regulation ontology that maps the rules in regulations to OWL classes. For water quality monitoring, we developed a regulation ontology in which the allowable ranges of regulated characteristics are encoded via numeric range restrictions on datatype properties. The rule-compliance results are reflected by whether an observational data item is a member of the class mapped from the rule. Figure 2 illustrates the OWL representation of one rule from EPA's NPDWRs, i.e. drinking water is polluted if the concentration of Arsenic is more than

[13] http://escience.rpi.edu/ontology/2/0/water.owl#

0.01 mg/L. In the regulation ontology, ArsenicDrinkingWaterRegulationViolation is a water measurement with value greater than or equal to 0.01 mg/L of the Arsenic characteristic. Regulations in other environment domains can be similarly mapped to OWL2 restrictions if they represent violations as ranges of measured characteristics.

Reasoning Domain Data with Regulations

Combining observational data items collected at water-monitoring sites and the domain and regulation ontologies, an OWL2 reasoner can decide if any sites are polluted. This design provides several benefits. First, the upper ontology is small and easy to maintain; it consists of only 7 classes, 4 object properties, and 10 data properties. Secondly, the ontology design is extensible. The upper ontology can be extended to other domains, e.g. air quality monitoring[14]. Regulation ontologies can be extended to incorporate more regulations as needed. We wrote converters to extract federal and four states' regulation data from web pages and translated them into OWL2 [7] constraints that align with the upper ontology. The same workflow can be used to obtain the remaining state regulations using either our existing converters or potentially new converters if the data are in different forms. The design leads to flexible querying and reasoning: the user can select the regulations to apply to the data and the reasoner will classify using only the ontology for the selected regulations. For example, when Rhode Island regulations are applied to water quality data for zip code 02888 (Warwick, RI), the portal detects 2 polluted water sites and 7 polluting facilities. If the user chooses to apply California regulations to the same region, the portal identifies 15 polluted water sites, including the 2 detected with Rhode Island regulations, and the same 7 polluting facilities. One conclusion is that California regulations are stricter than Rhode Island's (and many other states), and the difference could be of interest to environmental researchers and local residents.

3.2 Data Integration

When integrating real world data from multiple sources, monitoring systems can benefit from adopting the data conversion and organization capabilities enabled by the TWC-LOGD portal [6]. The open source tool **csv2rdf4lod**[15] can be used to convert datasets from heterogeneous sources into Linked Data [9].

Linking to ontological terms: Datasets from different sources can be linked if they reuse common ontological terms, i.e. classes and properties. For instance, we map the property "CharacteristicName" in the USGS dataset and the property "Name" in the EPA dataset to a common property water:hasCharacteristic. Similarly, we map spatial location to an external ontology, i.e. wgs84[16]:lat and wgs84:long.

Aligning instance references: We promote references to characteristic names from literal to URI, e.g. "Arsenic" is promoted to "water:Arsenic", which then can be linked to external resources like "dbpedia:Arsenic" using owl:sameAs. This design is based on the observation that not all instance names can be directly mapped to

[14] http://escience.rpi.edu/ontology/2/0/air.owl#

[15] http://purl.org/twc/id/software/csv2rdf4lod

[16] http://www.w3.org/2003/01/geo/wgs84_pos

DBpedia URIs (e.g., "Nitrate/Nitrite" from the Massachusetts water regulations[17] maps two DBpedia URIs), and some instances may not be defined in DBpedia (e.g., "C5-C8" from the Massachusetts water regulations). By linking to DBpedia URIs, we reserve the opportunity to connect to other knowledge base, e.g. disease database.

Converting complex objects: As discussed in section 2.2, we often need to compose a complex data object from multiple cells in a table. We use the cell-based conversion capability provided by csv2rdf4lod to enhance EPA data by marking each cell value as a subject in a triple and bundling the related cell values with the marked subject. The details can be found in [18].

3.3 Provenance Tracking and Provenance-Aware Computing

SemantEco provenance data come from two sources: (i) provenance metadata embedded in the original datasets, e.g. measurement location and time; (ii) metadata that describe the derivation history of the data. We automatically capture provenance data during the data integration stages and encode them in PML 2 [11] due to the provenance support from csv2rdf4lod. At the retrieval stage, we capture provenance, e.g. data source URL, time, method, and protocol used in data retrieval. We maintain provenance at the conversion stage, e.g. engine performing the conversion, antecedent data, and roles played by those data. At the publication stage, we capture provenance, e.g. agent, time, and context for triple store loads and updates. When we convert the regulations, we capture their provenance programmatically. We reveal these provenance data via pop up dialogs when the user selects a measurement site or facility, and utilize them to enable new applications like dynamic data source (DS) listings and provenance-aware cross validation.

Data Source as Provenance

We utilize data source provenance to support dynamic data source listing as follows:

1. Newly gathered water quality data are loaded into the system as RDF graphs.
2. When new graphs come, the system generates an RDF graph, namely the DS graph, to record the metadata of all the RDF graphs in the system. The DS graph contains information such as the URI, classification and ranking of each RDF graph.
3. The system tells the user what data sources are currently available by executing a SPARQL query on the DS graph to select distinct data source URIs.
4. With the presentation of the data sources on the interface, the user is allowed to select the data sources he/she trusts (see Figure 4). The system would then only return results within the selected sources.

Provenance information can allow the user to customize his/her data retrieval request, e.g. some users may be only interested in data published within a particular time period. The SPARQL queries used in each step are available at [18].

Provenance-Aware Cross-Validation over EPA and USGS Data

[17] The "2011 Standards & Guidelines for Contaminants in Massachusetts Drinking Water" at http://www.mass.gov/dep/water/drinking/standards/dwstand.htm

Provenance enables our system to compare and cross-validate water quality data originating from different source agencies. Figure 3 shows pH measurements collected at an EPA facility (at 41:59:37N, 71:34:27W) and a USGS site (at 41:59:47N, 71:33:45W) that are less than 1km apart. Note that the pH values measured by USGS fell below the minimum value from EPA often and went above the maximum value from EPA once. We found two nearby locations using a SPARQL filter:

```
FILTER ( ?facLat < (?siteLat+"+delta+")
&& ?facLat > (?siteLat-"+delta+")
&& ?facLong < (?siteLong+"+delta+")
&& ?facLong > (?siteLong-"+delta+"))
```

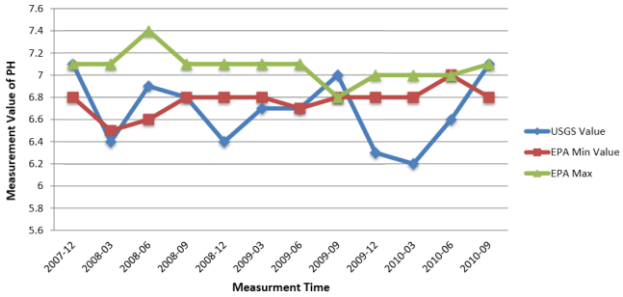

Fig. 3. Data Validation Example

4 SemantAqua: Semantic Water Quality Portal

4.1 System Implementation

Figure 4 shows one example where the semantic water quality portal supports water pollution identification. The user specifies a geographic region of interest by entering a zip code (mark 1). Users can customize queries from multiple facets: data source (mark 3), water regulations (mark 4), water characteristic (mark 6) and health concern (mark 7). After the portal generates the results, it visualizes the results on a Google map using different icons to distinguish between clean and polluted water sources and facilities (mark 5). The user can access more details about a site by clicking on its icon. The information provided in the pop up window (mark 2) include: names of contaminants, measured values, limit values, and time of measurement. The window also provides a link that displays the water quality data as a time series.

The portal retrieves water quality datasets from EPA and USGS and converts the heterogeneous datasets into RDF using csv2rdf4lod. The converted water quality data are loaded into OpenLink Virtuoso 6 open-source edition [18] and retrieved via SPARQL queries. The portal utilizes the Pellet OWL Reasoner [16] together with the Jena Semantic Web Framework [2] to reason over the water quality data and water ontologies in order to identify water pollution events.

[18] http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/

The portal models the effective dates of the regulations, but only at the granularity of a set of regulations rather than per contaminant. We use provenance data to generate and maintain the data source facet (mark 3), enabling the user to choose data sources he/she trusts.

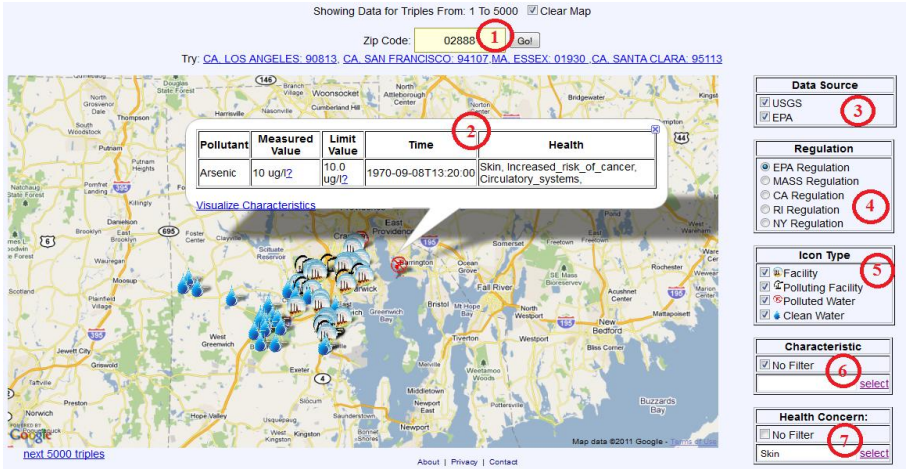

Fig. 4. Water Quality Portal In Action

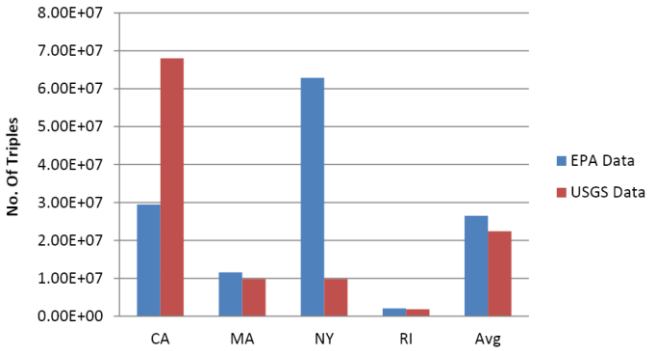

Fig. 5. Triple numbers for our initial four states with average computation.

4.2 Scaling Issues

We wanted to test our approach in a realistic setting so we gathered data for an initial set of states to determine scaling issues. We have generated 89.58 million triples for the USGS datasets and 105.99 million triples for the EPA datasets for 4 states, which implies that water data for all 50 states would generate at least a few billion triples. The sizes of the available datasets are summarized in Figure 5. Such size suggests that a triple store cluster should be deployed to host the water data.[19]

[19] We have recently obtained the data for the remaining 46 states and are working on the completed US portal.

Table 2 includes class counts for our initial four state regulations. Our programmed conversion provides a quick and low cost approach for encoding regulations. It took us about 2 person-days to encode hundreds of rules.

Table 2. Number of classes converted from regulations

EPA	CA	MA	NY	RI
83	104	139	74	100

5 Discussion

5.1 Linking to a Health Domain

Polluted drinking water can cause acute diseases, such as diarrhea, and chronic health effects such as cancer, liver and kidney damage. For example, water pollution co-occurring with new types of natural gas extraction in Bradford County, Pennsylvania has been reported to generate numerous problems[20, 21]. People reported symptoms ranging from rashes to numbness, tingling, and chemical burn sensations, escalating to more severe symptoms including racing heart and muscle tremors.

In order to help citizens investigate health impacts of water pollution, we are extending our ontologies to include potential health impacts of overexposure to contaminants. These relationships are quite diverse since potential health impacts vary widely. For example, according to NPDWRs, excessive exposure to lead may cause kidney problems and high blood pressure in adults whereas infants and children may experience delays in physical or mental development.

Similar to modeling water regulations, we programmatically extracted the relationships between contaminants and health impacts from a web page[22] and encoded them into OWL classes. We used the object property "hasSymptom" to connect the classes with their symptoms, e.g. health:high_blood_pressure. The classes of health effects are related to the classes of violations, e.g. LeadDrinkingWaterRegulationViolation, with the object property hasCause. We can query symptom-based measurements using this SPARQL query fragment:

```
?healthEffect water:hasSymptom health:high_blood_pressure.
?healthEffect rdf:type water:HealthEffect.
?healthEffect water:hasCause ?cause.
?cause owl:intersectionOf ?restrictions.
?restrictions list:member ?restriction.
?restriction owl:onProperty water:hasCharacteristic.
?restriction owl:hasValue ?characteristic.
```

[20] http://protectingourwaters.wordpress.com/2011/06/16/black-water-and-brazenness-gas-drilling-disrupts-lives-endangers-health-in-bradford-county-pa/

[21] http://switchboard.nrdc.org/blogs/amall/one_familys_life_in_the_gas_pa.html

[22] As obtained from the NPDWRs at http://water.epa.gov/drink/contaminants/index.cfm

```
?measurement water:hasCharacteristic ?characteristic.
```

Based on this modeling, the portal has been extended to begin to address health concerns: (1) the user can specify his/her health concern and the portal will detect only the water pollution that has been correlated the particular health concern; (2) the user can query the possible health effects of each contaminant detected at a polluted site, which is useful for identifying potential effects of water pollution and for identifying appropriate responses (e.g., boiling water to kill germs, using water only for bathing but not for drinking, etc.)

5.2 Time as Provenance

Temporal considerations were non-trivial in regulation modeling. The thresholds defined in both the NPDWRs' MCLs and state water quality regulations became effective nationally at different times for different contaminants[23]. For example, in the "2011 Standards & Guidelines for Contaminants in Massachusetts Drinking Water", the date that the threshold of each contaminant was developed or last updated can be accessed by clicking on the contaminant's name on the list. The effective time of the regulations has semantic implications: if the collection time of the water measurement is not in the effective time range of the constraint, then the constraint should not be applied to the measurement. In principle, we can use OWL2 RangeRestriction to model time interval constraints as we did on threshold.

5.3 Regulation Mapping and Comparison

The majority of the portal domain knowledge stems from water regulations that stipulate contaminants, pollution thresholds, and contaminant test options. Besides using semantics to clarify the meaning of water regulations and support regulation reasoning, we can also perform analysis on regulations. For example, we compared regulations from five different sources and shows substantial variation.

By modeling regulations as OWL classes, we may also leverage OWL subsumption inference to detect the correlations between thresholds across different regulatory bodies and this knowledge could be further used to speed up reasoning. For example, the California regulations are stricter than the EPA regulations concerning Methoxychlor so we can derive two rules: 1) with respect to Methoxychlor, if a water site is identified as polluted according to the EPA regulations, it is polluted according to the California regulations; and 2) with respect to Methoxychlor, if the available data supports no threshold violation according to the California regulations, it will not exceed thresholds according to the EPA regulations. Since regulations such as these can be subclassed, reasoning efficiencies may be realized when multiple regulations are used to evaluate pollution.

[23] Personal communication with the Office of Research and Standards, Massachusetts Department of Environmental Protection

5.4 Scalability

The large number of triples generated during the conversion phase prohibits classifying the entire dataset in real time. We have tried several approaches to improve reasoning speed: organize observation data by state, filter relevant data by zip code (we can derive county using zip code), and reasoning over the relevant data on one (or a small number of) selected regulation(s).

The portal assigns one graph per state to store the integrated data. The triple count at the state level is still quite large: we currently host 29.45 million triples from EPA and 68.03 million triples from USGS for California water quality data. Therefore, we refine the granularity to county level using a CONSTRUCT query (see below). This operation reduces the number of relevant triples to a manageable 10K to 100K size.

```
CONSTRUCT {
  ?s rdf:type water:MeasurementSite.
  ?s water:hasMeasurement ?measurement.
  ?s water:hasStateCode ?state.
  ?s wgs84:lat ?lat.      ?s wgs84:long ?long.
  ?measurement water:hasCharacteristic ?characteristic.
  ?measurement water:hasValue ?value.
  ?measurement water:hasUnit ?unit.
  ?measurement time:inXSDDateTime ?time.
  ?s water:hasCountyCode 085. }
WHERE { GRAPH <http://sparql.tw.rpi.edu/source/usgs-
gov/dataset/national-water-information-system-nwis-
measurements/06>
{ ?s rdf:type water:MeasurementSite.
  ?s water:hasUSGSSiteId ?id.
  ?s water:hasStateCode ?state.
  ?s wgs84:lat ?lat.      ?s wgs84:long ?long.
  ?measurement water:hasUSGSSiteId ?id.
  ?measurement water:hasCharacteristic ?characteristic.
  ?measurement water:hasValue ?value.
  ?measurement water:hasUnit ?unit.
  ?measurement time:inXSDDateTime ?time.
  ?s water:hasCountyCode 085. }}
```

5.5 Maintenance Costs for Data Service Provider

Although government agencies typically publish environmental data on the web and allow citizens to browse and download the data, not all of their information systems are designed to support bulk data queries. In our case, our programmatic queries of the EPA dataset were blocked. From a personal communication with the EPA, we were surprised to find that our previous continuous data queries were impacting their operations budget since they are charged for queries. Consequently, we filed an online form requesting a bulk data transfer from the EPA which has recently been processed.

In contrast, the USGS provides web services to facilitate periodic acquisition and processing of their water data via automated means.

5.6 System Evaluation

We provide an online questionnaire [24] to collect feedback from users. In the questionnaire, we ask the users to identify themselves as experts or lay users, then ask them to rate the data quality, responsiveness, and user interface of the portal. The questionnaire also solicits free text comments from users. We will report preliminary results of this ongoing user study at the conference.

6 Related Work

Three areas of work are considered most relevant to this work, namely knowledge modeling, data integration, and provenance tracking of environmental data.

Knowledge-based approaches have begun in environmental informatics. Chen et al. [5] proposed a prototype system that integrates water quality data from multiple sources and retrieves data using semantic relationships among data. Chau [4] presented an ontology-based knowledge management system (KMS) to enable novice users to find numerical flow and water quality models given a set of constraints. OntoWEDSS [3] is an environmental decision-support system for wastewater management that combines classic rule-based and case-based reasoning with a domain ontology. Scholten et al. [14] developed the MoST system to facilitate the modeling process in the domain of water management. The Registry of EPA Applications, Models and Databases (READ)[25] supports management of information resources. It collects life cycle phase information, how the resource supports environmental statutes, and whether the resource interfaces with other EPA information resources. A comprehensive review of environmental modeling approaches can be found in [17]. SemantEco and SemantAqua differ from these projects since they support provenance-based query and data visualization. Moreover, SemantAqua is built upon standard semantic technologies (e.g. OWL, SPARQL, Pellet, Virtuoso) and thus can be easily replicated or expanded.

Data integration across providers has been studied for decades by database researchers. In the area of ecological and environmental research, shallow integration approaches are taken to store and index metadata of data sources in a centralized database to aid search and discoverability. This approach is applied in systems such as KNB[26] and SEEK[27]. Our integration scheme combines a limited, albeit extensible, set of data sources under a common ontology family. This supports reasoning over the integrated data set and allows for ingest of future data sources.

[24] http://was.tw.rpi.edu/swqp/questionnaire/portal_questionnaire.php

[25] http://iaspub.epa.gov/sor_internet/registry/systmreg/home/overview/home.do

[26] Knowledge Network for Biocomplexity Project. http://knb.ecoinformatics.org/index.jsp

[27] The Science Environment for Ecological Knowledge. http://seek.ecoinformatics.org

There also has been a considerable amount of research efforts in semantic provenance, especially in the field of eScience. myGrid [19] proposes the COHSE open hypermedia system that generates, annotates and links provenance data in order to build a web of provenance documents, data, services, and workflows for experiments in biology. The Multi-Scale Chemical Science [12] (CMCS) project develops a general-purpose infrastructure for collaboration across many disciplines. It also contains a provenance subsystem for tracking, viewing and using data provenance. A review of provenance techniques used in eScience projects is presented in [15]. While these eScience projects design their own schemes for modeling provenance, the SemantAqua portal encodes provenance with PML 2, which is a general purpose interlingua for sharing explanations generated by various automated systems. These eScience projects keep provenance for uses like improving data quality, facilitating audits, and data replicability. Our portal demonstrates that provenance also can be used for developing and customizing web applications (e.g. generating the data source facet).

7 Conclusions and Future Work

We presented a semantic technology-based approach to ecological and environmental monitoring and described our work using this approach in the Tetherless World Constellation SemantEco approach and the SemantAqua Portal. SemantAqua supports both non-expert and expert users in water quality monitoring. We described the overall design and highlighted some benefits from utilizing semantic technologies, including: the design of the ontologies, the methodology used to perform data integration, and the encoding and usage of provenance information generated during data aggregation. The SemantAqua portal demonstrates some benefits and potential of applying semantic web technologies to environmental information systems.

A number of extensions to this portal are in process. First, only four states' regulations have been encoded. We intend to encode the regulations for the remaining states whose regulations differ from the federal regulations. Second, data from other sources, e.g. weather, may yield new ways of identifying pollution events. For example, a contaminant control strategy may fail if heavy rainfall causes flooding, carrying contaminants outside of a prescribed area. It would be possible with real-time sensor data to observe how these weather events impact the portability of water sources in the immediate area. We are also applying this approach to other monitoring topics, e.g. air quality, food safety, and health impacts.

References

1. Batzias, F. A., and Siontorou, C. G.: A Knowledge-based Approach to Environmental Biomonitoring. In: Environmental Monitoring and Assessment, vol.123, pp.167–197 (2006)
2. Carroll, J. J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., and Wilkinson, K.: Jena: Implementing the semantic web recommendations. In: 13th International World Wide Web Conference, pp. 74-83 (2004)

3. Ceccaroni, L., Cortes, U. and Sanchez-Marre, M.: OntoWEDSS: augmenting environmental decision-support systems with ontologies. In: Environmental Modelling & Software, vol. 19(9), pp. 785-797 (2004)
4. Chau, K.W.: An Ontology-based knowledge management system for flow and water quality modeling. In: Advances in Engineering Software, vol. 38(3), pp. 172-181 (2007)
5. Chen, Z., Gangopadhyay, A., Holden, S. H., Karabatis, G., McGuire, M. P.: Semantic integration of government data for water quality management. In: Government Information Quarterly, vol. 24(4), pp. 716–735 (2007)
6. Ding, L., Lebo, T., Erickson, J. S., DiFranzo, D., Williams, G. T., Li, X., Michaelis, J., Graves, A., Zheng, J. G., Shangguan, Z., Flores, J., McGuinness, D. L., and Hendler, J.: TWC LOGD: A Portal for Linked Open Government Data Ecosystems, In: JWS special issue on semantic web challenge'10, (2010)
7. Hitzler, P., Krotzsch, M., Parsia, B., Patel-Schneider, P., Rudolph, S.: OWL 2 Web Ontology Language Primer, <http://www.w3.org/TR/owl2-primer/> (2009)
8. Holland, D. M., Principe, P. P. and Vorburger, L.: Rural Ozone: Trends and Exceedances at CASTNet Sites. In: Environmental Science & Technology, vol. 33 (1), pp. 43-48 (1999)
9. Lebo, T., Williams, G.T.: Converting governmental datasets into linked data. Proceedings of the 6th International Conference on Semantic Systems. In: I-SEMANTICS '10, pp. 38:1–38:3 (2010)
10. Liu, Q., Bai, Q., Ding, L., Pho, H., Chen, Kloppers, C., McGuinness, D. L., Lemon, D., Souza, P., Fitch, P. and Fox, P.: Linking Australian Government Data for Sustainability Science - A Case Study. In: Linking Government Data (chapter), accepted (2011)
11. McGuinness, D.L., Ding, L., Pinheiro da Silva, P., and Chang, C.: PML 2: A Modular Explanation Interlingua. In: Workshop on Explanation-aware Computing (2007)
12. Myers, J., Pancerella, C., Lansing, C., Schuchardt, K., and Didier, B.: Multi-scale science: Supporting emerging practice with semantically derived provenance. In: ISWC workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data (2003)
13. Raskin, R. G., and Pan, M. J.: Knowledge representation in the semantic web for Earth and environmental terminology (SWEET). In: Computers & Geosciences, vol. 31(9), pp. 1119-1125 (2005)
14. Scholten, H., Kassahun, A., Refsgaard, J. C., Kargas, T., Gavardinas, C., and Beulens, A. J. M.: A Methodology to Support Multidisciplinary Model-based Water Management. In: Environmental Modelling and Software, vol. 22(5), pp. 743–759 (2007)
15. Simmhan, Y. L., Plale, B., and Gannon, D.: A survey of data provenance in e-science. In: ACM SIGMOD Record, vol. 34(3), pp. 31-36 (2005)
16. Sirin, E., Parsia, B., Cuenca-Grau, B., Kalyanpur, A., and Katz, Y.: Pellet: A practical OWL-DL reasoner. In: Journal of Web Semantics, vol. 5(2), pp. 51-53 (2007)
17. Villa, F., Athanasiadis, I. N., and Rizzoli, A. E.: Modelling with knowledge: A Review of Emerging Semantic Approaches to Environmental Modelling. In: Environmental Modelling and Software, vol. 24(5), pp. 577-587 (2009)
18. Wang, P., Zheng, J.G., Fu, L.Y., Patton E., Lebo, T., Ding, L., Liu, Q., Luciano, J. S., McGuinness, D. L.: TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring[28]. Technical Report, http://tw.rpi.edu/media/latest/twc-swqp.doc (2011)
19. Zhao, J., Goble, C. A., Stevens, R. and Bechhofer S.: Semantically linking and browsing provenance logs for e-science. In: Semantics of a Networked World, vol. 3226, pp. 158-176 (2004)

[28] SWQP has been renamed SemantAqua that is one instantiation of the SemantEco design for Ecological and Environmental Monotoring.