TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring

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 environmental information systems. We used our linked data approach in the Tetherless World Constellation Semantic Water Quality Portal (TWC-SWQP). Our integration scheme uses a core domain ontology and integrates water data from different authoritative sources along with multiple regulation ontologies to enable pollution detection and monitoring. An OWL-based reasoning scheme identifies pollution events relative to user chosen regulations. Our approach also captures and leverages provenance to improve transparency. In addition, semantic water quality portal features provenance-based facet generation, query answering and data validation over the integrated data via SPARQL. We introduce the approach and the water portal, and highlight some of its potential impacts for the future of environmental monitoring systems.

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

1 Introduction

Concerns over environmental issues such as biodiversity loss [1], water problems [13], atmospheric pollution [8], and sustainable development [9] have highlighted the need for reliable information systems to support monitoring of environmental trends, support scientific research and inform citizens. In particular, semantic technologies have been used in environment monitoring information systems to facilitate domain knowledge integration across multiple sources and support collaborative scientific workflows [16]. Meanwhile, growing interests have been observed from citizens, demanding direct and transparent access to 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]

[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

In this paper, we describe a semantic technology-based approach to environmental monitoring. We deployed the approach in the Tetherless World Constellation's Semantic Water Quality Portal (TWC-SWQP). TWC-SWQP is an exemplar next generation environmental monitoring portal that simultaneously supports water quality investigation for lay people as well as experts and also helps us evaluate our linked data approach in real world environmental settings. The portal integrates water monitoring and regulation data from multiple sources following Linked Data principles, captures the semantics of water quality domain knowledge using a simple OWL2 [7] ontology, preserves provenance metadata using the Proof Markup Language (PML) ontology [11], and infers water pollution events using OWL2 inference. The web portal delivers water quality information and reasoning results to citizens via a faceted browsing capable map interface[2].

The contributions of this work are multi-faceted. The overall design provides a model that may be used for creating environmental monitoring portals. The design has been used to develop a water quality portal (TWC-SWQP) that allows anyone, including those who do not have in-depth knowledge of water pollution regulations or water data sources, to monitor water quality in any region of the United States. It also shows potential directions that environmental monitoring systems may take to empower citizen scientists and create partnerships between concerned citizens and environmental officials. 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 water quality 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 the rest of this paper, section 2 reviews selected challenges in the development of the TWC-SWQP 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.

2 Challenges in Water Quality Information System

We focus on water quality monitoring in our current project, and propose a publicly accessible semantically-enabled water information system that facilitates discovery of polluted water, polluting facilities and the specific contaminants. However, to construct such an information system, we need to overcome the following challenges.

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

2.1 Modeling Domain Knowledge in Water Quality Monitoring

We focus on three types of water quality monitoring knowledge: observational data items (e.g., the amount of arsenic in water) collected by sensors and humans, regulations (e.g., safe drinking water acts) published by authorities, and water domain knowledge maintained by scientists (e.g., water-relevant contaminants, bodies of water, etc).

The observational data includes water quality characteristics together with the corresponding descriptive metadata including the type and unit of the data item as well as the provenance metadata such as the locations of sensor sites, the time when the data item was observed and optionally the test methods and devices used to generate the observation. A light-weight extensible domain ontology is needed to enable reasoning on observational data while limiting ontology development and understanding costs. We identified some relevant ontologies. A small portion of SWEET[3] models general water concepts (e.g. bodies of water). Chen et al. [5] models relationships among water quality datasets. Chau et al. [3] models a specific aspect of water quality. While all provide relevant terms, none covers our needs (while they simultaneously model notions we do not need).

Table 1. Subset of contaminant thresholds.

Contaminants	Rhode Island	EPA	New York	Massachusetts	California
Acetone	-	-	-	6.3 mg/l	-
Nitrate+Nitrite	-	-	-	-	0 mg/l
Tetrahydrofuran	-	-	-	1.3 mg/l	-
Methyl isobutyl ketone	-	-	-	0.35 mg/l	-
1,1,2,2-Tetrachloroethane	0.0017 mg/l	-	-	-	0.001 mg/l
1,2-Dichlorobenzene	0.42 mg/l	-	-	-	0.6 mg/l
Acenaphthene	0.67 mg/l	-	-	-	-
Aldicarb sulfoxide	-	-	0.004 mg/l	0.004 mg/l	-
Chlorine Dioxide (as ClO2)	-	-	0.8 mg/l	-	-

Water 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). In addition to federal regulations,

[3] Semantic Web for Earth and Environmental Terminology. http://sweet.jpl.nasa.gov/

[4] NPDWRs information can be found at http://water.epa.gov/drink/contaminants/index.cfm

individual states can enforce their own water regulations and guidelines. For instance, the Massachusetts Department of Environmental Protection (MassDEP) issued the "2011 Standards & Guidelines for Contaminants in Massachusetts Drinking Water"[5], which contains rules specifying thresholds for 139 contaminants. The water regulations are diverse in that they define different sets of contaminants with different contaminant thresholds as shown in Table 1[6]. Therefore, we need an interoperable model that represents a diverse collection of regulations together with the observational data and domain knowledge from different sources. According to our survey, regulations concerning water quality have not been modeled as part of any existing ontology so far. The best we found is regulation specifications organized in HTML tables.

2.2 Collecting Real World Data

Both the EPA and the US Geological Survey (USGS) released 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[7]) under the Clean Water Act (CWA) from ICIS-NPDES, an EPA system. The NPDES datasets contain descriptions of the facilities (e.g. name, permit number, address, and geographic location) and measurements of contaminants in the water discharged by the facilities for up to five test types per contaminant. USGS provides the National Water Information System (NWIS[8]) to publish data about water monitoring sites (e.g. identifier, geographic location) and the measurements of water characteristics from samples.

Although datasets from the EPA and USGS are organized as data tables, it is not easy to mash up them due to syntactic and semantic differences. In particular, we observe a need 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 chemicals, may be used in domains other than water quality monitoring, so it would be useful to link to other accepted models such as chemical element descriptions, e.g. ChemML. (iii) Most observational data are complex data objects. For example, Table 2 shows a table 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" indicates 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.

[5] http://www.mass.gov/dep/water/drinking/standards/dwstand.htm

[6] See complete table at http://tw.rpi.edu/web/project/TWC-SWQP/compare_five_regulation

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

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

Table 2. 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

2.3 Provenance Tracking and Provenance-Aware Computing

In order to enhance transparency and encourage community participation, a citizen facing information system should track provenance metadata in data processing and leverage provenance metadata in its computational services.

A water quality monitoring system that mashes up data from different sources should maintain and expose data sources on demand. This enables data curators to get credit for their contributions and allows users to choose data from their trusted sources. The data sources can be automatically refreshed if we update 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 environmental monitoring. We are testing this approach with a water quality monitoring application at scale.

3.1 Domain Knowledge Modeling and Reasoning

We use an ontology-based approach to model domain knowledge in environmental information systems. A core ontology[9] includes the terms of interest in a particular environmental area (e.g., water quality) for encoding observational data, and regulation ontologies[10] include terms required for describing compliance levels (and pollution levels). The two types of ontologies are connected to leverage OWL inference to reason about the compliance of observations with regulations.

[9] http://purl.org/twc/ontology/swqp/core

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

Core Ontology Design

Complete ontology reuse is rare because most predefined ontologies do not completely cover all the domain concepts involved in an environmental information system. As mentioned in section 2.1, existing water ontologies are insufficient for our water quality monitoring application. We therefore defined a light-weight water quality ontology that reuses and is compliant with existing ontologies. For example, although the SWEET ontology does not contain water pollution terms, it does contain water body terms, such as Lake and Stream, which are reused in our ontology. This allows us to lower the cost of ontology development and maintenance since we can rely on other authoritative sources to define and maintain core terms - in this case re-using SWEET's BodyOfWater subhierarchy. The core ontology models domain objects (e.g. water sites, facilities, measurements, and characteristics[11]) as classes, and the relationships among the domain objects (e.g. hasMeasurement, hasValue, hasUnit) as properties. A subset of the ontology is illustrated in Figure 1. We model a polluted water site (PollutedWaterSource) as the intersection of a water site (WaterSource) and something that has a characteristic measurement with a value that exceeds its threshold, i.e., it satisfies an owl:Restriction that encodes the specific definition of an excessive measurement for a characteristic as a numeric range constraint. However, the thresholds for the characteristic measurements are not defined in the core ontology, but in the regulation ontology, so that polluted water sites can be detected with different regulations.

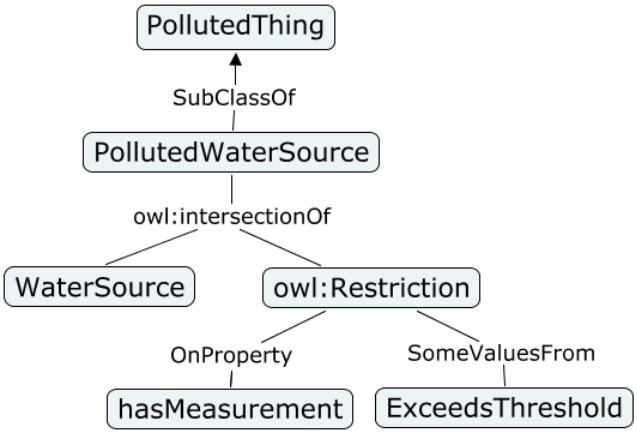

Fig. 1. Portion of the TWC Core Water Ontology.

Regulation Ontology Design

In order to support the diverse collection of federal and state water quality regulations, the core ontology is extended to map each rule in the regulations into an OWL class. The conditions of a rule are mapped to owl restrictions. We use numeric range restrictions on a datatype property to encode the allowable ranges of the water

[11] Our ontology uses characteristic instead of contaminant based on the consideration that characteristics measured like PH, temperature are not contaminants.

characteristics defined in the regulations. The rule-compliance results is 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. Drinking water is considered polluted if the concentration of Arsenic is more than 0.01 mg/L. In the mapped regulation ontology, we create the class ExcessiveArsenicMeasurement as a water measurement with value greater than or equal to 0.01 mg/L.

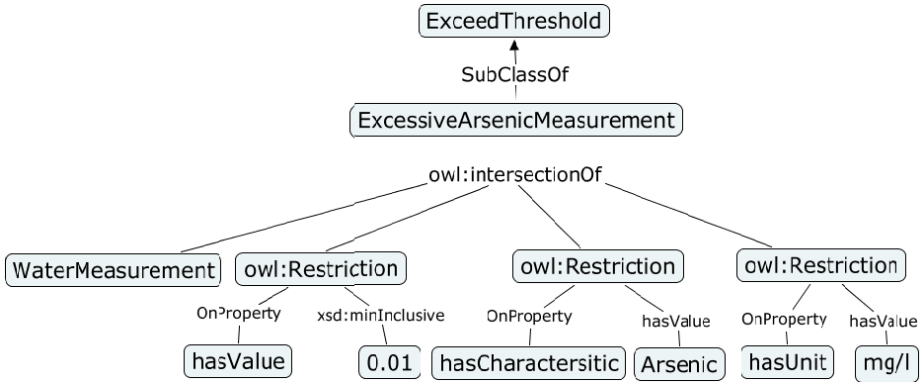

Fig. 2. Portion of EPA Regulation Ontology.

Reasoning Domain Data with Regulations

Combining the observational data items collected at a water monitoring site, the core ontology and the regulation ontology, a reasoner can decide if the corresponding water body is polluted using OWL2 classification inference. This design provides several benefits. First, the core ontology is small and easy to maintain. Our core ontology consists of only 18 classes, 4 object properties, and 10 data properties. Secondly, the ontology component can be easily extended to incorporate more regulations. We wrote converters to extract federal and four states' regulation data from HTML web pages and translated them into OWL 2 [7] constraints that align with the core 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 is in different forms. The design leads to flexible querying and reasoning: the user can select the regulation to apply on the data and the reasoner will reason using only the ontology for the selected regulation together with the core ontology and the water quality data. 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 on the same region, the portal identified 15 polluted water sites containing the 2 detected with Rhode Island regulations and 7 same polluting facilities. The results show that California regulations are stricter than Rhode Island's, 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, environmental monitoring systems can benefit from adopting the data conversion and organization capabilities enabled by the TWC-LOGD portal [6]. In the TWC-SWQP project, we used the open source tool **csv2rdf4lod**[12] to convert the data from EPA and USGS into Linked Data.

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 twcwater:hasCharacteristic. Similarly, we map spatial location data to properties from an external ontology, e.g. wgs84[13]:lat and wgs84:long.

Aligning instance references: We promote references to chemicals in our water quality data from literal to URI, e.g. "Arsenic" is promoted to "twcwater: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 DBpedia URI (e.g., "Nitrate/Nitrite" from MassDEP's regulations maps two DBpedia URIs), and some instances may not be defined in DBpedia (e.g., "C5-C8" from MassDEP's regulations). By linking to DBpedia URIs, we reserve the opportunity to connect to other knowledge base such as disease database.

Converting complex objects: As discussed in section 2.1, we may 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, which is done by first marking each cell value that should be treated as a subject in a triple, and then bundling the related cell values with the marked subject. The details can be found in [17].

3.3 Provenance Tracking and Provenance-Aware Computing

The provenance data in the system come from two sources: (i) provenance metadata can be embedded in the original EPA and USGS datasets, e.g. measurement location and time; (ii) the portal automatically captures provenance data during the data integration stages and encodes them in PML2 [11] due to the provenance support from csv2rdf4lod. At the retrieval stage, we capture provenance, e.g. the URL of the data source, who fetched the source data at what time, and what agent and protocol are used for retrieving the data. At the conversion stage, we keep provenance, e.g. what engine performs the conversion, what antecedent data are involved, and what roles those data play. At the publication stage, we capture provenance, e.g. who loaded the data to the triple store at what time. When we convert the regulations, we capture their provenance pragmatically. We reveal these provenance data via pop up window when the user selects a measurement site or facility.

[12] http://purl.org/twc/id/software/csv2rdf4lod
[13] http://www.w3.org/2003/01/geo/wgs84_pos

In the TWC-SWQP, we used the provenance metadata to enable dynamic data source listing and provenance-aware cross validation over EPA and USGS data.

Data Source as Provenance

TWC-SWQP utilizes provenance information about data sources 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 only the data sources he/she trusts (see Figure 4). The system would then only return results within the selected sources.

This is just one usage of the provenance information, which can also be used to give the user more options to specify 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 [17].

Provenance-Aware Cross-Validation over EPA and USGS Data

Since we maintain the information of where each piece of data comes from, our system can compare water quality data originating from different sources for the purpose of cross-validation. When we performed the comparison between EPA and USGS data, we got interesting results. Figure 3 shows the measurement of PH collected by an EPA facility (at 41:59:37N, 71:34:27W) and a USGS site (at 41:59:47N, 71:33:45W) that are located within 1KM from each other for a common period. Note that the PH values measured by USGS went below the minimum value from EPA quite often and went above the maximum value from EPA once. The way we find two locations close to each other is through the SPARQL filter shown in (1).

FILTER (?facLat < (?siteLat+"+delta+") && ?facLat > (?siteLat-"+delta+") && ?facLong < (?siteLong+"+delta+") && ?facLong > (?siteLong-"+delta+")) (1)

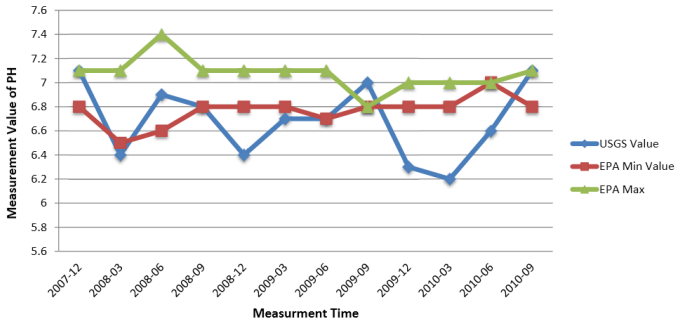

Fig. 3. Data Validation Example

4 Semantic Water Quality Portal

4.1 System Implementation

Figure 4 shows the semantic water quality portal as it supports water pollution identification. The user specifies a geographic region of interest by entering a zip code (mark 1), and can customize queries from multiple facets: data source (mark 3), water regulation (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: the names of contaminants, the measured values, the limit values, and time of measurement. The window also provides a link that displays a trend graph of the water quality over time.

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[14] and retrieved via SPARQL queries. The portal utilizes the Pellet OWL Reasoner [15] 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.

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.

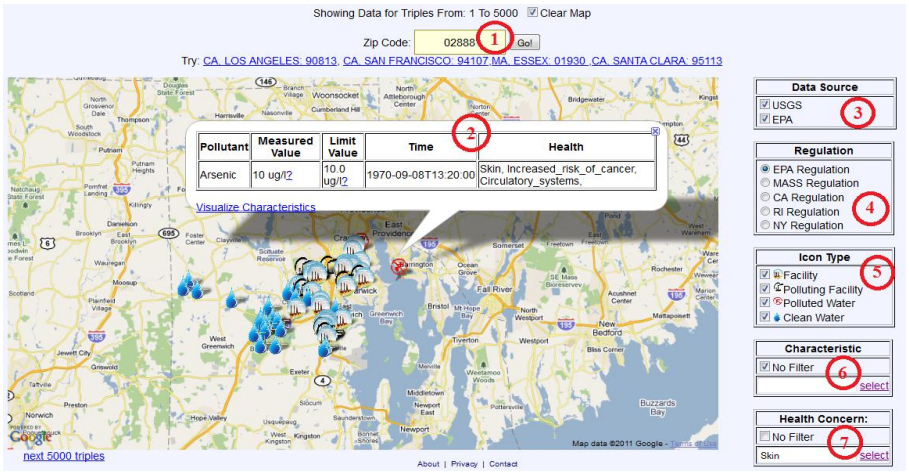

Fig. 4. Water Quality Portal In Action

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

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 on the order of billions of 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.

The numbers of the classes we generated for modeling the rules from the different regulations are listed in Table 3. Our programmed conversion provides a quick and low cost approach for encoding regulations. It took us about 2 person-days to encoding hundreds of rules.

Table 3. Number of threshold classes converted from regulations

EPA	CA	MA	NY	RI
83	104	139	74	100

Fig. 5. Number of triples for the four states and the average number

5 Discussion

5.1 Linking to 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, PA has been reported to generate numerous problems[15]. 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

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

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 can map relationships between contaminants and health impacts to an OWL class. We use the object property "hasSymptom" to connect the classes with their symptoms, e.g. twchealth:high_blood_pressure. The classes of health effects are related to classes of thresholds, e.g. ExcessiveLeadMeasurement, with the object property hasCause. We can query symptom-based measurements using the SPARQL query fragment below.

```
?healthEffect twcwater:hasSymptom twchealth:high_blood_pressure;
?healthEffect rdf:type twcwater:HealthEffect. ?healthEffect twcwater:hasCause
?cause. ?cause owl:intersectionOf ?restrictions.
?restrictions list:member ?restriction. ?restriction owl:onProperty
twcwater:hasCharacteristic. ?restriction owl:hasValue ?characteristic.
?measurement twcwater:twcwater: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 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 speed up the reasoning process: organize observation data by county, filter relevant data by zip code (we can derive county using zip code), and reasoning over the relevant data on one selected regulation.

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 twcwaterMeasurementSite . ?s twcwaterhasMeasurement ?measurement.
?s twcwaterhasStateCode ?state. ?s wgs:lat ?lat. ?s wgs:long ?long.
```

?measurement twcwaterhasCharacteristic ?element. ?measurement twcwaterhasValue ?value. ?measurement twcwaterhasUnit ?unit. ?measurement time:inXSDDateTime ?time. ?s twcwaterhasCountyCode 085. }
WHERE { GRAPH <http://sparql.tw.rpi.edu/source/usgs-gov/dataset/national-water-information-system-nwis-measurements/36>
{?s rdf:type twcwaterMeasurementSite . ?s twcwaterhasUSGSSiteId ?id.
?s twcwaterhasStateCode ?state. ?s wgs:lat ?lat. ?s wgs:long ?long.
?measurement twcwaterhasUSGSSiteId ?id. ?measurement
twcwater:hasCharacteristic ?element.?measurement twcwaterhasValue ?value.
?measurement twcwaterhasUnit ?unit. ?measurement time:inXSDDateTime ?time. ?s twcwaterhasCountyCode 085.} }

5.3 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[16]. 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 threshold constraint, then the threshold 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.

5.4 Regulation Mapping and Comparison

The majority of the portal domain knowledge stems from water regulations that stipulate contaminants, thresholds for pollution, and pollutant 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, Table 1 compares 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, California is stricter than the EPA 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, it is polluted according to the CA regulation; and 2) with respect to Methoxychlor, if the available data supports no pollution threshold violation according to the California regulation, then it will not exceed thresholds according to the EPA regulation. We can use subclass to model such rules in order to

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

evaluate subsuming relationships. This could spare some reasoning time when multiple sets of regulations are applied to detect the pollution.

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 out that our continuous data queries have impacted their operations budget since they are charged for queries. We have filed an online form requesting a bulk data transfer from the EPA which is being processed. In contrast, the USGS provides web services to facilitate periodic acquisition and processing of their water data via automated means.

6 Related Work

Three areas of work are considered related to this paper, 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. [13] developed the MoST system to facilitate the modeling process in the domain of water management. A comprehensive review of environmental modeling approaches can be found in [16]. SWQP differs from these projects in that it supports provenance based query and data visualization. Moreover, SWQP 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[17] and SEEK[18]. Our integration scheme combines a limited, albeit extensible, set of data sources under a common ontology. This supports reasoning over the integrated data set and allows for ingest of future data sources.

There also has been a considerable amount of research efforts in semantic provenance, especially in the field of e-Science. myGrid [18] proposes the COHSE open hypermedia system that generates, annotates and links provenance data in order

[17] Knowledge Network for Biocomplexity Project. http://knb.ecoinformatics.org/index.jsp
[18] The Science Environment for Ecological Knowledge. http://seek.ecoinformatics.org

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 the provenance techniques used in e-science projects is presented in [14].

7 Conclusions and Future Work

We presented a semantic technology-based approach to environmental monitoring and described our work using this approach in the Tetherless World Constellation Semantic Water Quality Portal. TWC-SWQP supports both non-expert and expert users in water quality monitoring. We described the overall design and highlighted some ways that it benefits from utilizing semantic technologies, including: the design of the water ontology and its roots in SWEET, the methodology used to perform data integration, and the encoding and usage of provenance information generated during data aggregation. The portal example demonstrates the benefits and potential of applying semantic web technologies to environmental information systems.

A number of extensions to this portal are ongoing. First, only a handful of states' regulations have been encoded, and we intend to encode the regulations for the remaining states that have regulations that 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 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. Lastly, we would like to apply this architecture to other applications, such as the Clean Air Status and Trends demo[19], by enabling these applications to expose regulation and sample data.

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 manage ment system for flow and water quality modeling. In: Advances in Engineering Software, vol. 38(3), pp. 172-181 (2007)

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

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, accepted (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., 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. 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)
14. 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)
15. 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)
16. 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)
17. 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. Technical Report, http://tw.rpi.edu/media/latest/twc-swqp.doc (2011)
18. 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)