Semantically Enabling Next Generation Environmental Informatics Portals

by

Ping Wang

A Thesis Submitted to the Graduate

Faculty of Rensselaer Polytechnic Institute

in Partial Fulfillment of the

Requirements for the degree of

MASTER OF SCIENCE

Major Subject: Computer Science

Approved:

_____

Deborah L. McGuinness, Thesis Adviser

Rensselaer Polytechnic Institute
Troy, New York

April 2012
(For Graduation May 2012)

CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENT

I would like to express my sincere gratitude to Prof. Deborah L. McGuinness, my graduate advisor. Prof. McGuinness has been a fantastic advisor and great mentor to me, guiding and motivating me from the very inception of the research work. This research project would not have been possible without all the assistance, support and guidance from Prof. McGuinness. I am very thankful to Prof. Jim Hendler and Prof. Peter Fox for their advice and support during the course of the research work. I would also like to thank Prof. Joanne S. Luciano for the valuable help and suggestions that she gave me on the work. I also wish to convey thanks to the professors for providing me with the greatest research environment I have ever been to: Tetherless World Constellation at RPI.

I would also like to thank the staff members of the Department of Computer Science and the Graduate School for helping me with the administrative details concerning the thesis.

Special thanks also to all my friends, especially lab members: Evan W. Patton, Timothy Lebo, Jin Guang Zheng, Li Ding and Linyun Fu for all the help I received from them.

Finally, yet importantly, I would like to express my heartfelt thanks to my beloved parents for their understanding, support and encouragement.

ABSTRACT

Environmental informatics web portals, which have environmental information systems as back end and web portals as front end, can be used as the right platform to enable both professionals and citizens better utilize environmental data and investigate environmental problems. We present a semantic technology-based approach for building environmental informatics portals, and deployed the approach in the Tetherless World Constellation's Semantic Ecology and Environment Portal (SemantEco). The exemplar portal captures the semantics of domain knowledge using a family of modular simple ontologies, integrates environmental data from multiple sources following Linked Data principles, and infers environment pollution events using OWL2 inference. The portal captures provenance, and leverages provenance in multiple ways, including data lineage rendering, provenance-based facet generation, and validation over the integrated data via SPARQL. We then describe the implementation which has been built out in the domain of water quality monitoring, and highlight some of the potential extensions and enhancements for future semantically-enabled environmental informatics portals.

1. Introduction

The extent of environmental change over the past decades has evoked concerns over ecological and environmental issues, such as biodiversity loss [1], water problems [2], atmospheric pollution [3], and sustainable development [4], in both environmental scientists and citizens who understand the importance of environmental sustainability.

Environmental problems are usually very complicated problems to tackle in that they require significant domain knowledge and are involved with large amounts of data. For example, to monitor and control the quality of drinking water, government agencies establish regulations to define pollution in terms of acceptable levels of the predefined water characteristics[1], set up sites where water samples are taken, and regularly measure the quantity of the water characteristics in water samples. After government agencies generate the water quality data, they open the environmental data to the public on the web. However, such data cannot be readily utilized by citizens and professionals due to several reasons [4]. First, given that citizens and professionals often do not understand the responsibilities and operations of various government agencies well, it is often difficult for them to find relevant and useful data quickly. Secondly, datasets from multiple sources are released in different formats, e.g. CSV, HTML, TXT. The users need to either transform the different data formats into a common representation or to create mappings between the representations before they can conduct analysis over the heterogeneous datasets. Lastly, the semantics of the data are not explicitly encoded, and the users from the web often have a hard time making sense of the data.

To deal with the above challenges, semantic technologies have been used in environmental monitoring information systems. Ontologies are found to have multiple

---

[*] Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "A Semantic Portal for Next Generation Monitoring Systems," in *Proceedings of the 10th International Semantic Web Conference*, Bonn, Germany, 2011, pp. 253-268.
Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring," TWC RPI, Troy, NY, 2011.
Portions of this chapter previously appeared as: J. G. Zheng, P. Wang, E. W. Patton, T. Lebo, J. Luciano, and D. L. McGuinness, "A Semantically-Enabled Provenance-Aware Water Quality Portal," presented at the Environmental Information Management Conference 2011, Santa Barbara, CA, 2011.

[1] We use the term characteristic instead of contaminant based on the consideration that some characteristics measured like pH and temperature are not contaminants.

usages for environmental domains [5] such as providing controlled vocabularies, thesauri and taxonomies, formalizing common knowledge and enabling machine reasoning. Meanwhile, the linked data principle [6] presents a set of best practices for publishing and connecting structured data on the web to build a world scale data space. Supporting technologies that facilitate integration of data from different sources have been developed and practiced in aggregating government data [7], including environmental data, fiscal data, etc.

Meanwhile, demand has increased for direct and transparent access to ecological and environmental information systems. Such demand is reflected in a real world use case. In 2009, after a recent water quality episode in Bristol County, Rhode Island where E. coli was reported in the water [8], residents requested information concerning when the contamination began, how it happened, and what measures were being taken to monitor and prevent future occurrences. Motivated by the use case, we believe that environmental informatics web portals, which have environmental information systems as back end and web portals as front end, can be used as the right platform to enable both professionals and citizens better utilize environmental data and investigate environmental problems.

However, building environmental informatics systems as web portals introduces a new challenge, which is how to gain trust from the broad community of users from the web. Some users may not trust the analysis results and the data from a web portal if they are not provided with the option to examine how the results and data are obtained. As pointed out in [9], knowledge provenance, which includes source identification, source authoritativeness, and a supporting graph, can be used to provide explanations. With the explanations that help users understand where responses come from, and what they depend on, an environmental informatics web portal becomes more transparent and reliable from the perspective of the users.

In this thesis, we describe a semantic technology-based approach for building environmental informatics portals. We deployed the approach in the Tetherless World Constellation's Semantic Ecology and Environment Portal (SemantEco). SemantEco is an exemplar next generation environmental informatics portal that provides investigation support for lay people as well as experts while also providing a real world environmental

evaluation test bed for our linked data approach. The portal captures the semantics of domain knowledge using a family of modular simple OWL2 [10] ontologies, integrates environmental data from multiple sources following Linked Data principles, 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 users via a faceted browsing map interface.

The contributions of this work are as follows. The overall design provides an operational specification model that may be used for creating environmental informatics portals. It includes a simple ontology for modeling pollution in general and initial domain ontologies for water and air. This design has been used to develop a water quality portal (SemantAqua[2]) 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 [12]. Next, our work also exposes potential directions for environmental informatics portals as they may empower citizen scientists and enable dialogue between concerned citizens and professionals. For example, these portals 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, domain experts 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).

While the Tetherless World Constellation's semantic water quality monitoring project has been done collaboratively, there are significant portions of the work that were done individually by the thesis author. The original design was done as a group project however taking the work beyond a class project to include data for more than four states needed to be done. Issues related to scale required redesign of the system and modularization of the ontology. Additionally, we needed to automate encoding of environmental regulation rules as ontology classes, and improve the data ingestion by

[2] http://tw.rpi.edu/web/project/SemantAQUA

switching from ad hoc programs to a standard data converter. Further, we generalized the work to provide a more extensible model for semantically-enabled monitoring instead of just semantic water quality monitoring.

In chapter 2, we review some of the existing work related to our research. In chapter 3, we present our semantic technology-based approach for building environmental informatics portals, including domain knowledge modeling and reasoning, integrating environmental data, and provenance support. The individual work of the author mainly focuses on the data integration and provenance support as well as making the foundational work more extensible. In chapter 4, we describe the implementation details and scaling issues related to the water quality portal SemantAqua. In chapter 5, we discuss potential extensions and enhancements for SemantAqua and their relation to the more general system - SemantEco. We summarize and conclude in chapter 6.

2. Historical Review

In this chapter we discuss research efforts that are considered most relevant to this work from three perspectives, namely knowledge modeling, data integration, and provenance management.

2.1 Knowledge Modeling

Knowledge modeling is essential in various informatics research disciplines. In environmental informatics, a number of knowledge-based approaches have been developed. Ceccaroni et al. [13] proposed OntoWEDSS, which is an environmental decision-support system for wastewater management that combines classic rule-based and case-based reasoning with a domain ontology. Chau [14] 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. Chen et al. [15] focused on enhancing integration of data flows and business processes to enable higher levels of e-government. They took water quality management as an example and developed a prototype system that integrates water monitoring data from federal, state, and local government organizations and retrieves data using semantic relationships among data. Scholten et al. [2] developed an ontological knowledge base (KB) and a Modeling Support Tool (MoST) to facilitate the complex, usually multidisciplinary modeling process in the domain of water management. A comprehensive review of environmental modeling approaches could be found in [5].

---

[*] Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "A Semantic Portal for Next Generation Monitoring Systems," in *Proceedings of the 10th International Semantic Web Conference*, Bonn, Germany, 2011, pp. 253-268.
Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring," TWC RPI, Troy, NY, 2011.
Portions of this chapter previously appeared as: J. G. Zheng, P. Wang, E. W. Patton, T. Lebo, J. Luciano, and D. L. McGuinness, "A Semantically-Enabled Provenance-Aware Water Quality Portal," presented at the Environmental Information Management Conference 2011, Santa Barbara, CA, 2011.

2.2 Data Integration

Data integration across providers has been a long-standing challenge for researchers from multiple areas including database, artificial intelligence and semantic web science. Meanwhile, data integration is a necessity for applications that need to query across multiple heterogeneous data sources [16]. One of the major bottlenecks in data integration is generating schema matching between the source and the target data schema. Researchers have developed rule-based and learning based solutions for semi-automatic schema matching. Furthermore, modularized systems that employ both rule-based and learning-based solutions have been proposed to exploit various types of information [17]. While, most previous approaches focus on computing direct schema matches, there also have been a few works for complex matching. For example, Xu and Embley [18] developed a technique for automating match discovery via leveraging domain ontologies. In SemantAqua, we also utilize ontologies to conduct semantic data integration. However, our schema matching is done manually based on information that could be collected, such as property names, data dictionary, and communication with domain experts. It would be interesting direction to explore how to apply automating match discovery techniques in our SemantEco framework.

2.3 Provenance Support

Provenance information can be used to estimate data quality and reliability, trace audit trail, repeat data derivation and set up attribution [19]. Given the wide range of application, provenance support has received increasing attention from researchers, especially in the field of eScience. In the myGrid project, Zhao et al. [20] utilized ontologies to annotate provenance data of biological experiments and link the data via inference over associated concepts. They proposed the COHSE open hypermedia system for building dynamically generated web of provenance documents, data, services, and workflows based on the annotated and linked provenance data. The Multi-Scale Chemical Science [21] (CMCS) project develops a general-purpose infrastructure for collaboration across many disciplines. It also contains a provenance subsystem that supports configurable services for extracting metadata, dynamically inferring additional

relationships, and browsing provenance. Simmhan et al. presents a survey of provenance systems used in eScience projects in [19].

3. Method of Procedure

In this capture, we discuss the challenges of building environmental informatics portals and then present our semantic technology-based approach to solve these challenges from three aspects: domain knowledge modeling and reasoning, integrating environmental data, and provenance support.

3.1 Domain Knowledge Modeling and Reasoning

3.1.1 Challenges

Environmental informatics systems are involved with 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 sampling 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. [22] propose the SWEET ontology family for earth system science. Chen et al. [15] model relationships among water quality datasets. Chau [14] models a specific

---

[*] Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "A Semantic Portal for Next Generation Monitoring Systems," in *Proceedings of the 10th International Semantic Web Conference*, Bonn, Germany, 2011, pp. 253-268.
Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring," TWC RPI, Troy, NY, 2011.
Portions of this chapter previously appeared as: J. G. Zheng, P. Wang, E. W. Patton, T. Lebo, J. Luciano, and D. L. McGuinness, "A Semantically-Enabled Provenance-Aware Water Quality Portal," presented at the Environmental Information Management Conference 2011, Santa Barbara, CA, 2011.

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.

Table 3.1 Subset of Contaminant Thresholds.

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

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) [23] 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 [24] while the threshold for Antimony is 0.006 mg/L according to the Drinking Water Protection Program [25] from the New York Department of Health. The water regulations are diverse in that they define different sets of contaminants with different contaminant thresholds. We generate a comparison table of contaminant thresholds at federal and state levels in [26]. The subset of the comparison table is shown in Table 3.1. We need an interoperable model that represents a diverse collection of regulations together with the domain knowledge and observational data 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.

3.1.2 Ontology Design and Semantic Reasoning

As we discussed, we need to develop ontologies for SemantEco. Developing ontologies is a process that attempts to encode domain knowledge with languages that machines can understand. Such practice may involve with many factors: the intended usage of the ontology, how domain knowledge can be obtained, the resources available, etc. Although there is no straightforward methodology for ontology design, Noy and McGuinness [27] give some fundamental rules and practical guidance in ontology design such as asking competency questions to determine the scope of the ontology, reusing existing ontologies to increase interoperability and taking ontology development as an iterative process.

To model domain knowledge in environmental information systems, we designed a set of ontologies: an upper ontology[3] that defines the basic terms for environmental monitoring, domain ontologies[4] that extend the upper ontology to model domain specific terms and regulation ontologies[5] that include terms required for describing compliance and pollution levels.

3.1.2.1 Upper Ontology

The upper ontology models the objects and their relationships in the domain of environmental monitoring with the following major classes and their properties.

Measurement: The class is imported from SWEET 2.1 [22] and is used for modeling observational data. We extend the class by adding properties including: hasCharacteristic that represents the characteristic measured, hasValue that represents the measured value and hasUnit that represents the unit of the measurement. hasUnit is also imported from SWEET 2.1.

Point: The class is imported from W3C SWIG Geo vocabulary [28]. It has properties corresponding to latitude and longitude to capture location data.

---

[3] http://escience.rpi.edu/ontology/semanteco/2/0/pollution.owl#
[4] http://escience.rpi.edu/ontology/semanteco/2/0/water.owl#
[5] http://purl.org/twc/ontology/swqp/region/ny; others are at http://purl.org/twc/ontology/swqp/region/

MeasurementSite: The class describes sites where measures are taken. Two important properties of MeasurementSite are: hasMeasurement that links measurements with its source site and hasUsage that describes the usage of the site. To capture the location of the site, MeasurementSite is a subclass of Point.

RegulationViolation: The class represents measurements that violate some regulation rule.

Facility: The class describes the facilities regulated by government agencies. It is a subclass of MeasurementSite.

PollutedThing: The class represents things that are polluted. It is defined as something that has at least one measurement that is a RegulationViolation.

PollutedSite: The class describes polluted measurement sites and is defined as the intersection of PollutedThing and MeasurementSite.

PollutingFacility: The class represents facilities that have violated regulatory requirements concerning pollution. It is defined as the intersection of PollutedThing and Facility.

A subset of the ontology is illustrated in Figure 3.1.

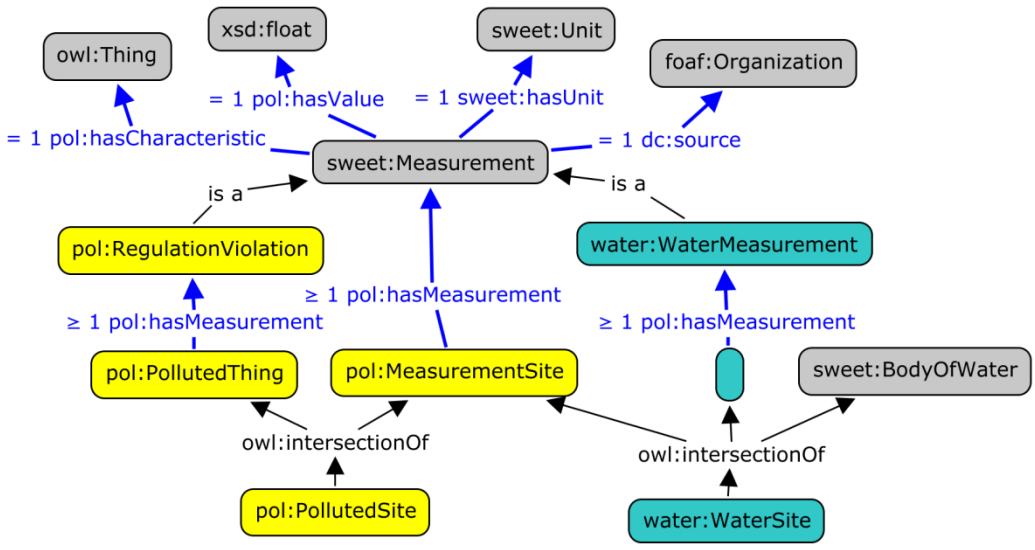

Figure 3.1 Portion of the TWC Environment Monitoring Ontology.

3.1.2.2 Water Ontology

There are various subjects of environmental monitoring, e.g. water quality, air quality, endangered animals, and deforestation. Each of these subjects is involved with its own domain specific knowledge. While the upper ontology is designed for capturing the general concepts in environmental monitoring, it can be extended to different fields by adding domain specific classes.

We take the field of water quality monitoring as an example, and develop the water ontology to encode the domain objects and their relationships with the following classes and their properties.

WaterMeasurement: The class is a subclass of Measurement, and represents measurements of one characteristic about a water sample.

WaterSite: The class is a subclass of MeasurementSite and represents sites where water quality was measured. It is also a subclass of BodyOfWater, which is defined in SWEET 2.1.

WaterFacility: The class is a subclass of Facility and represents facilities where water quality data are collected and regulated by EPA or other agencies.

Our water quality extension is also shown in Figure 3.1.

3.1.2.3 Regulation Ontology

To enable an environmental monitoring system to detect regulation violations, we need to model environmental regulations. Most rules in environmental regulations are defined by giving the allowable range for the concentration/quantity of a characteristic. For example, EPA's National Secondary Drinking Water Regulations contain the rule that the allowable pH value in drinking water is 6.5 - 8.5. Such rules can be encoded as OWL classes via numeric range restrictions on datatype properties provided by OWL 2. The rule-compliance results are reflected by whether an observational data item is a member of the class mapped from the rule or not.

Figure 3.2 shows the OWL class for the rule from EPA's NPDWRs that drinking water is polluted if the concentration of Arsenic is more than 0.01 mg/L. As we can see, the ArsenicDrinkingWaterRegulationViolation class is defined as the set of water measurements with greater than or equal to 0.01 mg/L of Arsenic concentration using

properties hasCharacteristic, hasValue and hasUnit. To connect with the upper and water ontology, the ArsenicDrinkingWaterRegulationViolation class is specified as a subclass of RegulationViolation and WaterMeasurement.

Regulations in other environment domains can be similarly mapped if they represent violations as ranges of measured characteristics.

After we design the schema of the regulation ontology, we can automate or semi-automate the process of encoding the water regulations as OWL classes. We developed a regulation converter which performs the following two steps. First, it transforms regulation data embedded in web pages into CSV files calling python scripts. Next, it encodes the intermediate CSV files into OWL classes that align with the upper and domain ontologies with java code.

The same workflow can be used to obtain the remaining state regulations using our converter if the regulation data are in the same format. If the data are in different formats, we need to update our converter in order to process the data. The current version of our converter can extract regulation data from web pages. Regulation data in more complex formats like PDF requires manual data extraction.

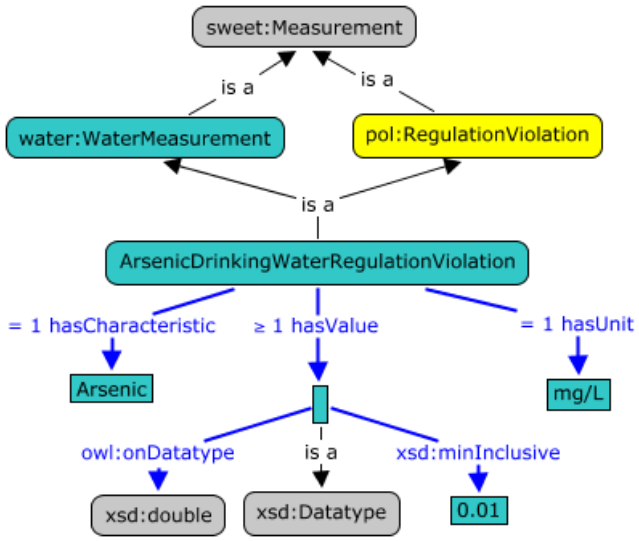

Figure 3.2 Portion of the EPA Regulation Ontology.

3.1.2.4 Reasoning Domain Data with Regulations

Combining observational data items collected at monitoring sites with the domain and regulation ontologies, an OWL2 reasoner is able to detect polluting facilities and water sites, and also the corresponding water measurements that violate the regulations.

For example, as we discussed in the previous subsections, in the upper ontology a polluted site is modeled as a measurement site that has at least one regulation violation, while the regulation ontology encodes rules such as any water measurements with greater than or equal to 0.01 mg/L of Arsenic concentration is an instance of ArsenicDrinkingWaterRegulationViolation. With these two pieces of knowledge, an OWL2 reasoner can automatically classify measurement sites as polluted sites or not with respect to the concentration of Arsenic.

Our ontology design provides several benefits.

First, the upper ontology is light weight: it consists of only 7 classes, 4 object properties, and 10 data properties. The development and maintenance cost for such light weight ontology is low.

Secondly, the ontology design is extensible. The upper ontology can be extended to other domains, e.g. air quality[6], soil geochemistry, and biodiversity. Regulation ontologies can be extended to encode regulations from various domains as long as the regulations specify violations as ranges of measured values for characteristics.

Thirdly, the design leads to flexible querying and reasoning: the user can select the ontologies to apply to the data and the reasoner will classify using only the selected ontologies. For example, a Rhode Island resident can choose to apply California regulations to his/her hometown by choosing to use the regulation ontology corresponded with the California regulations.

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

3.2 Integrating Environmental Data

3.2.1 Challenges

Government agencies, research labs, and citizen scientists all collect and publish environmental data on the web. For the purpose of water quality monitoring, we can get data from both the EPA and the US Geological Survey (USGS).

EPA Data: Permit compliance and enforcement status of facilities is regulated by EPA's National Pollutant Discharge Elimination System (NPDES) under the Clean Water Act (CWA). The NPDES water data include the descriptions of the facilities (e.g. name, permit number, address, and geographic location), measurements of contaminants in the water discharged by the facilities for up to five test types per contaminant, also the threshold values for these test types. The five test types are: concentration minimum, concentration average, concentration maximum, quantity average, and quantity maximum.

The descriptions of the facilities can be queried at EPA's Enforcement & Compliance History Online (ECHO) system. The facility data also can be downloaded as an Integrated Data for Enforcement Analysis (IDEA) dataset.

The measurements of contaminants can be accessed with EPA's ECHO system. However, if one needs to obtain a large amount of water measurement data, it is suggested by EPA to request the data with a Freedom of Information Act (FOIA) request.

USGS Data: USGS provides its water quality data through its National Water Information System (NWIS). The NWIS water quality data includes the descriptions of the water data collection stations (e.g. site id, address, and geographic location) and measurements of substances contained in water samples.

These water quality data are presented in tabular form: each column represents a property; each row corresponds to a record, and each table cell contains the value of the corresponding property in the corresponding record.

Although these datasets are organized to some extent, it is not easy to integrate them into one information system due to the following reasons:

(1) Branches from a common dataset can have big syntactic differences. For instance, EPA's NPDES has two branches: the Integrated Compliance Information

System (ICIS) is used by 49 states/districts, while the Permit Compliance System (PCS) is used by other 22 states/districts. While the two subsystems manage very similar water quality data, they use very different the file formats and layout for their data. What's more, if we fetch the water quality data through a web service, the data then is returned as HTML pages, which leads to more syntactic differences. Data in different formats need to be processed before they can be readily consumed by the information system.

(2) One concept may have multiple local names, and the local names may not be easily understood by data consumers unless they are the data providers at the same time. For example, the notion "name of measured characteristic" is represented by "CharacteristicName" in the USGS datasets and "Name" in the EPA datasets.

(3) We observe a need for linking data to domains other than water quality monitoring. Some popular concepts, e.g. names of chemicals, may be used in domains such as health care and food safety, and it would be useful to link them to other accepted models such as chemical element descriptions, e.g. ChemML.

(4) The semantics of the water quality data are not explicitly encoded in the data files. While it is possible to figure out the meaning of simple data objects, it can be very difficult to determine the semantics of complex data objects. For example, Table 3.2 shows a table fragment from the EPA ECHO measurement dataset, where four table cells in the first two columns together form 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". An information system cannot interpret such data correctly unless either the system is endowed with some prior knowledge about the data, or the semantics of the data is explicitly encoded as part of the data.

Table 3.2 Example of Complex Data Objects.

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

3.2.2 Data Integration Methodology

When integrating real world data from multiple sources, environmental monitoring systems can benefit from adopting the data organization and conversion capabilities enabled by the TWC-LOGD portal [29]. In the semantic water quality monitoring project, we used the open source tool csv2rdf4lod [7] to convert the data from EPA and USGS into Linked Data.

3.2.2.1 Integration Phases

The Tetherless World convertor tool named csv2rdf4lod can be used to produce RDF by performing four phases: catalog, retrieve, convert, and publish [29].

Creating a catalog for a dataset involves assigning two identifiers: the source identifier and dataset identifier. The source identifier is for identifying the source organization providing the dataset, the dataset identifier is for identifying the particular dataset being converted. After we have the identifiers, we make a local directory structure for each dataset as follows: base-dir/source/<source_identifier>/<dataset_identifier>.

Retrieving a version of a dataset is usually done using purl.sh, which is one of the utilities provided by csv2rdf4lod for crawling data specified by a URL and capturing the provenance of the crawling (e.g. source URL, HTTP Post parameters, and download time) in PML2 [11]. In this phase, we also assign a version identifier for identifying the version (or release) of the dataset, which is usually the date when we start crawling the dataset. We make a local directory for each dataset version as follows: base-dir/source/<source_identifier>/<dataset_identifier>/version/<version_identifier>. This directory is called the conversion cockpit. In the conversion cockpit, we make two more subdirectories: source, which is for the raw data we get from data sources; manual, which is for the data that has been processed by us and the conversion configuration files.

Converting tabular source data into RDF is performed according to configuration parameters encoded using a conversion vocabulary. Using parameters instead of custom code to perform conversions allows repeatable, easily inspectable, and queryable transformations; provides more consistent results; and includes a wealth of metadata. We

can apply multiple configurations of parameters to one dataset version, and generate multiple conversion layers. We can differentiate these conversion layers by conversion identifiers. Conversion creates more subdirectories in the conversion cockpit: automatic, which is for converted data, and publish, which is for the data dump to be published and the scripts that load data into triple store and publish data on the web.

Publishing phase is for hosting dump files on our web servers, loading data into a triple store, and exposing data as Linked Data via query end point.

3.2.2.2 Data Organization Model

The integration phases require us to create identifiers for data source, dataset, dataset version and conversion layer. The four identifiers are an important part of our data organization model. One more ingredient is a base URI, which is usually the hostname of a web server. With the following syntax, the dataset source, dataset, data version and conversion layer get its own URI.

<source_uri> ::= <base_uri>/source/<source_identifier>

<dataset_uri> ::= <source_uri>/dataset/<dataset_identifier>

<version_uri> ::= <dataset_uri> /version/<version_identifier>

<conversion_uri> ::= <version_uri>/conversion/<conversion_identifier>

3.2.3 Data Integration in SemantAqua

Next, we introduce how we perform data integration according to the integration phases and data organization model specified by csv2rdf4lod.

3.2.3.1 Catalog Phase

We create an inventory of the datasets. According to the data organization model, we assign two identifiers for each dataset: source identifier, and dataset identifier. Firstly, we have two data sources for water quality data: EPA and USGS, and we use “epa-gov” and “usgs-gov” as their source identifiers. Secondly, we have six types of datasets: facility data from EPA's ECHO system and FOIA request, measurement data from EPA's ECHO system and FOIA request, and water site and measurement data from USGS's NWIS system. For the six types, we use the following dataset prefix: “echo-facilities”, “foia-facilities”, “echo-measurements”, “foia-measurements”, “nwis-sites”

and “nwis-measurements”. The remaining part of the dataset identifier is a two-character abbreviation of state name. We divide the water quality data for the whole country by state this way, because the size of data for the whole country is too large to put in one dataset. Take Rhode Island as an example, we make the following local directory structure for the water quality data as a result of the catalog stage.

base-dir/source/epa-gov/echo-facilities-ri

base-dir/source/epa-gov/echo-measurements-ri

base-dir/source/usgs-gov/nwis-sites-ri

base-dir/source/usgs-gov/nwis-measurements-ri

3.2.3.2 Retrieve Phase

To get the EPA facilities, we access the IDEA service of the ECHO system and download two compressed folders: one for ICIS and one for PCS. Each compressed folder contains several TXT files and one of them is the description data for the EPA facilities.

As the downloaded data does not contain measurements of contaminants, we go to EPA's ECHO system to fetch water measurements. The ECHO measurements are organized by permit, i.e. it provides the water measurements for one facility permit in a file. As one state can contain thousands of facility permits, we need to crawl thousands of files for the EPA measurements of one state. To achieve this, we wrote a bash script 2source-echo-measurements.sh to query the web interface of ECHO. The script takes two parameters: the version identifier and the abbreviation of a state. The script executes the following four steps.

(1) It goes into the conversion cockpit of the dataset version specified by the input, e.g. source/epa-gov/echo-measurements-ri/version/2011-Mar-19. Note that the version identifier of the data snapshot is 2011-Mar-19, which is the date that we started to fetch the data.

(2) In the manual directory, we put the file that contains the facility data, from which the script reads the record of the facilities row by row.

(3) For each row it reads, it checks if the state of the facility equals to the given state.

(4) For each facility located in the given state, it invokes another script pcurl.sh to crawl the measurement file for the facility and put the file into the source directory. pcurl.sh allows us to specify the name of the crawled file with "-n" and the postfix of the crawled file with "-e". It also supports submitting POST requests with "-F variable=value". For example, the call of pcurl.sh to get the measurement data for the facility with permit RI0100005 is as follows:

pcurl.sh http://www.epa-echo.gov/cgi-bin/effluentdata.cgi -F "permit=RI0100005" -F "hits=1" -n RI0100005 -e csv

After we crawled the measurement data for four states (namely CA, MA, NY, RI), our programmatic queries of the EPA dataset were blocked. From our communication with the EPA, we were told that the ECHO system was designed to be used by human users not software agents and we should file a FOIA request to get the bulk water data. We then filed a FOIA request and received the water data for the remaining states via both DVDs and online storage.

The water datasets from USGS are organized by county, i.e. the description of the water sites located in one county is in one file and the measurements taken at these water sites are in other file.

As one state can have hundreds of counties, we also wrote a bash script 2source-nwis-sites.sh to crawl the data from USGS. The script takes two parameters: the version identifier and the Federal Information Processing Standard (FIPS) code of a state. The script executes the following steps. We use the workflow of getting the site data for Bristol, RI as an example.

(1) It goes into the directory for the dataset version specified by the input, e.g. base-dir/source/usgs-gov/nwis-sites-ri/version/2011-Mar-20. The version identifier of the data snapshot is 2011-Mar-20, which is the date that we started to fetch the data.

(2) It invokes pcurl.sh to get a XML file that contains the FIPS codes for the counties in the given state from a web service provided by USGS, and put the XML file to the source directory. For our example, after this step, we get US-44-county-code.xml and US-44-county-code.xml.pml.ttl in base-dir/source/usgs-gov/nwis-sites-ri/version/2011-Mar-20/source.

(3) It invokes the python script extract-county-code.py to extract the FIPS codes from the XML file and put the extracted data in a TXT file in the manual directory. In our example, we produce US-44-county-code.txt in base-dir/source/usgs-gov/nwis-sites-ri/version/2011-Mar-20/manual.

(4) It invokes justify.sh to encode the provenance of the extraction, i.e. the TXT file is generated by parsing field of the XML file. This step produces a provenance file e.g. US-44-county-code.txt.pml.ttl in base-dir/source/usgs-gov/nwis-sites-ri/version/2011-Mar-20/manual.

(5) It reads the FIPS codes of the counties from the TXT file.

(6) For each county in the given state, it invokes pcurl.sh to crawl the file for the county into the source directory. The call of pcurl.sh to get the site data for Bristol is as follows:

pcurl.sh "http://qwwebservices.usgs.gov/Station/search?statecode=US:44&countycode=001&mimeType=csv&zip=yes" -n US-44-001-site -e "zip"

From this step, we get US-44-001-site.zip and US-44-001-site.zip.pml.ttl in base-dir/source/usgs-gov/nwis-sites-ri/source.

(7) The crawled file is compressed, so the script calls punzip.sh to uncompress the file. punzip.sh is another tool provided by csv2rdf4lod for uncompressing and capturing the provenance of the uncompressing (e.g. name of the original compressed file, and uncompress time). From this step, we get US-44-001-site.csv and US-44-001-site.csv.pml.ttl in base-dir/source/usgs-gov/nwis-sites-ri/version/2011-Mar-20/source.

For water measurements, we wrote the script 2source-nwis-measurements.sh. This script accepts the same parameters as 2source-nwis-sites.sh and performs almost the same steps as 2source-nwis-sites.sh. The only difference is that the two scripts call two different water quality web services of USGS: one is for getting site information and the other is for getting measurements. The call of pcurl.sh to get the measurement data for Bristol is as follows:

pcurl.sh "http://qwwebservices.usgs.gov/Result/search?statecode=US:44&countycode=001&mimeType=csv&zip=yes" -n US-44-001-result -e "zip"

3.2.3.3 Pre-process Phase

Before converting the datasets with csv2rdf4lod, we need to preprocess them due to two factors. (1) Some datasets may contain incomplete or inconsistent data. (2) Some datasets may require additional domain information to be interpreted properly.

Firstly, the facility data from the IDEA datasets contains incomplete information in that the records of some facilities have addresses but no valid latitude and longitude values, while the records of some other facilities have latitude and longitude values, but no valid addresses. One way to fix incompleteness or inconsistency of the datasets is to leverage data from additional sources. In our case, we use the Google Geocoding service to get extra location data. For the former type of the incomplete facility records, we call the Google Geocoding Service to get the latitudes and longitudes from the facility addresses. For the latter type, we invoke the Google Reverse Geocoding Service to obtain the facility addresses from the geographic coordinates.

Next, as a result of our FOIA request, the water measurement data of EPA's PCS system are sent to us as a flat TXT file. We preprocess this file as follows.

(1) In this file, data properties such as characteristic name and unit are encoded. Thus, we need to look up two additional tables to translate the codes into human-readable names.

(2) Each measurement has two units: one is the default unit and the other is called the reported unit. When the reported unit is absent, both the measured value and the threshold value are under the default unit. When the data object has a reported unit, the measured value is under the reported unit while the threshold value is under the default unit. On such occasions, we need to do unit conversion for the measured value, i.e. multiply the measured value with a conversion rate. EPA provides us with a table for finding the conversion rate between two units.

(3) While the measurement objects from the ICIS branch include comparison operators, the measurement objects from PCS have no comparison operators with them. We need to infer the comparison operator of a data object from its test types and override flag. For the test type of concentration minimum, the default comparison operator is >=, i.e. the value should be larger than or equal to the threshold. But when the override flag of concentration minimum is set, the comparison operator is flipped

into <=, i.e. the value should be less than or equal to the threshold. Conversely, for the other four test types (concentration average, concentration maximum, quantity average, and quantity maximum), the default comparison operator is <=. When the override flag of concentration average is set, the comparison operator is flipped into >=. Note that not all the test types have override flags. Only concentration minimum and concentration average have these flags.

3.2.3.4 Conversion Phase

At the conversion stage, we create configurations and convert the dataset snapshots to conversion layers, which are our representations of the datasets. csv2rdf4lod provides the basic conversion configuration and can automatically generate the corresponding raw layer, in which each table cell is converted into a triple in the form of (current_row, current_column, cell_value). While the raw layer minimizes the need for human involvement for data conversion, csv2rdf4lod supports user-contributed conversion configurations. By conducting the enhanced conversion, we can resolve the issues posed by the heterogeneous datasets and tailor the converted data according to the needs of the environmental information system.

File format: The default input format of the converter is CSV. While the majority of our datasets are in this format, the dataset of the ECHO facility is a well formatted TXT file, in which each row describes one facility and the cells are separated by "|". Fortunately, the converter supports multiple delimits as long as the input file is well formatted. After we specify the character that delimits cells as "|" in the enhancement configuration, the converter can separated the cells properly.

Data type: While the values of the cells are interpreted as literals by default, the converter allows us to specify the data types of some property ranges so that the values of the domain objects are better modeled. For instance, we specify that the data type of the water measurement values is xsd:decimal. With the explicit data type, we can use the Pellet reasoner [30] to enforce the numerical range constraints from the water regulations.

Linking to ontological terms: One effective mechanism for linking heterogeneous datasets is through reusing 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 predicate pol:hasCharacteristic. We use “pol” as the namespace prefix of our upper ontology, and “water” as the namespace prefix of our water ontology. Similarly, we map spatial location properties, such as “LatitudeMeasure” and “LongitudeMeasure” in the USGS dataset and “FCLGLAT” and “FCLGLON” in the EPA dataset, to predicate from an external ontology, e.g. wgs84:lat and wgs84:long.

Table 3.3 Example of Cell Based Conversion.

Configuration snippet	Conversion result
conversion:enhance [ov:csvCol 20; ov:csvHeader "C1_VALUE"; a scovo:Item; conversion:label "Test type"; conversion:object "[/sd]typed/test/C1";]; conversion:enhance [ov:csvCol 21; ov:csvHeader "C1_UNIT"; conversion:bundled_by [ov:csvCol 20] ;];	:measurement_469_20 rdf:type water:WaterMeasurement ; pol:hasPermit FacilityPermit-RI0100005 ; pol:hasCharacteristic pol:Coliform_fecal_general ; dcterms:date "2010-09-30"^^xsd:date ; e1:test_type typed-test:C1 ; rdf:value "34.07"^^xsd:decimal ; reprSciUnits:hasUnit "MPN/100ML" .

Aligning instance references: We promote references to chemicals in our water quality data from literal to URI, e.g. “Arsenic” is promoted to “pol:Arsenic”, which then can be linked to external resources like “http://dbpedia.org/resource/Arsenic” using owl:sameAs. This design choice is made based on the observation that not all chemical names can be directly mapped to DBpedia URI (e.g., “Nitrate/Nitrite” from Massachusetts water regulations [31] maps two DBpedia URIs), and some instances may not be defined in DBpedia (e.g., “C5-C8” from Massachusetts water regulations). By linking to DBpedia URIs, we reserve the opportunity to connect to other knowledge base such as disease databases. In another case, we promote the permit of a facility to URI to

connect the water measurements to their facilities. If a water measurement and a facility point to the same permit, the measurement is for the facility.

Cell based conversion: The default conversion is row-based conversion in that the converter only creates subjects out of row items. However, there are occasions that we need to compose a complex data object from multiple cells in a table. One example is given in Table 3.2. In order to properly interpret such measurement data, we switch from row-based conversion to cell-based conversion, 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. We name the subject made out of a cell as cell subject. For instance, to convert the first two columns in Table 3.2, we mark the first column (column 20 of the original CSV file) as "scovo:Item" , then the converter creates a new subject measurement_469_20 out of the cell and the cell value is attached to the subject with the predicate "rdf:value". Meanwhile, we use "conversion:bundled_by" to associate the other columns (e.g. column 21) to the cell subject. The left column in Table 3.3 shows the configuration snippet and the right column shows the conversion result.

3.2.3.5 Publish Phase

We take the dataset "nwis-measurements-ri" as example to illustrate the steps of the publish stage.

Firstly, we publish RDF dump files on the web for the dataset so that anyone on the web can access our RDF data. For "nwis-measurements-ri", we publish the RDF file: http://sparql.tw.rpi.edu/source/usgs-gov/file/nwis-measurements-ri/version/2011-Mar-20/conversion/usgs-gov-nwis-measurements-ri-2011-Mar-20.ttl.gz.

To enable web users to repeat the data conversion, we also publish source files and conversion configuration files such as the following two files.

http://sparql.tw.rpi.edu/source/usgs-gov/provenance_file/nwis-measurements-ri/version/2011-Mar-20/source/US-44-001-result.csv

http://sparql.tw.rpi.edu/source/usgs-gov/provenance_file/nwis-measurements-ri/version/2011-Mar-20/manual/US-44-001-result.csv.e1.params.ttl

Next, we load the converted data in our triple store as a named graph: http://sparql.tw.rpi.edu/source/usgs-gov/dataset/nwis-measurements-ri/version/2011-Mar-20.

Lastly, we expose the data via a query interface. So the user can access the dataset in RDF format by querying the interface http://sparql.tw.rpi.edu/virtuoso/sparql.

3.3 Provenance

The information system contains provenance data from two sources. (1) Provenance metadata can be embedded in the original datasets, e.g. measurement location and time. (2) The system automatically captures provenance data during the data integration phases and encodes them in PML2 [11] due to the provenance support from csv2rdf4lod and our regulation converter.

At the retrieval phase, csv2rdf4lod captures provenance, e.g. the URL of the data source, who fetches the source data at what time, and what agent and protocol are used for retrieving the data. At the conversion phase, it keeps provenance, e.g. what engine performs the conversion, what antecedent data are involved, and what roles those data play. At the publication phase, it captures provenance such as who loads the data to the triple store at what time.

Our regulation converter captures the following provenance, e.g. the URL of the regulation source, who fetches the regulation at what time, and who converts the regulation at what time.

Provenance data can be used for a variety of purposes such as increasing data reliability, improving transparency and thus potentially increasing trust from users, and giving credit to data contributors [19]. In our work, we utilize provenance in three ways. To make our portal more transparent, we display data lineage using a pop up window when the user selects a measurement site or facility. We also use provenance data to enable dynamic data source listing and provenance-aware cross validation over EPA and USGS data.

3.3.1 Data Lineage

While the system captures, encodes and stores provenance data, the user still does not have easy access to the provenance data. Thus, we developed the provenance feature that reveals lineage of the water quality datasets via a pop up window. As we have the provenance data in the triple store, we can write SPARQL [32] queries to retrieve them. For example, the following query fetches provenance including the inference engine, the URL of the data source and download time of the data. After the provenance is retrieved, we format and render them as HTML pages.

```
PREFIX pmlp: <http://inference-web.org/2.0/pml-provenance.owl#>
PREFIX pmlj: <http://inference-web.org/2.0/pml-justification.owl#>
PREFIX rlinput: <http://inference-web.org/registry/ROLE/Input.owl#>
Select distinct ?infEngine ?inf3source ?inf3dateTime
Where {
Graph <http://sparql.tw.rpi.edu/source/usgs-gov/dataset/nwis-measurements-ri/version/2011-Mar-20> {
 ?node1 pmlj:hasConclusion <http://sparql.tw.rpi.edu/source/usgs-gov/file/nwis-measurements-ri/version/2011-Mar-20/conversion/usgs-gov-nwis-measurements-ri-2011-Mar-20.ttl.gz>.
 ?node1 pmlj:isConsequentOf ?inf1.
 ?inf1 pmlj:hasInferenceEngine ?infEngine.
 ?inf1 pmlp:hasAntecedentRole ?atRoles1.
 ?atRoles1 pmlp:hasAntecedent ?antecedent1.
 ?atRoles1 pmlp:hasRole rlinput:Input.
 ?node2 pmlj:hasConclusion ?antecedent1.
 ?node2 pmlj:isConsequentOf ?inf2.
 ?inf2 pmlj:hasSourceUsage ?inf2sourceUsage.
 ?inf2sourceUsage pmlp:hasSource ?inf2source.
 ?node3 pmlj:hasConclusion ?inf2source.
 ?node3 pmlj:isConsequentOf ?inf3.
 ?inf3 pmlj:hasSourceUsage ?inf3sourceUsage.
 ?inf3sourceUsage pmlp:hasUsageDateTime ?inf3dateTime.
```

```
  ?inf3sourceUsage pmlp:hasSource ?inf3source.
}}
```

3.3.2 Data Source as Provenance

We can utilize provenance information about data sources to support dynamic data source listing as follows.

(1) 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.

(2) The system loads additional water quality data into the system as a new RDF graph.

(3) After the new graph is loaded, the system updates the DS graph to add the metadata of the new graph.

(4) The system tells the user what data sources are currently available by executing a SPARQL query on the DS graph to obtain distinct data source URIs.

(5) With the presentation of the data sources on the interface, the user is allowed to select only the data sources he/she trusts. The system would then return results with data only from the selected sources.

Starting with the above example, we can further utilize provenance information to give the user more options to specify his/her data retrieval request, e.g. some users may be only interested in data released within a particular time period.

3.3.3 Provenance-Aware Cross-Validation

Our system can compare water quality data originating from different sources for the purpose of cross-validation, since we maintain the information of where each piece of data comes from.

Environmental data depends on the geographic location and the measurement time, thus only data measured at close locations and time could be used to validate each other. We can write SPARQL queries to express such requirements on data retrieval. For example, the filter feature of SPARQL can be used to find two locations close to each other as shown as below.

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

After we obtain comparable data, we performed the comparison between the EPA and USGS data and obtained interesting results. Figure 3.3 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 multiple times and went above the maximum value from EPA once.

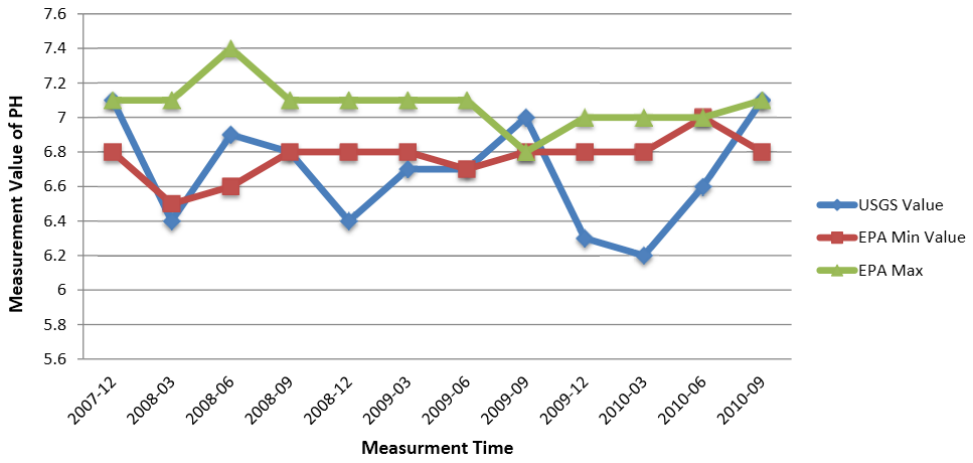

Figure 3.3 Data Validation Example.

4. Results

In this chapter, we describe the implementation details of our exemplar environmental informatics portal, including its system architecture and workflow. We also present the scaling issues of the water quality portal SemantAqua.

4.1 System Architecture and Components

The system architecture of the SemantAqua is illustrated in Figure 4.1. The system comprises six major components: (1) ontology, (2) data conversion, (3) storage, (4) reasoning, (5) provenance and (6) visualization.

Ontology Component: The ontology component is for capturing the domain knowledge. We develop our ontologies in OWL2 [10] using the Protege ontology editor and knowledgebase framework.

Data Conversion Component: This component is composed of two tools, namely csv2rdf4lod which converts water quality data into RDF triples and our regulation converter which converts water regulations into OWL classes.

Storage Component: The RDF data are stored in OpenLink's Virtuoso 6 open source community edition triple store, which includes a web-accessible endpoint that answers SPARQL [32] queries from web clients.

Reasoning Component: We utilize the Pellet OWL Reasoner [30] together with the Jena Semantic Web Framework [33] to reason over the data and ontologies in order to identify water pollution.

Provenance Component: The provenance Component refers to the provenance capture support that the system gets from the converters and the provenance

---

[*] Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "A Semantic Portal for Next Generation Monitoring Systems," in *Proceedings of the 10th International Semantic Web Conference*, Bonn, Germany, 2011, pp. 253-268.
Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring," TWC RPI, Troy, NY, 2011.
Portions of this chapter previously appeared as: J. G. Zheng, P. Wang, E. W. Patton, T. Lebo, J. Luciano, and D. L. McGuinness, "A Semantically-Enabled Provenance-Aware Water Quality Portal," presented at the Environmental Information Management Conference 2011, Santa Barbara, CA, 2011.

applications, including display data lineage, dynamic data source listing, and cross validation.

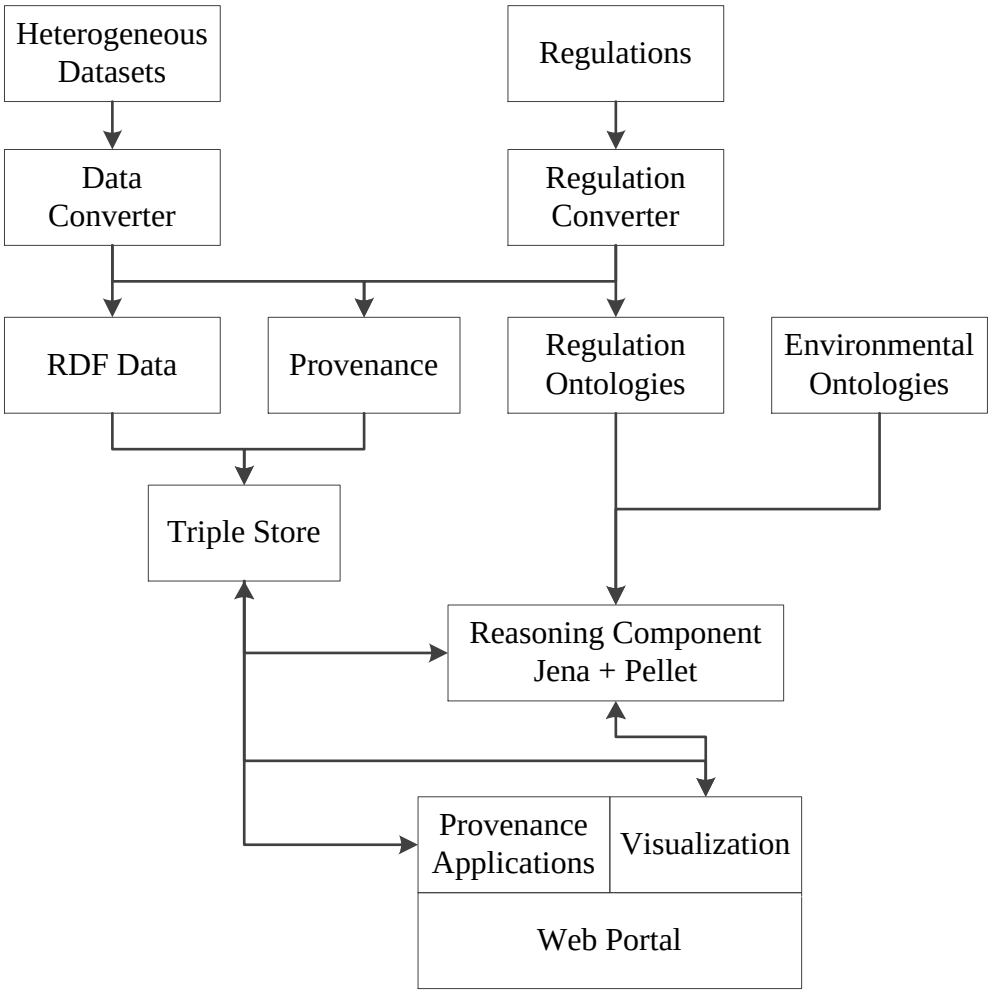

Figure 4.1 System Architecture and Workflow.

Visualization Component: This component is responsible for mashing up and representing the data collected from various sources. The system supports two types of visualizations: (1) map visualization that displays the sources of the water pollution in the context of geographic regions and (2) time series visualization that depicts pollution levels over time with respect to a particular water source or facility.

The map visualization receives the reasoning results for a user query from the back-end reasoner and renders the results on a Google Map. Figure 4.2 shows a screenshot of the map visualization of SemantAqua. The user can specify 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), health concern (mark 7). We use different icons to distinguish between clean and polluted water sites, and between clean and polluting 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, time of measurement, etc. The window also provides a link to the time series visualization of the water quality data of the selected site. The results of applying the EPA federal water regulation on the region with the zip code 02888 (Warwick, RI) is visualized in this example. Two polluted water sources and eight polluting facilities are indicated with icons.

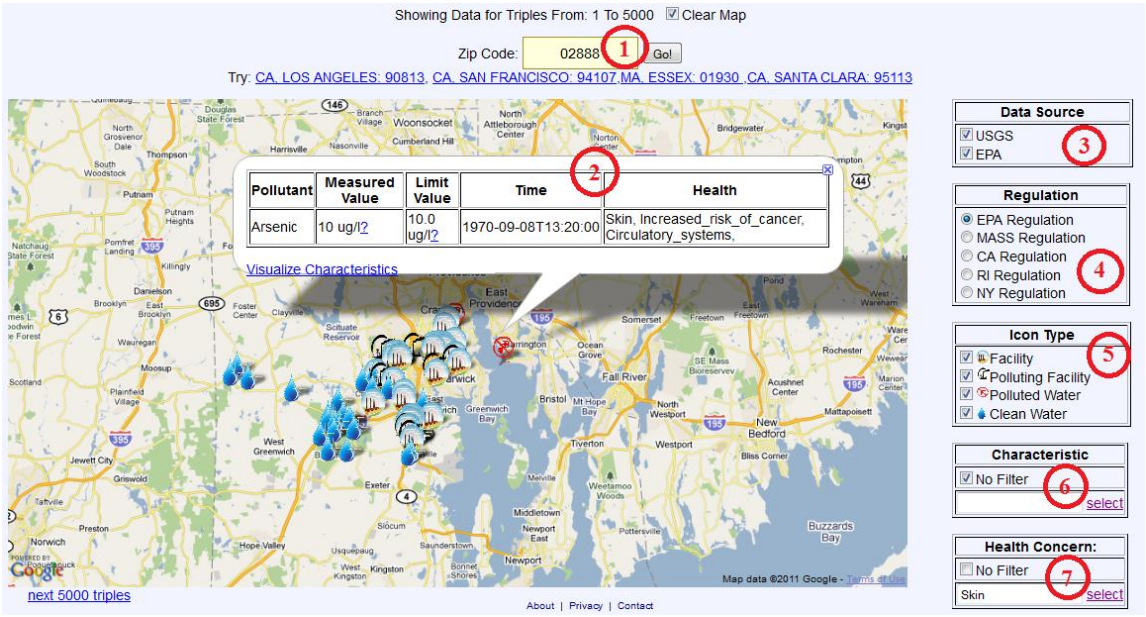

Figure 4.2 Map Visualization. The results of applying the EPA federal water regulations on the region with zip code 02888 is visualized on a Google Map.

The time series visualization fetches the water quality data about the selected water site or facility by querying the triple store and depicts the water quality data as a time series using the Protovis visualization toolkit. Figure 4.3 shows the phosphorus measurements from 2007 to 2010 in green and the regulation defined limit in blue. Note that the data show one violation in 2009 (in red) and no subsequent violations.

A live demo of the water quality portal can be found at [34].

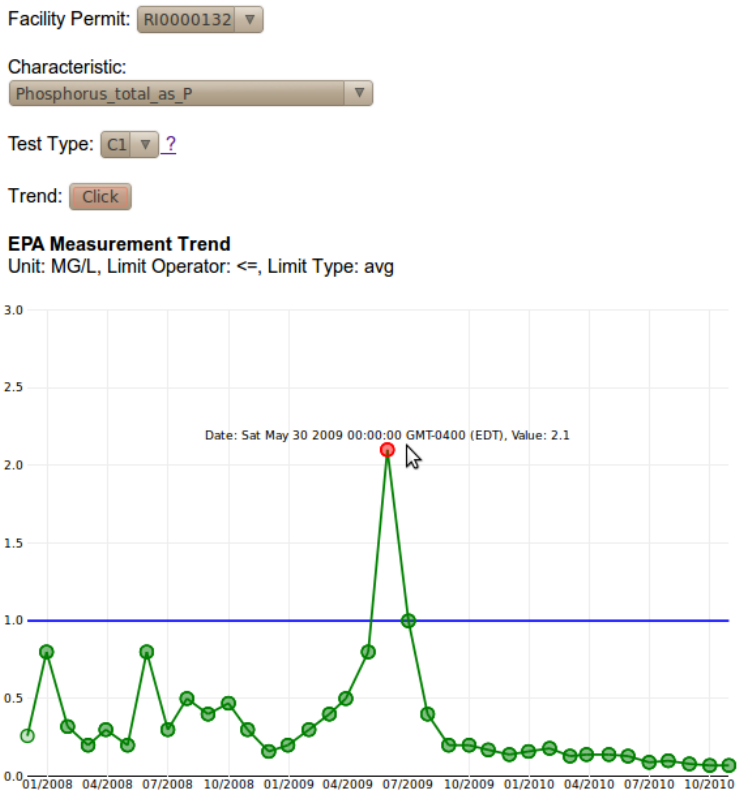

Figure 4.3 Time Series Visualization. The phosphorus measurements from 2007 to 2010 and the regulation defined limit for the selected facility are visualized.

4.2 System Workflow

We now introduce how the components work together to identify polluted water sources, polluting facilities, and pollution occurrences. The system workflow is also shown in Figure 4.1. Firstly, the system retrieves data from data providers like EPA, USGS and state regulation agencies, and converts the heterogeneous datasets into RDF using the data conversion component. During the data integration process, provenance information for the downloaded and converted data is captured. The system then loads the converted data into the storage component, which is a Virtuoso triple store. When the user accesses the front-end interface of the system and issues a request, the request is sent to the back-end reasoning component. The reasoning component then loads the pollution ontology, water ontology, and selected regulation ontology, retrieves water quality data from the storage component, and performs reasoning over the retrieved data based on the

ontologies. After the reasoning component completes its reasoning, the results are sent to the visualization component for user presentation.

Table 4.1 Number of Triples of EPA Data.

State	Number of triples from EPA
AK	24258901
AR	69453257
AS	1057286
CA	8515112
CO	53328833
CT	9185278
DC	4044800
GA	15631809
HI	6467289
ID	27946658
LA	72870574
MA	5508699
MD	83912342
NE	51737294
NH	15326905
NM	15576955
NV	3946323
NY	30607170
OK	19312028
PA	28416419
PR	29502305
RI	2862392
SD	15171908
UT	16269152
VI	4438257
WI	87279163

4.3 Scaling Issues

We tested our approach in a realistic setting and gathered data for about half the states. We have generated 495.83 million triples for the USGS datasets for 20 states, and 702.63 million triples for the EPA datasets for 26 states. These numbers imply that water data for all 50 states would generate on the order of billions of triples, which in turn suggests that a triple store cluster should be deployed to host the water data.

Table 4.2 Number of Triples of USGS Data.

State	Number of triples from USGS
AK	14426447
AS	14983
CA	79289481
CT	17449231
DC	451506
FL	67428800
ID	35340412
IN	16739709
MA	11327861
MD	18571620
MI	14642177
NC	26596915
NH	3262711
NY	69248384
OH	22924460
PA	51329766
RI	2158796
SC	6439567
VI	205160
WI	37979240

The sizes of the converted data are summarized in Table 4.1 and Table 4.2. We have obtained the data for all the 50 states and are working on integrating the data into the SemantAqua portal. We maintain the statistics about the water quality data at [35].

The numbers of the classes we generated for modeling the rules from the different regulations are given in Table 4.3. Our programmed conversion provides a quick and low cost approach for encoding regulations.

Table 4.3 Number of Threshold Classes Converted from Regulations.

EPA	CA	MA	NY	RI
83	104	139	74	100

5. Discussion

In this chapter, we discuss potential extensions and enhancements for the environmental informatics portal from three perspectives: linking to the health domain, approaches used to increase scalability, and observations about water regulation comparison.

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 [36], [37]. The reported symptoms range 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 need to model 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.

We obtained the health effects of contamination from NPDWRs, and designed a small health ontology which models the health effects with OWL classes, e.g. health:High_blood_pressure. We use the object property "hasHealthEffect" to connect the contaminants with their health effects. Then, we can query the health effects of water contamination with respect to a particular characteristic using the SPARQL query fragment below.

?violation pol:hasCharacteristic ?characteristic.

? characteristic health:hasHealthEffect ?effect.

---

[*] Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "A Semantic Portal for Next Generation Monitoring Systems," in *Proceedings of the 10th International Semantic Web Conference*, Bonn, Germany, 2011, pp. 253-268.
Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring," TWC RPI, Troy, NY, 2011.

While our health modeling is simplified, it enables the system to address health concerns to some extent: (1) the user can specify his/her health concern and the portal will detect only the water pollution that has been correlated to 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.

5.2 Scalability

The large number of triples generated during data conversion prohibits reasoning over the entire dataset in real time. Several approaches have been applied 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 four named graphs for each state to store the integrated data, i.e. one graph per dataset for each state.

However, the triple count at the state level is still quite large: we currently host 8.52 million triples from EPA and 79.29 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:WaterSite.
  ?s pol:hasMeasurement ?measurement.
  ?s water:hasStateCode ?state.
  ?s wgs84:lat ?lat.      ?s wgs84:long ?long.
  ?measurement pol:hasCharacteristic ?characteristic.
  ?measurement pol:hasValue ?value.
  ?measurement reprSciUnits:hasUnit ?unit.
  ?measurement time:inXSDDateTime ?time.
  ?s water:hasCountyCode 085. }
```

```
WHERE { GRAPH <http://sparql.tw.rpi.edu/source/usgs-gov/dataset/nwis-measurements-ca/version/2011-Mar-20>
{ ?s rdf:type water:WaterSite.
 ?s water:hasUSGSSiteId ?id.
 ?s water:hasStateCode ?state.
 ?s wgs84:lat ?lat.  ?s wgs84:long ?long.
 ?measurement water:hasUSGSSiteId ?id.
 ?measurement pol:hasCharacteristic ?characteristic.
 ?measurement pol:hasValue ?value.
 ?measurement reprSciUnits:hasUnit ?unit.
 ?measurement time:inXSDDateTime ?time.
 ?s water:hasCountyCode 085. }}
```

5.3 Regulation Mapping and Comparison

The majority of the portal domain knowledge stems from water regulations that stipulate contaminants, thresholds for pollution, 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, Table 3.1 compares regulations from five different sources and shows substantial variation.

By modeling regulations as OWL classes, we may 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 a subclass relation to model such rules in order to evaluate subsuming relationships. This could spare some reasoning time when multiple sets of regulations are applied to detect the pollution.

6. Conclusion

We presented a semantic technology-based approach for building environmental informatics portals and described our work using this approach in the Tetherless World Constellation SemantEco portal and the exemplar SemantAqua portal. We described the overall design, including: the design of the ontologies, the methodology for data integration, and the encoding and usage of provenance information generated during data integration. The SemantAqua portal demonstrates some benefits and potential of applying semantic web technologies to environmental information systems as follows. First, ontologies can be used to model domain knowledge and to enable machine reasoning over environmental data. Secondly, semantic data integration provides an effective, systematic, and low cost approach for aggregating data from multiple sources. Lastly, provenance information encoded using semantic web technology not only improves the transparency and trust of web portals, but also enables provenance-aware applications. In this research project, the individual work of the author includes: integrating the water quality data from heterogeneous sources, designing and implementing the provenance applications, scaling the application to go from four states to twenty, and developing the time series visualization of the data.

A number of extensions to this portal are in process. First, we are extending our SemantEco portal by applying our approach to other environmental topics, e.g. air quality, biodiversity, and health impacts of environmental pollution. 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

---

[*] Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "A Semantic Portal for Next Generation Monitoring Systems," in *Proceedings of the 10th International Semantic Web Conference*, Bonn, Germany, 2011, pp. 253-268.
Portions of this chapter previously appeared as: P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring," TWC RPI, Troy, NY, 2011.
Portions of this chapter previously appeared as: J. G. Zheng, P. Wang, E. W. Patton, T. Lebo, J. Luciano, and D. L. McGuinness, "A Semantically-Enabled Provenance-Aware Water Quality Portal," presented at the Environmental Information Management Conference 2011, Santa Barbara, CA, 2011.

in the immediate area. Lastly, we are expanding SemantAqua to support all 50 US states. Water quality data have been obtained from EPA and USGS. After we finish processing and converting these data, the portal could identify water pollution events in all the states according to federal water regulations, or other state regulations we have already encoded such as CA and RI.

7. References

[1] F. A. Batzias and C. G. Siontorou, "A knowledge-based approach to environmental biomonitoring," *Environmental monitoring and assessment*, vol. 123, no. 1-3, pp. 167-97, Dec. 2006.

[2] H. Scholten, A. Kassahun, J. C. Refsgaard, T. Kargas, C. Gavardinas, and A. J. M. Beulens, "A methodology to support multidisciplinary model-based water management," *Environmental Modelling & Software*, vol. 22, no. 5, pp. 743-759, May 2007.

[3] D. M. Holland, P. P. Principe, and L. Vorburger, "Rural Ozone: Trends and Exceedances at CASTNet Sites," *Environmental Science & Technology*, vol. 33, no. 1, pp. 43-48, Jan. 1999.

[4] Q. Liu, Q. Bai, L. Ding, H. Pho, Y. Chen, C. Kloppers, D. L McGuinness, D. Lemon, P. Souza, P. Fitch, and P. Fox, "Linking Australian Government Data for Sustainability Science - A Case Study," in *Linking Government Data*, D. Wood, Ed. New York, NY: Springer New York, 2011, pp. 181-204.

[5] F. Villa, I. N. Athanasiadis, and A. E. Rizzoli, "Modelling with knowledge: A review of emerging semantic approaches to environmental modelling," *Environmental Modelling & Software*, vol. 24, no. 5, pp. 577-587, May 2009.

[6] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1-22, 2009.

[7] T. Lebo and G. T. Williams, "Converting governmental datasets into linked data," in *Proceedings of the 6th International Conference on Semantic Systems*, Graz, Austria, 2010, pp. 38:1-38:3.

[8] T. J. Morgan, "Bristol, Warren, Barrington residents told to boil water," 2009. [Online]. Available: http://news.providencejournal.com/breaking-news/2009/09/residents-of-3.html. (Date Last Accessed on March 5, 2012).

[9] P. Pinheiro Da Silva, D. L. McGuinness, and R. McCool, "Knowledge Provenance Infrastructure," *IEEE Data Engineering Bulletin*, vol. 25, no. 2, pp. 179-227, 2003.

[10] P. Hitzler, M. Krötzsch, B. Parsia, P. F. Patel-Schneider, and S. Rudolph, "OWL 2 Web Ontology Language Primer," *W3C Recommendation 27 October 2009*, 2009. [Online]. Available: http://www.w3.org/TR/owl2-primer/. (Date Last Accessed on March 5, 2012).

[11] D. L. Mcguinness, L. Ding, P. P. D. Silva, and C. Chang, "PML 2: A Modular Explanation Interlingua," in *Proceedings of the AAAI 2007 Workshop on Explanation-aware Computing*, Vancouver, Canada, 2007, pp. 22 - 23.

[12] D. DiFranzo and L. Ding, “Clean Air Status and Trends - Ozone.” [Online]. Available: http://logd.tw.rpi.edu/demo/clean_air_status_and_trends_-_ozone. (Date Last Accessed on March 5, 2012).

[13] L. Ceccaroni, U. Cortes, and M. Sanchezmarre, “OntoWEDSS: augmenting environmental decision-support systems with ontologies,” *Environmental Modelling & Software*, vol. 19, no. 9, pp. 785-797, Sep. 2004.

[14] K. W. Chau, “An ontology-based knowledge management system for flow and water quality modeling,” *Advances in Engineering Software*, vol. 38, no. 3, pp. 172-181, Mar. 2007.

[15] Z. Chen, a Gangopadhyay, S. Holden, G. Karabatis, and M. Mcguire, “Semantic integration of government data for water quality management,” *Government Information Quarterly*, vol. 24, no. 4, pp. 716-735, Oct. 2007.

[16] A. Halevy, A. Rajaraman, and J. Ordille, “Data integration: the teenage years,” in *Proceedings of the 32nd International Conference on Very Large Data Bases*, Seoul, Korea, 2006, pp. 9-16.

[17] A. Doan and A. Y. Halevy, “Semantic Integration Research in the Database Community: A Brief Survey,” *AI magazine*, vol. 26, no. 1, pp. 83 - 94, 2005.

[18] L. Xu and D. W. Embley, “Using Domain Ontologies to Discover Direct and Indirect Matches for Schema Elements,” presented at the 2nd International Semantic Web Conference Semantic Integration Workshop, Sanibel Island, FL, 2003.

[19] Y. L. Simmhan, B. Plale, and D. Gannon, “A survey of data provenance in e-science,” *ACM SIGMOD Record*, vol. 34, no. 3, pp. 31-36, Sep. 2005.

[20] J. Zhao, C. Goble, R. Stevens, and S. Bechhofer, “Semantically Linking and Browsing Provenance Logs for E-science,” in *Proceedings of the 1st International Conference on Semantics of a Networked World*, Paris, France, 2004, pp. 158-176.

[21] J. Myers, C. Pancerella, C. Lansing, K. Schuchardt, and B. Didier, “Multi-scale science: supporting emerging practice with semantically derived provenance,” presented at ISWC 2003 Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data, Sanibel Island, FL, 2003.

[22] R. G. Raskin and M. J. Pan, “Knowledge representation in the semantic web for Earth and environmental terminology (SWEET),” *Computers & Geosciences*, vol. 31, no. 9, pp. 1119-1125, Nov. 2005.

[23] US EPA, “National Primary Drinking Water Regulations.” [Online]. Available: http://water.epa.gov/drink/contaminants/index.cfm. (Date Last Accessed on March 5, 2012).

[24] State of Rhode Island and Providence Plantations, "Water Quality Regulations," 2009. [Online]. Available: http://www.dem.ri.gov/pubs/regs/regs/water/h20q09.pdf. (Date Last Accessed on March 5, 2012).

[25] Department of Health New York State, "Part 5, Subpart 5-1 Public Water Systems - Tables." [Online]. Available: http://www.health.ny.gov/regulations/nycrr/title_10/part_5/subpart_5-1_tables.htm. (Date Last Accessed on March 5, 2012).

[26] J. G. Zheng and P. Wang, "Side-by-site comparison on water regulations from five different sources | Tetherless World Constellation." [Online]. Available: http://tw.rpi.edu/web/project/TWC-SWQP/compare_five_regulation. (Date Last Accessed on March 5, 2012).

[27] N. F. Noy and D. L. Mcguinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford KSL and SMI, Stanford, CA, 2001.

[28] W3C Semantic Web Interest Group, "Basic Geo (WGS84 lat/long) Vocabulary," 2003. [Online]. Available: http://www.w3.org/2003/01/geo/. (Date Last Accessed on March 5, 2012).

[29] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams,X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangguan, J. Flores, D. L. McGuinness, and J. A. Hendler, "TWC LOGD: A portal for linked open government data ecosystems," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 9, no. 3, pp. 325-333, Sep. 2011.

[30] E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, and Y. Katz, "Pellet: A practical OWL-DL reasoner," *Web Semantics: Science, Services and Agents on the World Wide Web*, vol. 5, no. 2, pp. 51-53, Jun. 2007.

[31] Department of Environmental Protection Massachusetts Government, "2011 Standards & Guidelines for Contaminants in Massachusetts Drinking Water," 2011. [Online]. Available: http://www.mass.gov/dep/water/drinking/standards/dwstand.htm. (Date Last Accessed on March 5, 2012).

[32] E. Prud'hommeaux and A. Seaborne, "SPARQL Query Language for RDF," *W3C Recommendation*, 2008. [Online]. Available: http://www.w3.org/TR/rdf-sparql-query/. (Date Last Accessed on March 6, 2012).

[33] J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, and K. Wilkinson, "Jena: implementing the semantic web recommendations," in *Proceedings of the 13th International World Wide Web Conference*, New York, NY, 2004, pp. 74-83.

[34] E. W. Patton, P. Wang, J. G. Zheng, L. Fu, T. Lebo, J. S. Luciano, and D. L. McGuinness, "SemantAqua Live Demo." [Online]. Available: http://aquarius.tw.rpi.edu/projects/semantaqua/. (Date Last Accessed on March 20, 2012).

[35] P. Wang, "Statistics of Water Quality Data." [Online]. Available: http://tw.rpi.edu/web/semantaqua/statistics. (Date Last Accessed on March 20, 2012).

[36] I. M. Bloom, "Black Water and Brazenness: Gas Drilling Disrupts Lives, Endangers Health in Bradford County, PA," 2011. [Online]. Available: http://protectingourwaters.wordpress.com/2011/06/16/black-water-and-brazenness-gas-drilling-disrupts-lives-endangers-health-in-bradford-county-pa/. (Date Last Accessed on March 5, 2012).

[37] Mall Amy, "One family's life in the gas patch of Bradford County, Pennsylvania," 2011. [Online]. Available: http://switchboard.nrdc.org/blogs/amall/one_familys_life_in_the_gas_pa.html. (Date Last Accessed on March 5, 2012).

[38] P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "A Semantic Portal for Next Generation Monitoring Systems," in *Proceedings of the 10th International Semantic Web Conference*, Bonn, Germany, 2011, pp. 253-268.

[39] P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "TWC-SWQP: A Semantic Portal for Next Generation Environmental Monitoring," TWC RPI, Troy, NY, 2011.

[40] J. G. Zheng, P. Wang, E. W. Patton, T. Lebo, J. Luciano, and D. L. McGuinness, "A Semantically-Enabled Provenance-Aware Water Quality Portal," presented at the Environmental Information Management Conference 2011, Santa Barbara, CA, 2011.

[41] P. Wang, J. G. Zheng, L. Fu, E. W. Patton, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "Next Generation Environmental Informatics as exemplified by the Tetherless World Semantic Water Quality Portal," presented at the AGU Fall Meeting 2011, San Francisco, CA, 2011.

[42] E. W. Patton, P. Wang, J. G. Zheng, L. Fu, T. Lebo, L. Ding, Q. Liu, J. S. Luciano, and D. L. McGuinness, "Assessing Health Effects of Water Pollution Using a Semantic Water Quality Portal," presented at the 10th International Semantic Web Conference, Bonn, Germany, 2011.