SemantAqua: A Semantically-Enabled Provenance-Aware Water Quality Portal

Evan W. Patton
NSF Graduate Research Fellow
Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, NY, USA

Joint work with: Jin Guang Zheng, Ping Wang, Timothy Lebo, Li Ding, Qing Liu, Joanne Luciano, and Deborah L. McGuinness

Introduction

- Real Life Motivating Example:
  - In 2009, in Bristol County, Rhode Island, children became ill with symptoms such as diarrhea. The cause was found to be polluted water (*E. Coli*) and citizens were asked to boil water until the issue was resolved.
  - Public concerns: “When did the contamination begin?”, “How did this happen?”, “How can we keep it from happening again?”
  - We need environmental informatics systems that can automatically integrate and analyze water quality.

Challenges

1. Raw data from multiple sources and in different formats – **difficult to integrate and query**.
2. Semantics of the water quality data are not explicitly encoded in the data – **machine can't process data automatically**.
3. Large amount of data due to large spatial region, long time span, and large number of pollutants and regulated limit – **analysis can be time consuming and complex**.

Semantics Can Help

1. Raw datasets can be represented in RDF and ontologies enable integration of data between sources
2. Ontologies also add meaning using OWL2 Datatype Restrictions and ObjectIntersectionOf to support pollution recognition
3. SPARQL CONSTRUCT and classification using Pellet allow automated reasoning over small subsets of data for efficiency.

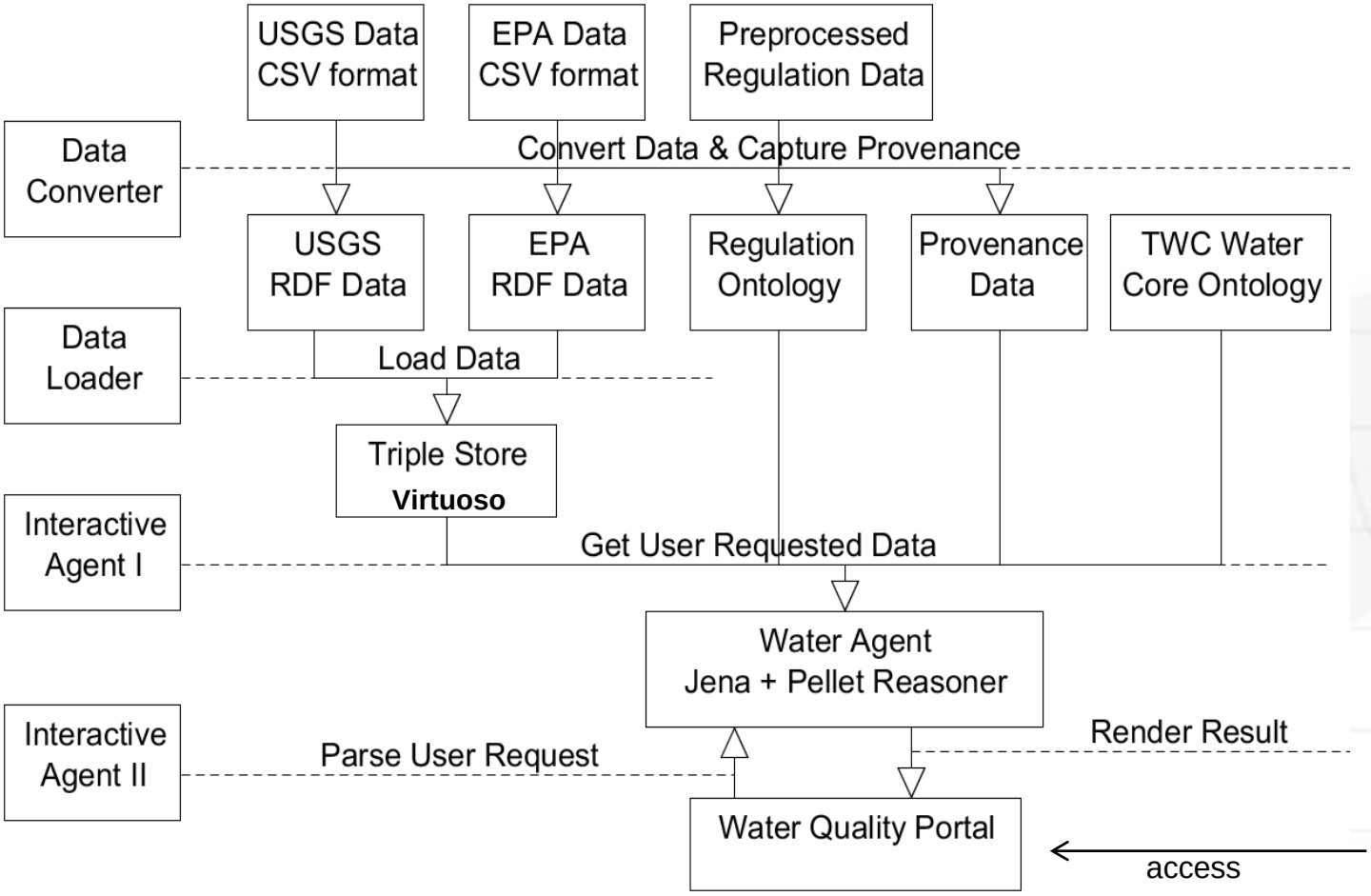
SemantEco

- Small OWL ontology, borrows from SWEET, OWL-Time, and SWIG Basic Geo, describing pollution concepts and associated metadata
- Combined with domain-specific ontologies, it can model pollution events (e.g. water pollution, air pollution, etc.)
- OWL 2 semantics simplify querying a SemantEco system

SemantAqua

- Identifies water pollution sites, including water sites monitored by USGS and polluting facilities regulated by EPA.
- Demonstrates the effectiveness of semantic web technologies in addressing the challenges faced by environmental informatics systems.
- Enable/Empower citizens & scientists to better explore water related information.

System Architecture

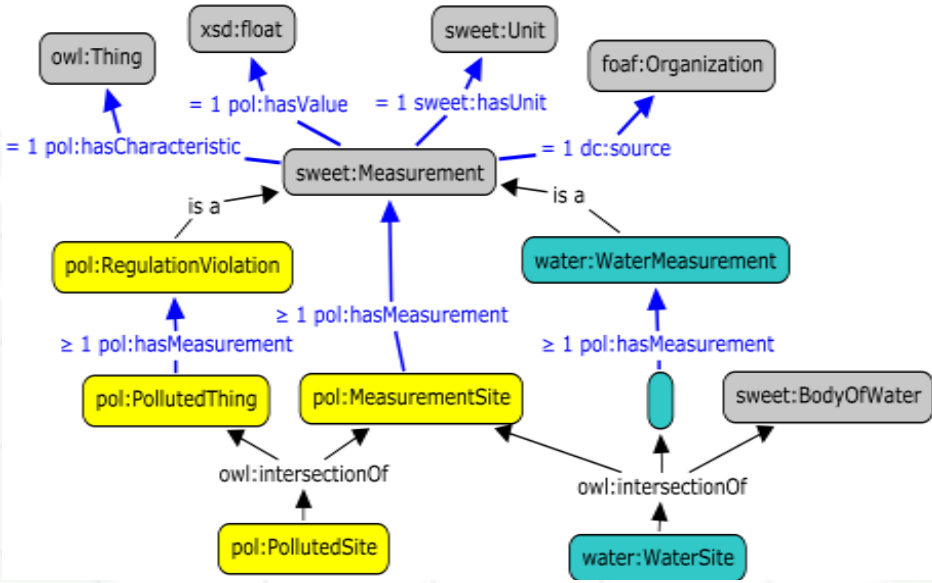


Some Statistics

- To date, encoded USGS and EPA data from 26 states (still ongoing)
  - 57,750 facilities
  - 493,504 water sites
  - 29,149,696 measurements
  - 3,048,378,871 triples
  - And that's just the data (no provenance)!
- Graph-level provenance adds one million more triples
- Option to include triple-level provenance although it is currently not needed

Ontology

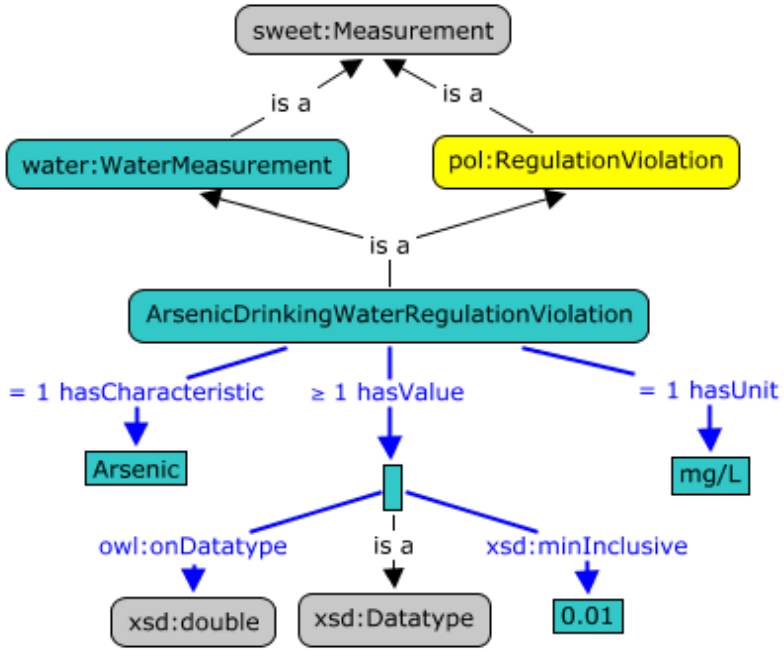
- Extends existing best practice ontologies, e.g. SWEET, OWL-Time.
- Includes terms for relevant pollution concepts
- Can be used to conclude: “any water source that has a measurement outside of its allowable range” is a polluted water source.



Portion of the SemantEco and SemantAqua ontologies.

Ontology

- Regulation Ontology
  - models the federal and state water quality regulations for drinking water sources
  - We can recognize pollution, e.g. “any water source that contains 0.01 mg/L of Arsenic or more is a polluted water source.”



Portion of Cal. Regulation Ontology.

Querying

- Traditional SPARQL query:

```
SELECT DISTINCT ?site ?lat ?lng
WHERE {
?site a pol:MeasurementSite ; pol:hasMeasurement ?m ;
  geo:lat ?lat ; geo:long ?lng .
?m pol:hasCharacteristic ?c ; pol:hasValue ?v ;
  units:hasUnit ?u .
?r a owl:Class ;
  rdfs:subClassOf [ owl:onProperty pol:hasCharacteristic ;
                    owl:hasValue ?c ] ;
    rdfs:subClassOf [ owl:onProperty units:hasUnit ;
                      owl:hasValue ?u ] ;
    rdfs:subClassOf [ owl:onProperty pol:hasValue ;
                      owl:someValuesFrom [ ?p ?l ] ] .
FILTER( isLiteral(?l) &&
  ((?l < ?v && str(?p) = xsd:minExclusive) || …))
}
```

Querying

- With OWL reasoning during query:

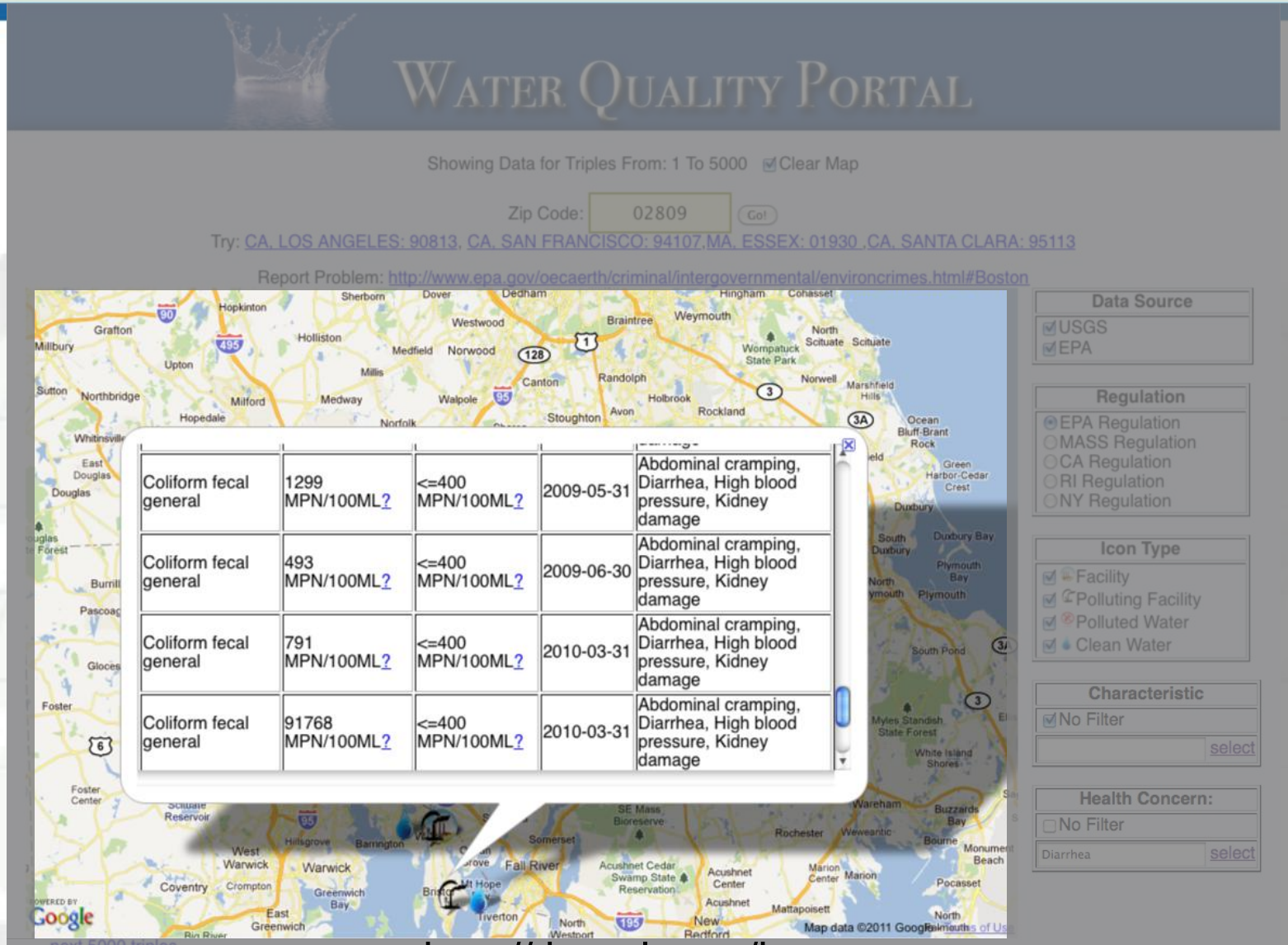
```
SELECT DISTINCT ?site ?lat ?lng
WHERE { ?site a pol:PollutedSite ;
        geo:lat ?lat ; geo:long ?lng . }
```

- OWL reasoning hides the complexity of the relationships allowing the developer (or user) to ask simple questions without requiring deep knowledge of environmental regulations

Provenance

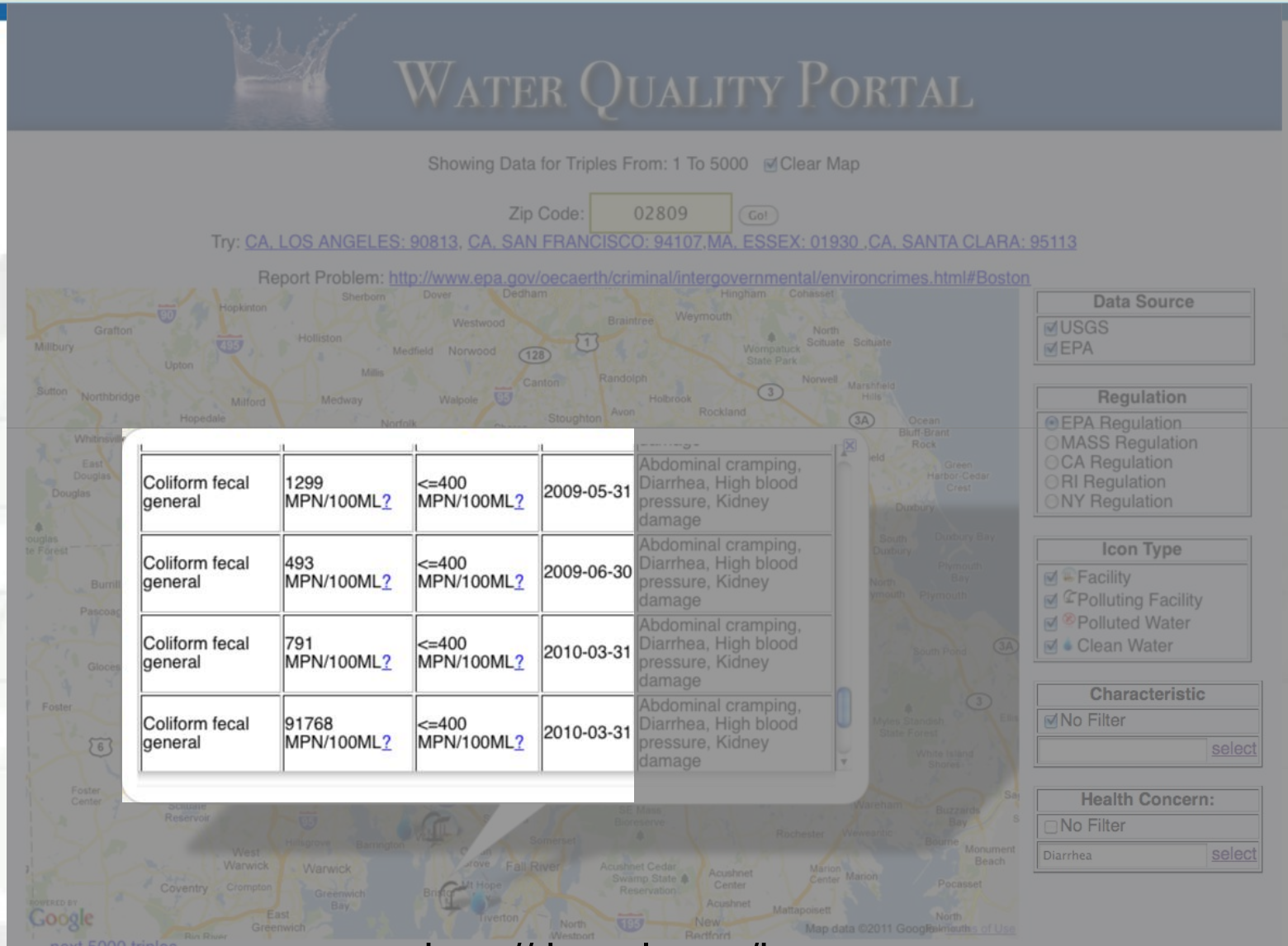
- Preserves provenance in the Proof Markup Language (PML).
- Data Source Level Provenance:
  - The captured provenance data are used to support provenance-based queries.
- Reasoning level provenance:
  - When water source been marked as polluted, user can access supporting provenance data for the explanations including the URLs of the source data, intermediate data, the converted data, and regulatory data.

Visualization



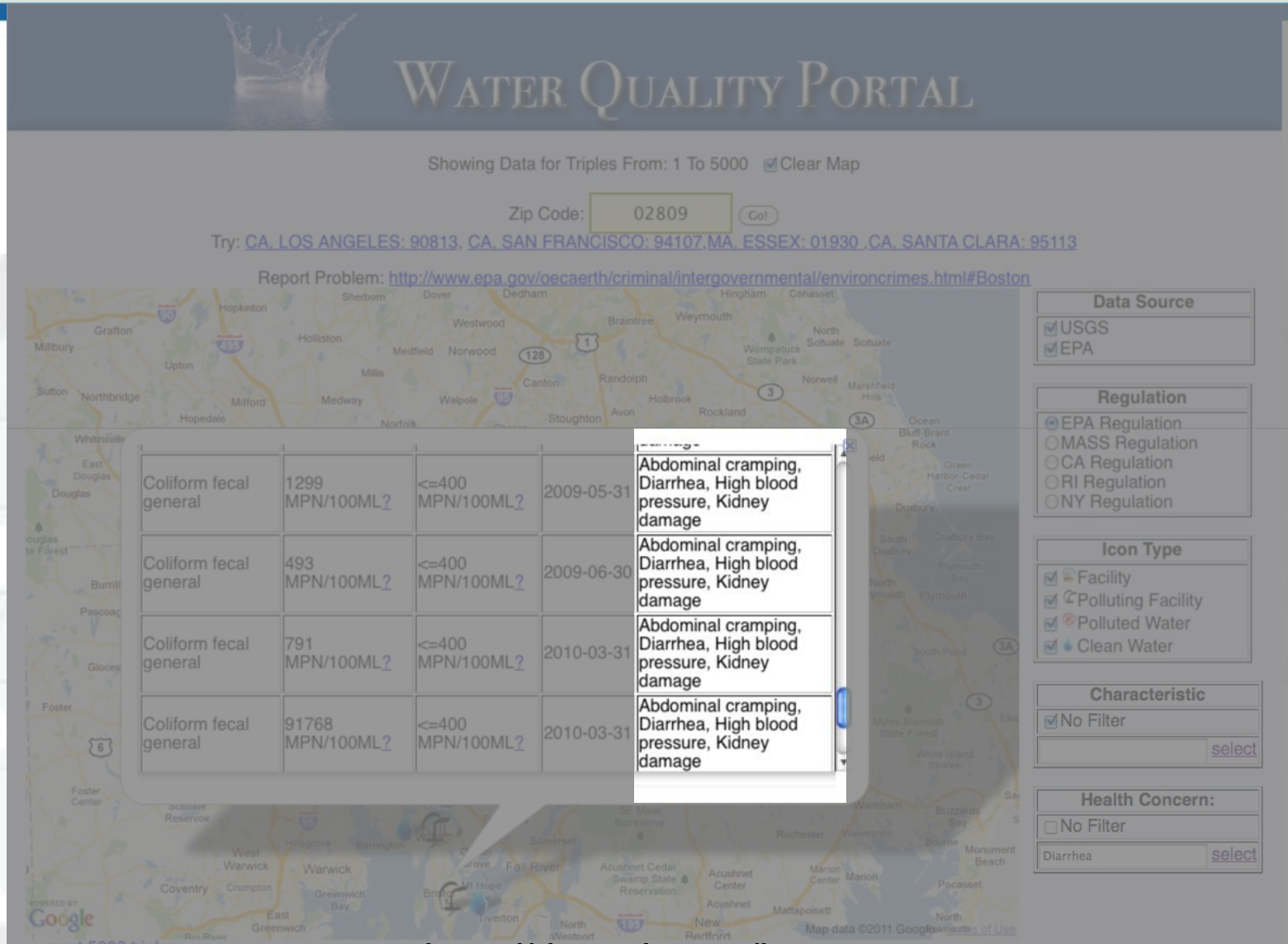
http://tinyurl.com/iswc-swqp

Visualization



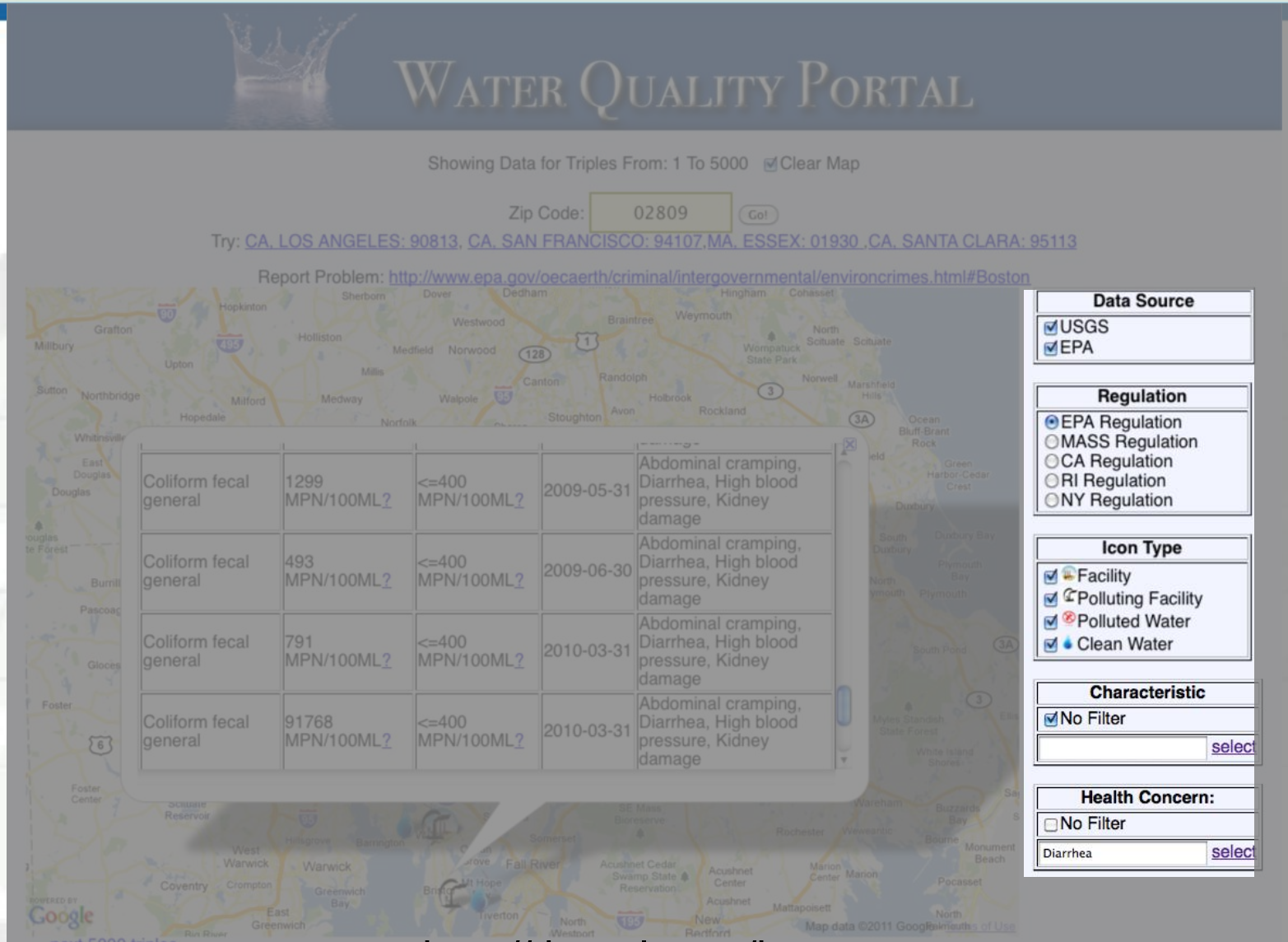
http://tinyurl.com/iswc-swqp

Visualization



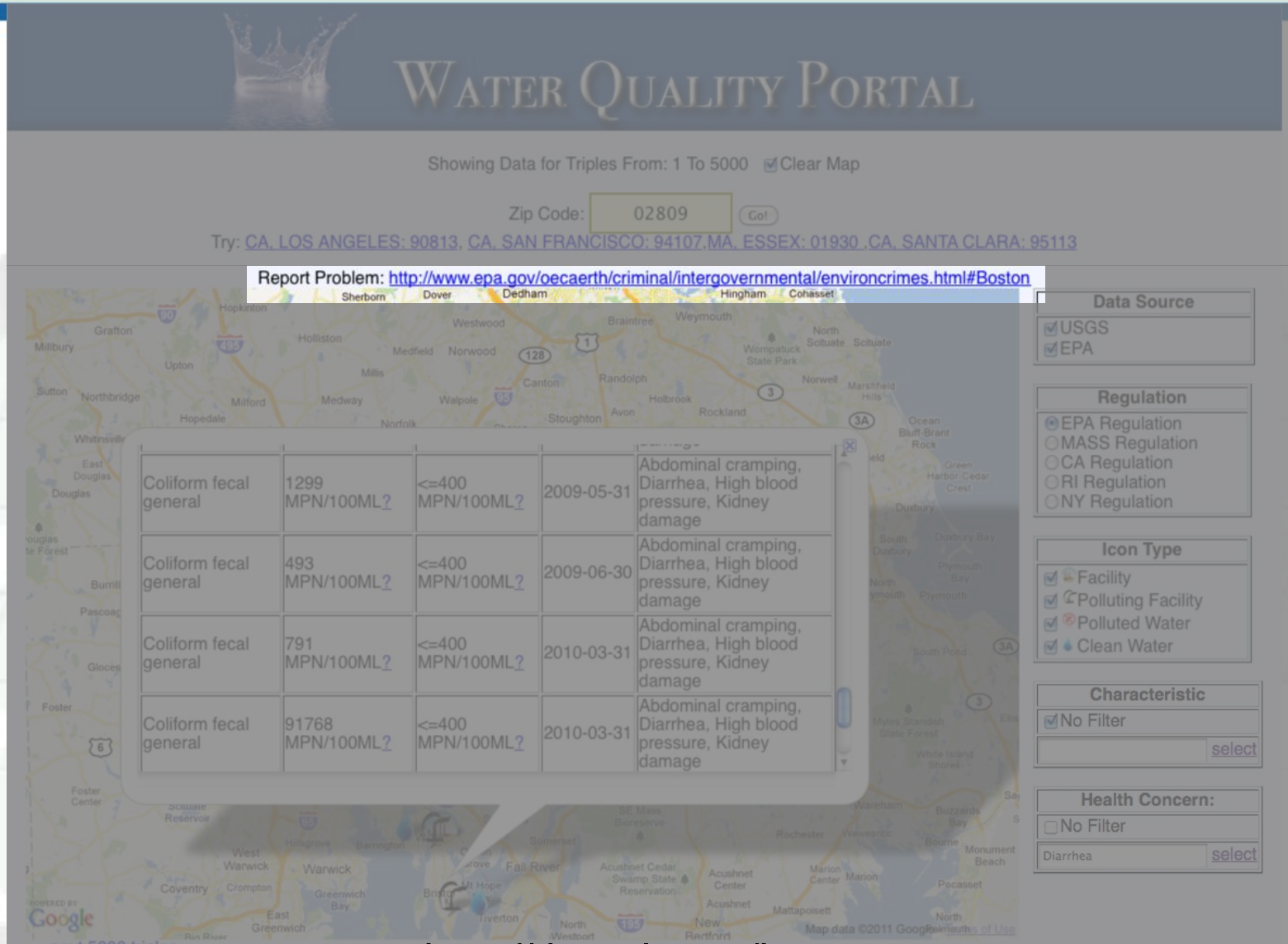
http://tinyurl.com/iswc-swqp

Visualization



http://tinyurl.com/iswc-swqp

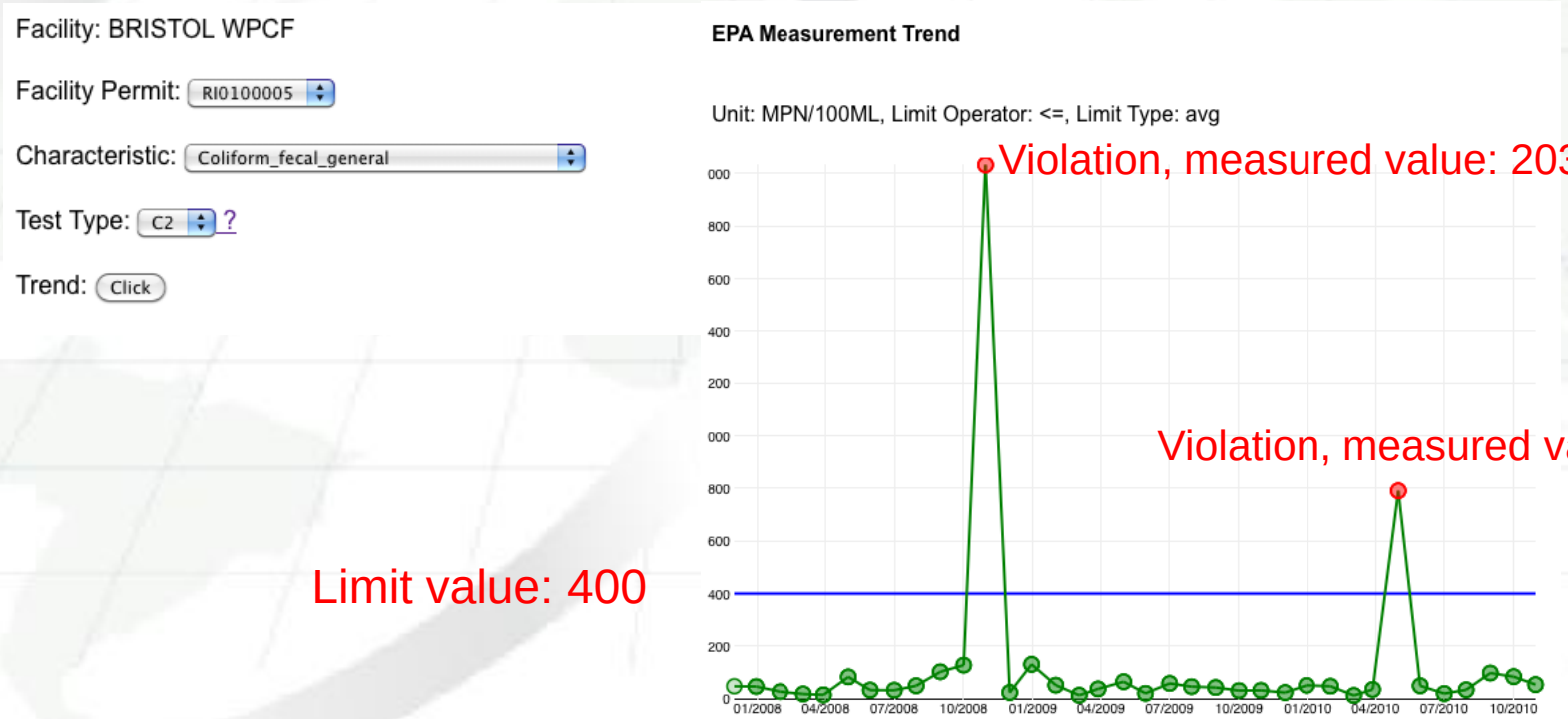
Visualization



http://tinyurl.com/iswc-swqp

Visualization

- Time series Visualization:
  - Presents data in time series visualization for user to explore and analyze the data



http://was.tw.rpi.edu/swqp/trend/epaTrend.html?state=RI&county=1&site=http%3A%2F%2Ftw2.tw.rpi.edu%2Fzhengj3%2Fowl%2Fepa.owl%23facility-110009444869

Data Reduction

- Potentially thousands of instances per county
  - Applying regulation ontology costly, cannot be done in a ‘reasonable’ timeframe
    - Tompkins County, NY has 140 sites, 9965 measurements
    - Classification takes 196 seconds on average
  - Potential large amount of targets, clutter
- Need to segment/page the data

Data Reduction

- Segment data into state-wide per-agency graphs
- Integration, pagination solved via SPARQL 1.1 Construct query:

```
CONSTRUCT {
   ?site a water:WaterMeasurementSite .
   …
   ?measurement a water:WaterMeasurement .
   …
}
WHERE {
   {
      SELECT ?site WHERE {
         GRAPH <…> { ?site a water:WaterMeasurementSite . }
      } order by ?site offset 0 limit 10
   }
   …
}
```

Data Reduction

- Results in less points to consider per display of interface
- Less triples pulled from triple store
- Faster reasoning by considering relevant data
- URIs make it straightforward to go back at user's request and obtain additional information (e.g. provenance)

Results

- Semantic Data Integration provides an effective and low cost approach for integrating data from various sources.
  - SemantAqua integrates data from various sources, including EPA, USGS, and state governments.
  - Linking to external data: “water:Arsenic”, linked to “dbpedia:Arsenic” using rdfs:seeAlso.

Results

- Query and reasoning supported by semantic technologies improves responsiveness and simplifies the development of web applications.
  - SPARQL queries narrow down the data allowing the application to reason over only the relevant data on one selected regulation.
  - Reasoning eases the complexity of queries a developer needs to write for software applications.

Results

- Provenance information encoded using semantic web technology supports transparency and trust.
  - SemantAqua provides detailed provenance information:
    - Original data, intermediate data, data source
  - “What if” Scenario:
    - User can apply a stricter regulation from another state to a local water source.
  - User may be interested only in certain sources and can use the interface to control queries

Discussion

- Future Work
  - Currently expanding SemantAqua to support all 50 states.
  - Add flood/weather information, and their effect on water sources; regulations can be different under flood conditions
  - Support reasoning over contaminants and their corresponding health effects.
  - Expand use of SemantEco ontology to other environmental topics: soil quality, air quality (e.g. support data from EPA's CASTNET)

Conclusion

- SemantEco provides a simple foundation for semantically-enabled monitoring systems
- SemantAqua is a web portal that allows citizens and professionals to easily explore water quality information from different sources.
- SemantAqua illustrates benefits of applying semantic web technologies to water quality research.
  - Data integration, provenance, automatic reasoning, simplified query structure.
- Magnitude of data forces us to address OWL 2 scalability issues and SemantAqua offers different approachs to improving performance.

Acknowledgements

- Advisors: Deborah McGuinness and Joanne Luciano
- Evan Patton is funded by a National Science Foundation Graduate Research Fellowship
- Tetherless World acknowledges funding from Microsoft Research, Qualcomm, Lockheed Martin, Fujitsu, and LGS Innovations

Questions?

http://tw.rpi.edu/web/project/SemantAQUA

http://inference-web.org/wiki/Semantic_Water_Quality_Portal

If you are interested in collaborating, please contact us:

pattoe @ rpi [dot] edu and dlm @ rpi [dot] edu