Towards Semantically-enabled Exploration and Analysis of Environmental Ecosystems

Deborah L. McGuinness

Tetherless World Senior Constellation Chair
Rensselaer Polytechnic Institute

In conjunction with Ping Wang, Linyun Fu, Evan W. Patton, F. Joshua Dein, and R. Sky Bristol

1RPI, 2USGS, 3Amazon

eScience 2012, Chicago, IL October 10, 2012
Motivation

- Semantic eScience class generated a water quality portal project
- Presented at Environmental informatics and AGU attracted interest of wildlife managers at USGS (among other places)
- Generalized a semantically-enabled water quality portal to be a semantic foundation for monitoring environmental and ecological data. This talk will be on the expanded more generalized version of our system.

- Wildlife and their habitats are deteriorating around the world
  - E.g. 40% of US freshwater fish are at risk [1]
- Scientists and resource managers need tools to integrate, interpret, and present data from both government and non-government sources
- Tools are need to support “the collaborative effort of multiple disciplines … to attain optimal health for people, animals, and the environment” [2]

Diverse (and growing) Data Sources

- Supporting resource managers requires integration of different data sources
  - Environmental quality data
  - Pollution regulatory data
  - Species counts
  - Health effects of pollutants on species
- Large amount of observations spread across different databases with different schema interpretations
- Performing data integration requires an automated process that can provide an understanding of the data and incorporate that into a shared schema
- EPA has two separate database formats for storing compliance history
  - Permit Compliance System (PCS) uses fixed-width ASCII tables for storing tabular data with a data dictionary for code lookups
  - Integrated Compliance Information System (ICIS) uses comma-separated values (CSV) for storing tabular data
- Different states report to different systems, so multiple methods of data conversion are required to aggregate the data into a single system to be queried
- Meaning is separated out into supplementary files (i.e. the data dictionaries)
- SemantAqua uses both compliance data and regulation data
- EPA also maintains documentation on the health effects caused by the pollutants tracked by the agency
National Water Information System

• Provides measurements for water quality at a large number of water bodies across the United States
• Data are retrieved as comma-separated values files using the US Geological Survey (USGS) National Water Information System (NWIS) web service
• Data are denormalized, so information is often repeated between measurements
Avian Knowledge Network

- Aggregates records of bird counts provided by government and non-government organizations that can be used for determining bird populations over time
- Tab-delimited files retrieved via an application programming interface (API)
- Includes a number of provenance-related items
  - Protocols used by the observing party
  - Whether submitted entries have been reviewed
  - Original reference system used by the observation team and its transformed coordinates in the final dataset
- Requires attribution of data to the appropriate parties
Wildpro

• Includes information on various species laid out taxonomically
• Provides references for habitat and migratory patterns of species
• Provides partial health effect information concerning contaminant effects on wildlife
(Selected) Value of Semantics

• Semantic technologies alleviate many of the issues of these different formats by incorporating the meaning and interpretation of the data into the data themselves
  – The conversion process we use incorporates the information from the data dictionaries so that artificial intelligence (AI) reasoners can make use of that external information

• Extract common information related to individual entries in tables to promote (i.e. normalize) them as objects that can be linked to and further described by others on the web

• Possibly biggest value is support for quick prototyping
  – students got an initial version up in a month
• Pollution ontology describes the relationship between a regulation violation (a measurement), a polluted thing, and a polluted site.

• Combined with other ontologies (e.g. W3C Geo) users can ask “Tell me all of the polluted things within 1 mile of my location.”
Ontologies

- Water quality ontology extends pollution to describe water-related pollution
- Further extended by regulation ontologies to provide “regulation violation” inference
- Allows the reasoner to match specific regulations to measurements that violate them
• Extensible Observation Ontology (OBOE) [1] provides a common schema for scientific observations to be shared.

• SemantEco provides a mapping ontology to recode from OBOE to the pollution ontology and vice versa to support interoperability with tools supporting OBOE.

Ontologies

- **Geospecies [1]**
  - Provides a semantic representation of species in the Linnaean taxonomy
- **Wildpro [2]**
  - Provides semantic encodings of the relationships between species and the observable health effects of exposure to pollutants

Architecture

• Converting the data from tabular formats into Resource Description Framework (RDF) is done using csv2rdf4lod [1]

• The converter produces RDF, provenance metadata to trace back to the data provider, and includes an RDF representation of the processes used to translate the data from tabular form to structured form for repeatability

Architecture
SemantEco/SemantAqua

- Enable/Empower citizens & scientists to explore pollution sites, facilities, regulations, and health impacts along with provenance.
- Demonstrates semantic monitoring possibilities.
- Map presentation of analysis
- Explanations and Provenance available

1. Map view of analyzed results
2. Explanation of pollution
3. Possible health effect of contaminant (from EPA)
4. Filtering by facet to select type of data
5. Link for reporting problems
6. Now joint with USGS resource managers; expanded to endangered species; now more virtual observatory style

http://was.tw.rpi.edu/swqp/map.html and http://aquarius.tw.rpi.edu/projects/semantaqua
### Interface

#### Water Body Properties

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ComID</td>
<td>148900411</td>
</tr>
<tr>
<td>Permanent</td>
<td>148900411</td>
</tr>
<tr>
<td>FDate</td>
<td>2011/08/29</td>
</tr>
<tr>
<td>Resolution</td>
<td>2</td>
</tr>
<tr>
<td>GNIS_ID</td>
<td>01531534</td>
</tr>
<tr>
<td>GNIS_Name</td>
<td>Lake Washington</td>
</tr>
<tr>
<td>AreaSqKm</td>
<td>89</td>
</tr>
<tr>
<td>Elevation</td>
<td>0</td>
</tr>
<tr>
<td>ReachCode</td>
<td>171100120005962</td>
</tr>
<tr>
<td>FType</td>
<td>390</td>
</tr>
<tr>
<td>FCode</td>
<td>39004</td>
</tr>
</tbody>
</table>

#### Provenance

- water body shapes: [ftp://www.ecy.wa.gov/gis_a/hydro/nhd/NHDmajor.zip](ftp://www.ecy.wa.gov/gis_a/hydro/nhd/NHDmajor.zip)
- terminology explanation: [National Hydrography Dataset Data Dictionary](http://nhd.usgs.gov/NHDDataDictionary_model2.0.pdf)
Interface


Species
- Canada Goose

Regulation
- EPA regulation for aquatic life
- NE regulation for aquatic life

Icon Type
- Clean Water
- Facility
- Polluted Water
- Polluting Facility

Industry
- All Data

Characteristic
- select

Health Concern
- select
Performance

• Restructuring data using OBOE required changing the structure of “what pollution is” in the ontology
• Such restructuring can adversely affect reasoning performance, so a handful of counties were chosen and their data were analyzed under the new schema
• We found up to 25% increase in triples and an order of magnitude increase in some reasoning.
Semantic engines used for quick prototypes, however scaling to 6 billion triples (to cover the entire country) caused scaling problems.

We began exploring performance enhancement strategies.

Custom rules to match certain characteristics of the ontology with the instance data could improve performance over the DL scenario by only considering relevant properties.

Obtained 40x speed up in cases.
Current and Future Work

• Incorporating air quality
  – Water quality is only one factor in the health of species
  – Air quality data is readily available through the EPA’s CASTNet project, among others

• Connecting with CUAHSI via the DataOne project
  – Additional sources of water data from a coalition of non-governmental institutions, in particular the Santa Barbara Coastal (SBC) LTER

• Further investigate reasoning tradeoffs to improve performance
Conclusions

• Semantic technologies provide:
  – Systematic means of integrating sparse data from various sources
  – Rapid prototyping of extensible applications via shared schemas for describing data

• Enhancing interoperability using mapping ontologies can decrease reasoning performance

• Performance can be improved by optimizing rule application to those that are essential for answering user queries
Conclusions, continued.

- Integration of many different types of data into a single access point allows stakeholders to ask more complex questions, e.g.,:

  Find **all water bodies** and nearby **facilities** around Lake Michigan, where **Mallard Duck** have been observed in the past 3 years and where **pollutants** that lead to **mercury toxicity** have been observed in the water.

Data on water bodies from USGS, facilities from EPA, places from USGS Geonames, ducks from species facet, time from time facet, health impacts from wildpro, contaminants from epa or usgs services to access it, …
• Acknowledgements:
  – Support for Tetherless World Constellation comes from Fujitsu, Lockheed Martin, LGS, Microsoft Research, Qualcomm in addition to sponsored research from DARPA, IARPA, NASA, NIST, NSF, and USGS.
  – Evan Patton is supported by a NSF Graduate Research Fellowship

• Contact Info:
  – Evan Patton  pattoe@cs.rpi.edu
  – Deborah McGuinness: dlm@cs.rpi.edu
Extra slides
Performance

• Using OBOE to encode the data in a more true-to-form representation increases the number of triples
• The complexity required in the ontology to support the pollution reasoning caused execution time to increase 14-54 fold
Performance

• Rewriting portions of the regulation ontologies as rules improves performance
  – 2-10x performance improvement

• Automated methods of simplifying ontologies into rule sets when certain semantic restrictions are used could improve other systems using semantic technologies to integrate data
Performance

Number of Triples

Kent RI Suffolk MA Yates NY San Fran. CA

County

# Triples

# Triples w/ OBOE
Performance

- Kent RI
- Suffolk MA
- Yates NY
- San Fran. CA

Time (s)
- Reasoning Time (s)
- Reasoning Time w/ OBOE (s)
Performance

Time (s)

Kent RI  Suffolk MA  Yates NY  San Fran. CA

County

Pellet  Rules

- Pellet
- Rules