SemantEco: A Next-Generation Web Observatory

A. Patrice Seyed*
DataONE
University of New Mexico
1 University Boulevard N.E.
Albuquerque, NM 87131

Tim Lebo
Evan W. Patton
Jim McCusker
Deborah L. McGuinness
Tetherless World Constellation
RPI
110 8th Street
Troy, NY 12180

ABSTRACT

A web observatory for empirical research of Web data benefits from software frameworks that are modular, has a clear underlying semantic model, and that includes metadata enabling a trace and inspection of the source data and justifications for derived datasets. We present SemantEco as an architecture that can serve as an exemplar abstraction for infrastructure design and metadata based on best practices in Semantic Web, Provenance, and Software Engineering, that can be employed in any Web Observatory, that may grow out of a community. We will describe how the SemantEco framework allows for searching, visualizing, and tracing a wide variety of data.

Categories and Subject Descriptors

D.2.8 [**Information Systems**]: Models and Principles—*General*

General Terms

Design

Keywords

Semantic Web, provenance, environmental sciences

1. INTRODUCTION

In this paper we discuss SemantEco, a portal for ecological and environmental data, and how recent work and planned provenance-based extensions exemplify a next-generation web observatory. SemantEco is a framework and portal for using the Web as a tool to study events and situations and has been leveraged for the comprehension of water quality with respect to EPA and state water quality regulations, and presentation of species population counts, both across geospatial and temporal regions. SemantEco's aim is to assist in improving our environment and health by providing tools to integrate ecological and environmental data, to monitor the interactions thereof, and investigate causes of pollution and its effects on human and animal health. This vision for SemantEco is aligned with that of DataONE[1],which is to enable discovery, acquisition, interpretation, and general usage of data for the biological, ecological, and Earth sciences.

Given the sheer amount of web-based data available to assist in this endeavor, these efforts must be undertaken with a large scale and multidisciplinary perspective to be effective. This requires access and use of disparate data sources and systems. For example, the U.S. Geological Survey (USGS) provides integrated science and technology to support resource managers in the U.S. Department of the Interior (DOI) through initiatives such as the Wyoming Landscape Conservation Initiative (WLCI): an effort to assess and enhance aquatic and terrestrial habitats at a landscape scale in southwest Wyoming. Decision support systems are one end result of scientific research that facilitates examination of the many tradeoffs and conflicting drivers that resource managers often wade through in their work, from energy and agricultural development to fish and wildlife conservation to recreational uses of public lands. Semantic technologies facilitate access to multidisciplinary information that aid resource managers in making decisions about complex ecosystems. These technologies also enhance reusability and address extensibility issues targeting challenges in the areas of data integration and scalability.

SemantEco is an exemplar for web observatories since its modular design supports a model that "anyone can play the role of data provider or data user" by leveraging software engineering best practices to implement semantically-driven data integration and search capabilities. One area for enhancement is provenance. Questions users may ask include: What happens to site classifications once the web session is refreshed and thus the inference engine is cleared? How can a scientist aggregate results, given these classifications, across web sessions? The answers to these questions is enabled through our use of PML/PROV[5, 3] as a model for capturing provenance; we will outline such an approach in Section 4. In what follows we provide an overview of the SemantEco architecture and focus on the derivation and use of provenance to enable data and information transparency.

2. ARCHITECTURE

The SemantEco framework is written in Java and makes use of the Jena semantic web framework [2]. It is designed to allow pluggable modules (provided as JAR files) to operate independently of one another while providing mechanisms

*This author has dual affiliation with University of New Mexico and RPI under the NSF DataONE Project.

WWW 2013 Companion, May 13–17, 2013, Rio de Janeiro, Brazil.
ACM 978-1-4503-2038-2/13/05.

[1]http://www.dataone.org/what-dataone

for them to interact through their user interfaces, data, ontologies, and queries. Data are provided in structured form using the Resource Description Framework (RDF) and domain knowledge is specified in the Web Ontology Language (OWL). The framework provides programmatic interfaces for enabling user interface controls, adding additional data, and incorporating new knowledge by extending the existing ontologies used in SemantEco. In previous work, we evaluated the applicability of this modular design to domains beyond water from early iterations of SemantEco [6] and have also enabled students to extend the system with new data and features for air quality and species data.

Besides this interaction, each module can provide any number of RESTful services which do not depend on other modules for processing (but may depend on other modules for generating the appropriate RDF and ontology model). These services can be called from the client at various points in the execution flow of the portal, and are particularly useful for recombining data into new visualizations based on a user's task requirements. For example, the characteristic module, which models the relationships between various water and air characteristics and contaminants, and the species module, which provides bird and fish data, each provide RESTful methods for accessing their data so that a time series plot can be generated for a particular geospatial point that combines both streams of data into a single plot.

Since access to SemantEco services is RESTful, application state is encoded on the client in the URL fragment. We make use of a javascript library, jQuery BBQ Plugin,[2] to encode parameters in a standardized way. These parameters are transmitted to the server to complete user requests. This also provides an additional benefit of allowing users to share the entire state of their session with colleagues via email or other communication channels simply by copying the URL from their browser. Some examples of parameters captured in the state include:

- zip - Current Zip Code being searched
- domain - List of enabled domains (e.g. water, air, bird)
- source - List of data sources selected by the user (e.g. EPA, USGS)
- regulation.water - Regulation ontology selected by the user to apply to data points in the water domain

Based on the features a module provides, it falls into one of three categories: *Core Modules*, *Data Provider Modules*, and *Query-Modifying Modules*. *Core modules* provide essential functionality for SemantEco's use case, such as maintaining information about data sources, entity types, and available domains. *Data Provider Modules* implement methods that modify the data and ontology models to incorporate new data and semantics. They also provide client side code for the entity type module to help it match classes with their representative icons on the map. Lastly, *Query-Modifying Modules* implement a query visitor method to modify queries prior to their execution and provide user interface elements that allow the user to scope queries.

Since modules are Java archives, it is possible to publish a module via HTTP that can be loaded into SemantEco,

[2]http://benalman.com/projects/jquery-bbq-plugin/

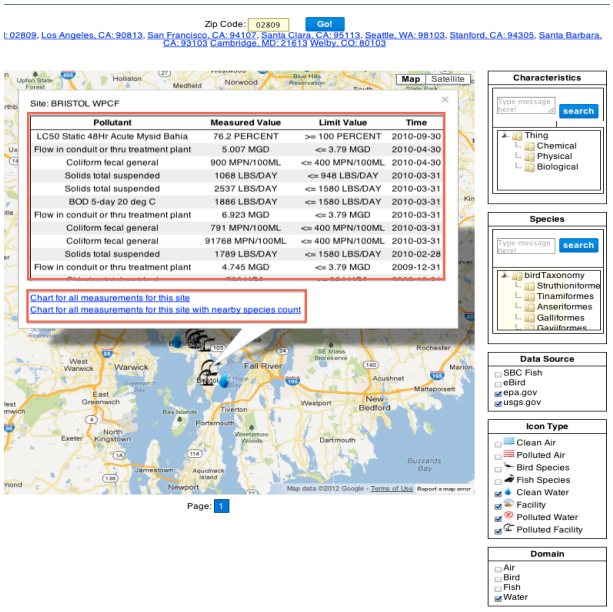

Figure 1: The SemantEco interface. The map presents locations of water, air, and species observations. Clicking on polluted water or air site presents a dialog that includes a table of found contaminants (top red box) and links to visualize the data (lower red box). Facets for controlling data search are on the right hand side.

providing a straightforward way for data providers to expose data in a usable, distributed fashion without systems maintainers needing to expend resources keeping track of module versions. This enables a "plugin and play" environment for multiple organizations to integrate data from multiple sources, and enables a user to leverage the UI interfaces according to whichever plugins are currently enabled.

3. SEMANTIC SEARCH

Web observatories are responsible for integration of massive amounts of multidisciplinary datasets and therefore must be able to handle the different types of data available and present them in a manner that is easily consumed by end users. To this end, SemantEco provides a number of tools for visualizing data. The primary interface of SemantEco is designed around rendering geospatial data (see Figure 1), but it provides mechanisms for rendering time series data as well (see Figure 2). Modules that incorporate data encoded with geospatial or time data using the W3C Geospatial[3] or OWL Time[4] ontologies will be able to easily expose their data using the existing visualization mechanisms in SemantEco. Furthermore, modules can supply client-side logic that offers new interfaces for data where existing visualization elements may not be appropriate. This flexibility allows for a more integrated view of different data types over the various domains exposed via SemantEco.

SemantEco also provides standardized mechanisms for exposing data encoded using taxonomies, such as species clas-

[3]http://www.w3.org/2003/01/geo/
[4]http://www.w3.org/TR/owl-time/

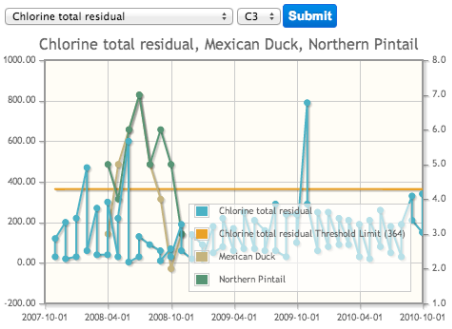

Figure 2: Water quality and Bird occurrence data.

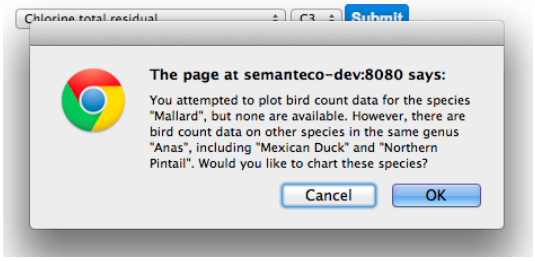

Figure 3: "Sibling suggester" feature.

sifications and chemical entities. The hierarchical search facet is a tool by which a module can render its internal class or role hierarchy to end users to allow them to browse data in a more structured way. Advanced users wishing to quickly find data can use these facets to find the appropriate class(es) and data loaded by the system will be constrained by any selections made. SemantEco will also provide suggestions of nearby classes in the hierarchy in the event a lookup does not return any data.

4. PROVENANCE IN SEMANTECO

Along our current efforts to explicitly describe the provenance of the data SemantEco uses and provides, and to further concretize its role as an exemplar web observatory, in what follows we: 1) apply the constructs of W3C PROV,[5] including properties **wasDerivedFrom** and **specializationOf**, in order to describe a derived dataset and its authoring agent; 2) provide these constructs from the same direct link to the application state afforded by jQuery BBQ plugin; and 3) outline the stewardship/authorship distinction for each dataset.

SemantEco uses OWL-DL ontology modeling, converted RDF data, and an OWL reasoner webapp to determine regulation violations from measurements that exceed chemical thresholds. This violation information is authored by SemantEco, but derived from accumulated third party sources. To implement #1 above, if for example **ExcessiveArsenic** is a regulation violation class **Measurement-12** satisfies, then the following provenance-infused data is newly asserted:[6]

```
semantecodata:violation-excessiveArsenic-Meas12
```

[5]http://www.w3.org/TR/prov-o/

[6]We express instance data using the Turtle Syntax [1].

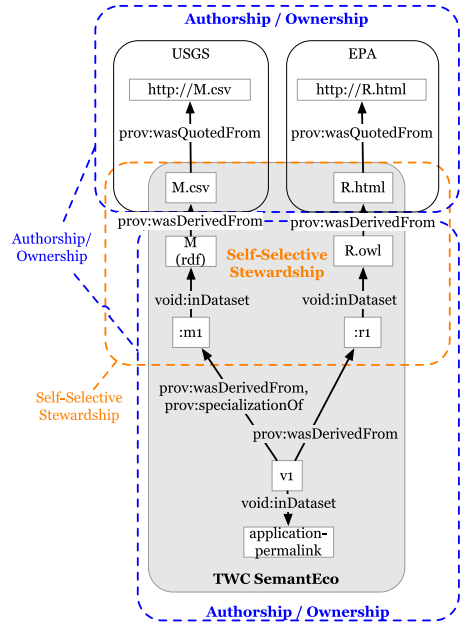

Figure 4: :v1 is assumed within the SemantEco namespace. Self-Selective Stewardship (orange box) here is SemantEco's stewardship of third-party data. :m1 stands for a particular measurement and :r1 stands for a particular regulation. 'M' stands for the collection of measurements, and 'R' stands for the collection of regulations. Application-permalink is a URI/URL that represents the application state.

```
a semantecovocab:Violation;
prov:specializationOf src1:Measurement-12;
prov:wasDerivedFrom   src1:Measurement-12;
prov:wasDerivedFrom src2:ExcessiveArsenic;
void:inDataSet semantecodata:bbqDataSet-223.
```

Authorship of this RDF data is captured in the following RDF statements:[7][8]

```
semantecodata:bbqDataSet-223
    pml3:authoredBy
        <http://purl.org/twc/semanteco/source/semanteco>;
    dcterms:contributor
        <http://purl.org/twc/semanteco/source/epa-gov>.
```

In the previous set of RDF triples, **ExcessiveArsenic** is an OWL class "punned"[9] as an OWL individual that represents the EPA Arsenic Regulation. From the PROV ontology, **wasDerivedFrom** captures the notion that one entity is transformed into another. **specializationOf** holds when one entity shares all aspect of another along with additional aspects of the former. In other words, its a contextualized notion of the more general entity. **violation-**

[7]pml3:authoredBy (http://inference-web.org/) is a subproperty of prov:wasAttributedTo.

[8]For the URI design of source organizations we follow the conventions established by our RDF converter CSV2RDF4LOD, within the purl.org namespace that we control. The URIs are redirected to a TWC server that is hosting the data. After 'source' is the provider of the original data files. In the case where we are the content provider, we name our own organization.

[9]http://www.w3.org/TR/owl2-new-features/

excessiveArsenic-Meas12 is a new datum that represents a specialization of and is derived from a measurement **Measurement-12**, is derived from a regulation **ExcessiveArsenic**, and is within a newly derived dataset generated by the SemantEco application, **bbqDataSet-223**, as is asserted above. Regarding #2, the URI for **bbqDataSet-223** is also a URL that represents the state of the application (i.e., application permalink), and is content-negotiable to obtain all provenance-infused RDF data used to create the web page view.

Since we are modelling inferred new data provided through an OWL-DL reasoner, it is also informative to explicitly encode data for cases that a measurement is not inferred to be a violation of a regulation. Although not currently in SemantEco's ontology, hypothetically if it were to include classes for non-violations (per regulation), based on a measurement being below a regulation's threshold, then these classes would be subclassed under a class **Non-Violation** that is disjoint from **Violation**.[10] Given this extension, a measurement is potentially an instance of both **Violation** and **Non-Violation**, for regulations at state and EPA levels with different thresholds for a chemical, which under this presumed model is an inconsistency.

This modeling approach fails because a measurement serves as (i.e., plays the role of) a violation only in the context of some specific regulation, and a site serves as a polluted site only in the context of a violation. To fit the above model of disjointness, then, we model these contextualized notions (i.e., specializations) of measurements, violations and non-violations, as members of the classes **Violation** and **Non-Violation**. We can further extend this model for sites, but reserve this discussion due to space constraints.

To provide the ability to trace the derived data back to the original source, we include also the following RDF statements:

```
src1:Measurement-12
  void:inDataSet semantecodata:graph11.
semantecodata:graph11
   prov:wasDerivedFrom src1:CSVFile112.
src1:CSVFile112
   prov:wasQuotedFrom http://epa.gov/DataURL4 .
http://epa.gov/DataURL4
   pml3:authoredBy
     <http://purl.org/twc/semanteco/source/epa-gov> .
```

In addition to showing these relationships using the example instance data in RDF, we also illustrate them in the corresponding schema in Figure 4. Also depicted in Figure 4 is that SemantEco is a *self-selected steward* of this measurement and regulation data (which in the corresponding RDF is within namespace prefixes 'src1' and 'src2', respectively). Self-selection is vital to this particular notion of stewardship, given that the organizations that originally authored the data have not assigned the SemantEco project as a handler of the data. Stewardship is a different, but a non-disjoint role from content authorship, and may or may not involve the stewarding organization adding new data elements. In addition, FRBR can be used to delineate the distinction between stewardship and authorship when deriving enhancements of third-party data [4].

By enabling content-negotiation of the application's permalinks, users have the option to both navigate measurement sites and measurements in visual style, or extract the equivalent data as RDF to inspect its derivational history (e.g., for claims of regulation violations), obtain the original data, or repurpose the derivations for a subsequent use.

[10]Since this would increase reasoning time, this sort of reasoning can be done outside the reasoner and on-demand for sites the user is interested, based on UI selections.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we described recent work on SemantEco's modular design and visualization capabilities, including planned provenance extensions on data generated and used by SemantEco.The modularity of SemantEco enables interoperability, and the state-based URIs (application permalinks) ease sharing and comprehending shared datasets, which are both key features of an exemplar web observatory that can enable the extension, refinement, reproduction, or elaboration of any existing visualization's dataset. We envision that this type of flexibility – combined with the plurality of the Web – can enable the development of follow-on analyses that better serve a web observatory's community.

In other planned work, we intend on using the SemantEco framework for developing an annotator interface, using CSV2RDF4LOD as a web service, for integrating new data via ontologies already in use by SemantEco. This work will serve to reduce the barrier of interoperability and data exchange and by providing good user interfaces for transcoding data into RDF, and also serve to enable immediate validation during or following an annotator session.

Acknowledgments

The Tetherless World Constellation is supported in part by Fujitsu, Lockheed Martin, LGS, Microsoft Research, Qualcomm, in addition to sponsored research from DARPA, IARPA, NASA, NIST, NSF, and USGS. E.W.P. is supported by a National Science Foundation Graduate Research Fellowship.

References

D. Beckett, T. Berners-Lee, E. Prud'hommeaux, and G. Carothers. Turtle: Terse RDF triple language. Technical report, W3C, 2013.

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 on Alternate track papers & posters*, pages 74–83, New York, NY, USA, 2004. ACM.

T. Lebo, S. Sahoo, and D. McGuinness. PROV-O: The PROV Ontology. http://www.w3.org/TR/prov-o/, 2013.

J. P. McCusker, T. Lebo, C. Chang, D. L. McGuinness, and P. P. da Silva. Parallel Identities for Managing Open Government Data. *IEEE Intelligent Systems*, 27(3):55, 2012.

D. McGuinness, L. Ding, P. Pinheiro Da Silva, and C. Chang. Pml 2: A modular explanation interlingua. In *Proceedings of AAAI*, volume 7, 2007.

E. W. Patton, A. P. Seyed, P. Wang, L. Fu, F. J. Dein, R. S. Bristol, and D. L. McGuinness. Semanteco: A semantically-powered modular architecture for integrating distributed environmental and ecological data. *Future Grid Computing Systems*, Submitted.