Towards Semantically-enabled Exploration and Analysis of Environmental Ecosystems

Ping Wang, Linyun Fu, Evan W. Patton, Deborah L. McGuinness
Tetherless World Constellation
Rensselaer Polytechnic Institute
Troy, NY USA
{wangp5, ful2, pattoe, dlm}@cs.rpi.edu

F. Joshua Dein, R. Sky Bristol
U.S. Geological Survey
{fjdein, sbristol}@usgs.gov

Abstract—We aim to inform the development of decision support tools for resource managers who need to examine large complex ecosystems and make recommendations in the face of many tradeoffs and conflicting drivers. We take a semantic technology approach, leveraging background ontologies and the growing body of open linked data. In previous work, we designed and implemented a semantically-enabled environmental monitoring framework called SemantEco and used it to build a water quality portal named SemantAqua. In this work, we significantly extend SemantEco to include knowledge required to support resource decisions concerning fish and wildlife species and their habitats. Our previous system included foundational ontologies to support environmental regulation violations and relevant human health effects. Our enhanced framework includes foundational ontologies to support modeling of wildlife observation and wildlife health impacts, thereby enabling deeper and broader support for more holistically examining the effects of environmental pollution on ecosystems. Our results include a refactored and expanded version of the SemantEco portal. Additionally, the updated system is now compatible with the emerging best in class Extensible Observation Ontology (OBOE). A wider range of relevant data has been integrated, focusing on additions concerning wildlife health related to exposure to contaminants. The resulting system stores and exposes provenance concerning the source of the data, how it was used, and also the rationale for choosing the data. In this paper, we describe the system, highlight its research contributions, and describe current and envisioned usage.

Keywords—Semantic Web; Semantic Environmental Informatics; Provenance; Ecological Data Integration;

I. INTRODUCTION

In many places around the world, wildlife and the habitats on which they depend are deteriorating. For instance, almost 40 percent of the United States' freshwater fish species are considered at risk or vulnerable to extinction according to [1]. Aiming at preserving the environment and wildlife, scientists and resource managers have initiated various efforts to monitor ecological and environmental trends, investigate causes and possible effects of pollution, and identify threats to wildlife and their habitats. To be effective, these efforts must be undertaken with a large scale and multidisciplinary perspective. However, by nature this approach requires access and use of disparate data sources and systems. For example, the 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.

Meanwhile, semantic science and technologies have been used in environmental monitoring to facilitate knowledge encoding, data integration and collaborative scientific workflows [2]. Previously we proposed the Tetherless World Constellation Semantic Ecology and Environment Portal (SemantEco) [3], [4] as both an environmental portal and as an example of a semantic infrastructure for environmental informatics applications. The portal captures the semantics of domain knowledge using a family of ontologies, integrates environmental data (e.g. water quality measurement and regulatory data), infers pollution events using OWL2 inference, and leverages provenance. In this paper, we extend the focus of SemantEco beyond water quality and related human health effects to a more comprehensive effort that includes fish and wildlife species and their related health effects. As such, this extension is in concert with the “One Health” concept, defined as “the collaborative effort of multiple disciplines — working locally, nationally, and globally — to attain optimal health for people, animals and the environment” [5]. Semantic technologies will facilitate access to multidisciplinary information that will 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, provenance, and scalability.

We use semantic technologies to provide responses to these challenges. The present work connects to the Extensible Observation Ontology (OBOE) [6] to support interoperable observation data. We investigated the reasoning implications of adopting OBOE for interpreting data and found out that OBOE brings greater interoperability at the cost of longer reasoning time. Additionally, we now integrate various ecological and
environmental data: wildlife observational data from the Avian Knowledge Network (AKN\(^1\)) and U.S. Geological Survey (USGS\(^2\)); environmental criteria for wildlife from the Environmental Protection Agency (EPA\(^3\)); water body data from USGS; and health effects of contaminants on wildlife from Wildpro\(^4\). These are provided in addition to the existing water quality and regulatory data presented in earlier versions of this work.

Our approach provides a formal encoding of the semantics of the data and provides services for automatic reasoning and visualizations. Furthermore, we compared the performance of a standard semantic web reasoner used in previous iterations with a new, customized rule based reasoner over our data. Lastly, we enhanced the provenance support of the existing portal by incorporating rationale as provenance. This extensibility enabled by semantic technologies allows the portal to consume data from many different domains and present it together for the purposes of informing environmental stakeholders.

In this paper, section 2 elaborates how semantic web technologies have been used to extend and improve the portal, including connecting to OBOE, extension for wildlife monitoring, upgrade the reasoning mechanism, and capturing rationale as provenance. Section 3 reviews related work. Section 4 discusses impacts, highlights, and future directions. Section 5 presents the conclusion.

II. SEMANTIC APPROACH

To inform the development of decision support tools for environmental and observational communities, we extend the focus of SemantEco to a more comprehensive effort including fish and wildlife species and their related health effects, and upgrade the portal with respect to three aspects: data integration; breadth and scalability, and provenance support.

A. Knowledge bases

To increase the interoperability of the portal, we update our ontologies and data conversion to make them compatible to OBOE. We name the previous version of our SemantEco ontologies as version 2 and the new version as version 3. The two versions differ in several ways. For instance, the SemantEco ontology capture measurement values with the data property hasValue, while OBOE encodes measured values using property chains on structured objects. More details about the differences between SemantEco and OBOE can be found in [7]. Fig. 1 and 2 depict the data presentation of an example water observation record generated according to the SemantEco version 3 and 2.

We updated our data converter and regulation converter for encoding the water observation data and regulation rules according to the new ontologies, and investigated the reasoning implications of adopting OBOE for interpreting data. Our system uses the Pellet OWL reasoner [8] together with the Jena

---

1 http://www.avianknowledge.net
2 http://www.usgs.gov/
3 http://www.epa.gov/
4 http://wildpro.twycrosszoo.org
Semantic Web Framework [9] to reason over the data and ontologies to answer the queries for polluted sites. Query (1) is under SemantEco version 2, and query (2) is under version 3. We execute each query 10 times to obtain an average query time. Table I summarizes the data sets and the query execution time. There is a clear tradeoff between interoperability and reasoning performance: while OBOE brings greater interoperability for the portal, it incurs much longer reasoning time. Reasoning performance is tested on a Silicon Mechanics running Debian Squeeze (6.0.4) configured with two six core Intel Xeon E5620 with hyper-threading operating at 2.67 GHz and 48 GB of 1333 MHz dual ranked memory. Tests using Pellet were conducted using the 64-bit Java 1.6.0_26 runtime environment (sun-javase-jre, version 6.26-0squeezel) and the virtual machine was configured with 1GB min heap size and 32GB max heap size.

select distinct ?pollutedSite
where{ ?violation a pol:RegulationViolation.
  ?violation pol:hasSite ?pollutedSite. }  (1)

select distinct ?pollutedSite
where{ ?observation a oboe:ObservationOfRegulationViolation.
  ?context oboe-core:hasMeasurement ?observation oboe-core:hasValue ?pollutedSite. }  (2)

<table>
<thead>
<tr>
<th>Region</th>
<th>Kent, RI</th>
<th>Suffolk, MA</th>
<th>Yates, NY</th>
<th>San Francisco, CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulations</td>
<td>RI</td>
<td>MA</td>
<td>NY</td>
<td>CA</td>
</tr>
<tr>
<td># of Triples v2</td>
<td>25293</td>
<td>53866</td>
<td>27631</td>
<td>74920</td>
</tr>
<tr>
<td># of Triples v3</td>
<td>31670</td>
<td>75059</td>
<td>38926</td>
<td>101969</td>
</tr>
<tr>
<td>Reasoning time v2</td>
<td>6.4 s</td>
<td>32.9 s</td>
<td>32.6 s</td>
<td>58.4 s</td>
</tr>
<tr>
<td>Reasoning time v3</td>
<td>301.9 s</td>
<td>1797.1 s</td>
<td>568.9 s</td>
<td>856.0 s</td>
</tr>
</tbody>
</table>

Table 1 demonstrates that the cost of adding interoperability through support of the OBOE ontology greatly increases the amount of time required to complete reasoning over all of the counties studied. There are a number of reasons for this difference, but the one that contributes the most is the complexity of the graph and how the changes in OBOE affect how the logical structure of a regulation increase the amount of time needed by Pellet to classify each observation (e.g. compare the order of the graphs in Figs. 1 and 2). Thus, one technique that may be important in further versions of the portal is a tool for converting OBOE observations into the more compact representation used in earlier versions so that reasoning performance is increased.

B. Extension for Wildlife Monitoring

We designed the following use case to identify necessary extensions to the portal for evaluating wildlife health effects from pollutants:

The resource manager chooses a geographic region of interest by entering a zip code and the species of concern in the species facet. The portal identifies polluted water sources and polluting facilities, and visualizes the results on a map using different icons. Meanwhile, the portal displays the distribution of the species in the region. Then, the resource manager views the map to find out if the selected species might be endangered by water pollution in the region. The resource manager can click on polluting facilities or polluted sites to investigate more about the pollution, e.g. the health effects of the pollution on the species.

To realize the use case, we enhanced our ontology for modeling the domain of wildlife observation, integrated ecological and environmental observation data according to linked data principles [10], and developed visualizations to present the data and provenance.

Ontology for Wildlife Monitoring: There are a number of existing ontologies for modeling and publishing RDF data about species descriptions [11], [12]. After reviewing these ontologies, we choose to reuse the Geospecies ontology for the purpose of modeling the domain of wildlife monitoring as it contains most of the concepts required by the use case. For example, the Geospecies ontology defines the classes Observation and SpeciesConcept, and links the two classes with object properties hasObservation and hasSpecies.

However, the Observation class from Geospecies does not capture the observed habitat of the wildlife nor the date of the observation. Thus, in the extension15, we introduce a new class

![Subset of wildlife ontology](http://escience.rpi.edu/ontology/semanteco/3/0/wildlife.owl)

---

15 http://escience.rpi.edu/ontology/semanteco/3/0/wildlife.owl
WildlifeObservation that subclasses Observation and provides a new property: hasDate. A subset of our ontology extension is illustrated in Fig. 4.

Each contaminant can cause different adverse effects on different wildlife species. For example, when exposed to excessive zinc concentrations, mallards exhibit leg paralysis and decreased food consumption while invertebrates exhibit decreased growth rate and increased mortality [13]. To help facilitate the investigating health impacts of contaminants on wildlife species, we refactored our ontologies to include the class HealthEffect to model the potential health effects of overexposure to contaminants [14]. The property isCausedBy establishes the relationship between each effect and its causing contaminant and the property forSpecies links each effect with its target species.

Wildlife Source Data: We integrate various ecological and environmental datasets from government and research institutions as follows.

Bird observation data: One major source of bird observation data is AKN, an international effort of government and non-government institutions to understand the patterns and dynamics of bird populations across the Western Hemisphere. We obtain a subset of the eBird Reference Dataset (3.0) [14] from AKN via its database query interface. The datasets are based on reported observations from novice and experienced bird observers. These datasets contain count data for bird species, the location where observation took place, and time observation started. The AKN is strengthened through the adoption of species codes from the Integrated Taxonomic Information System (ITIS), providing unique identification of species tied to an authoritative source.

Fish observation data: The National Fisheries Data Infrastructure brings together local and regional fisheries information systems and provides fisheries managers and decision-makers with one source of comprehensive data and information of fish species [16]. We fetch fish observation data from its query interface. The fish dataset includes the species name, the hydrological unit code (HUC) [17] of the watersheds where the fish species is observed, the date of the observation, and the originating database.

Regulation Data: We integrate EPA’s compilation of national recommended water quality criteria [15], which is presented as a summary table containing recommended water quality criteria for the protection of aquatic life and human health in surface water for approximately 150 contaminants.

Water Body Data: The USGS National Hydrography Dataset (NHD) services are used to get HUC codes given locations on the map [18], and the data for the water body shapes [19].

Health Data: We obtain the health effects on wildlife from the research effort Wildpro, an electronic encyclopedia and library for wildlife.

Data Conversion: The general-purpose csv2rdf4lod tool provides the capability of quick and easy data integration [16]. We provide the converter with declarative parameters that map properties of the raw data to terms defined in ontologies. For example, the field “Common Name” in the eBird dataset is mapped to the property geospecies :hasCommonName. Using this mapping, the converter is able to generate RDF triples compatible to our ontologies from the raw tabular data provided by our data sources. We use the same regulation ontology design and conversion workflow with our previous work [4] to map the rules in wildlife regulations to OWL [17] classes.

Data Visualization: We support two types of visualizations: (1) map visualization that displays the sources of the water pollution together with species habitat in the context of geographic regions and (2) time series visualization that depicts species count over time with respect to a particular geographic region.

The map visualization retrieves sources of water pollution and species habitats from the back-end reasoner and the triple store. We label clean and polluted water sources, and clean and polluting facilities with different markers. In the extension, we focus on waterfowl and fish. We visualize highlight the water bodies that are their habitats. When the user clicks one of the highlighted water bodies, the information of the water body and the provenance of the information are shown in the "Water Body Properties" tab of the pop up window. Fig. 5 shows an example of our map visualization and pop up window for water body. In the example, the portal applies EPA’s water quality criteria for aquatic life on the region with the zip code 98103 (Seattle, WA) and identifies polluting sites (outlined in red).

Figure 5. Map Visualization

Figure 6. Time Series Visualization

---

[18] http://services.nationalmap.gov/ArcGIS/rest/services/nhd/MapServer
that are close to bird habitats.

The time series visualization retrieves species count data within a particular geographic region by querying the triple store and displays the data as a time series using the d3.js library. Fig. 6 shows the count of Canada geese in the Washington state in 2007.

C. Reasoning engine

Due to forward-chaining closure computing, standard reasoners such as Pellet are much slower than general rule reasoners with only the necessary rules. For example, a standard reasoner for the RDF language would include the following rule (encoded with Jena rule syntax):

\[
\text{\texttt{[rdf1: (uuu aaaa yyy) \dashedrightarrow (aaa rdf:type rdf:Property)]}}
\]

Although this rule ensures the fulfillment of the semantics of RDF, it is not useful in the query answering task of our system. There are 14 such rules embedded in a standard RDFS reasoner and even more for an OWL-DL reasoner like Pellet. We avoid the invocation of these rules to improve query answering efficiency. Table II compares the performance of a specifically tailored rule reasoner implementing rules such as (3) and (4) to the Pellet reasoner. They both answer query (5) under the California water quality regulations on the same hardware as described in Section II.A.

\[
\begin{align*}
\text{Chloride: (?x pol:hasValue ?v) ge(?v, 10.0) (?x pol:hasCharacteristic pol:Chloride) (?x repr:hasUnit 'mg/l') \dashedrightarrow} \\
\text{(?x rdf:type pol:ExcessiveChlorideMeasurement)} \\
\text{Chloride:subclass: (?x rdf:type pol:ExcessiveChlorideMeasurement) \dashedrightarrow (?x rdf:type pol:RegulationViolation)} \\
\text{select ?s ?x ?v where} \\
\text{(?x a pol:RegulationViolation).} \\
\text{(?x pol:hasValue ?v. ?x pol:hasSite ?s)}
\end{align*}
\]

TABLE II. RULE REASONER VS. PELLET

<table>
<thead>
<tr>
<th>Region</th>
<th>Kent, RI</th>
<th>Suffolk, MA</th>
<th>Yates, NY</th>
<th>San Francisco, CA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rule</td>
<td>4014 ms</td>
<td>15408 ms</td>
<td>4333 ms</td>
<td>13093 ms</td>
</tr>
<tr>
<td>Pellet</td>
<td>11384 ms</td>
<td>162566 ms</td>
<td>31148 ms</td>
<td>31558 ms</td>
</tr>
</tbody>
</table>

These results show that the new custom rule implementation of the regulations outperforms Pellet by an average factor of four and range between 10% to 40% of the overall running time of the complete OWL-DL closure. This suggests that having automated mechanisms to convert from complete OWL-DL ontologies into rule-based approximations could provide significant performance improvements when completeness of the OWL-DL closure is not necessary to sufficiently answer application queries.

D. Rationale as provenance

During the construction of an information portal, we make various choices: what data sources to use, which integration tools, which data transformations, etc. As pointed by Bristol from USGS, rationale that explains how choices were made is critical. Rationale helps portal users to obtain a holistic understanding of the portal and facilitates the portal’s maintenance and reuse. Users are more likely to have confidence in the portal if the rationale behind design choices and the presentation of data are acceptable for them. If other portal builders are interested in reusing the architecture or workflow, they can more easily decide whether they would like to reuse one dataset or software agent when given the rationale for why we selected a particular dataset or software agent.

To encode rationale as provenance, we extend PML 2. PML 2 is a modular explanation interlingua and contains three ontologies that focus on three types of explanation metadata: provenance, information manipulation or justifications, and trust [18]. We introduce the property pmlp:hasRationale, whose domain is the class Identified Thing and range is String. Identified-things can be information, language and sources (including organization, person, agent, services). With hasRationale, we can provide the rationale for why we choose to adopt the identified things in simple text. By extending the scope of provenance to include rationale, we are able to capture some important information that may otherwise be lost when the original architects of the portal leave the project. Table III includes three rationale examples.

While we automatically capture provenance during the data integration stages, such as data source URL, time, and method used in data manipulation due to the provenance support from csv2rdf4lod [3], [4], the rationale as provenance are captured manually. The provenance captured using csv2rdf4lod are published as metadata graphs within the SemantEco triple store and can be used for generating faceted search interfaces, e.g. providing data source based search to allow users with a particular interest in EPA data to selectively access datasets provided by the EPA.

TABLE III. EXAMPLES OF THE RATIONALE CAPTURED

<table>
<thead>
<tr>
<th>Identified thing</th>
<th>Type</th>
<th>Rationale</th>
</tr>
</thead>
<tbody>
<tr>
<td>USGS</td>
<td>Data Organization</td>
<td>It is an authoritative government agency for science about the Earth, its resources, and the environment.</td>
</tr>
<tr>
<td>NWIS</td>
<td>Dataset</td>
<td>It is distributed via web services and can be accessed periodically with automated means.</td>
</tr>
<tr>
<td>csv2rdf4lod</td>
<td>Software</td>
<td>The open source tool provides an quick and easy way to convert tabular data into well-structured RDF. We have direct support from the author of the tool, who is our lab mate.</td>
</tr>
</tbody>
</table>

III. RELATED WORK

In ecology and environmental communities, there have been research efforts that facilitate domain knowledge integration via semantic approaches. These research projects focus on different fields of ecological and environmental science. OBOE focuses on encoding generic scientific
observation and measurement [6]. GeoSpecies is an effort for enabling species data to be linked together as part of the Linked Data network [11]. Chen et al. proposed a prototype system that integrates water quality data from multiple sources [19]. However, environmental decision support systems need to integrate data across multiple fields, facilitates data examination and assist natural resource managers to draw informative conclusions. In this work, we designed a family of ontologies for encoding water measurement, species observation and the health effects of pollution on species and utilize the ontologies for data integration and reasoning, and demonstrated the value of semantic approaches for building environmental decision support systems.

eScience can benefit from provenance for a number of reasons. For example, provenance provides a context for data and information interpretation, enabling evaluation of the experimental results and replication of scientific workflows [20]. Research projects such as myGrid [21] and CMCS [22] have been conducted to build infrastructure that generates provenance data and allows users view and use provenance in analyses. The provenance support of this work differs from that of previous projects in that we extend the scope of provenance support to include rationale.

IV. DISCUSSION AND FUTURE WORK

Ecology and Environmental information systems benefit from semantic science and technology from several aspects. Firstly, by encoding the domain knowledge required by the information system with ontologies, we make the information system easier to maintain and extend by third parties interested in integrating their data into our system. In SemantEco, we encode the environmental regulation rules and the health effects of pollution on wildlife as OWL classes. If one rule becomes stricter, we only need to update the OWL class to adopt the stricter threshold value. This extensibility can enable users to make use of latest finds in research to refine thresholds. While such a system could be implemented with a relational database, it would require additional code complexity for maintaining threshold rules and providing user interfaces for manipulating them. Consider as an example a state wanting to import the EPA ruleset and then provide a more strict set of regulations for half of the pollutants while maintaining the EPA’s thresholds for the remaining contaminants. An OWL-DL reasoner such as Pellet can infer that the stricter regulation is a subclass of the less stringent EPA regulation and thus skip processing of the additional class. In a database environment, a developer or other maintainer would need to make a determination of whether exclude such a rule (e.g. in the form of a SQL query) from the application to simulate the same behavior or potentially face duplicate or irrelevant results.

Similarly, adding a new OWL instances is sufficient for an extension such as introducing the health effect of a new contaminant. In contrast, if we embed the domain knowledge in the source information systems, these changes would require more costly modifications than changing the ontology files. The small number of example data sets used in this experiment for wildlife observations, pollution sources, regulatory thresholds, and health effects will need to be dramatically expanded and diversified to answer complex resource management questions. However, the methods pioneered by the SemantEco portal should prove useful in scaling to the necessary breadth and depth of such systems. In particular, SemantEco provides a platform that uses data published under a collection of interoperable ontologies, so data published by providers using an interface such as D2RQ [23] would allow SemantEco to be easily extended to incorporate external database systems. Such a system would be easier to maintain than writing ad-hoc data access APIs on a per-database level.

Semantic technologies facilitate data integration, a crucial step in building ecological and environmental information systems that are inherently interdisciplinary and sourced from many sectors – government, industry, private citizens, and others. Synthesizing observation data using our wildlife ontology and adoption of standards such as ITIS leads to the ability to integrate across disparate datasets. For example, we map the field "Latitude" and "Longitude" of the eBird Reference Dataset to the property wgs84:lat and wgs:lon.

Resource managers need the best available scientific information in a form that is easy to comprehend and assimilate into decision-making processes. For instance, evaluation of a pollution event in a watershed can benefit more from a focused list of species most likely to be impacted rather than a comprehensive list of all species that might be found in the watershed. This often involves large amounts of data and the analysis can require much time and effort to arrive at a decision with significant impacts. Semantic technologies can be used to lower the cost and shorten the time required by such decision-making processes by applying knowledge of known data connections under varying management and condition scenarios.

We use SPARQL [24] to retrieve data and perform appropriate data aggregation as guided by the semantics encoded in the ontologies. Not only does SPARQL enable us to specify the constraints of the data aggregation, it also supports aggregation functions including COUNT, SUM, AVG, MIN, and MAX. For example, we obtain the total counts of "Canada Goose" in Washington State in 2007 with the SPARQL query in (6). The query result can be provided in XML or JSON that can be readily consumed by different visualization toolkits (e.g. D3.js) to produce a time series plot. This way, with SPARQL and visualization toolkits, the portal enables users to review and interact with the growing data resource in the form of maps and other visualizations.

PREFIX geospecies: <http://rdf.geospecies.org/ont/geospecies#> .
PREFIX wildlife: <http://www.semanticweb.org/ontologies/2012/2/wildlife.owl#> .
SELECT ?month SUM(?count) as ?total
WHERE { ?obv wildlife:hasState "Washington";
  geospecies:hasCommonName "Canada Goose";
  wildlife:hasYearCollected "2007";
  wildlife:hasMonthCollected ?month;
  wildlife:hasObservationCount ?count.}
GROUP BY ?month

(6)

24 http://www.w3.org/2003/01/geo/wgs84_pos
It can be challenging to retrieve, aggregate and visualize data not encoded in semantic format. For instance, the Washington Department of Fish and Wildlife provide species distribution data in a spreadsheet. To retrieve the data to be aggregated and then perform the aggregation, a resource manager has three options: do it manually, write ad hoc programs, or write complex excel macros. All of the three options require considerable time and effort from the resource manager.

The SemantEco portal has many potential directions to explore. We can use semantic technologies to enable automatic analysis over ecological and environmental data. For example, we model "EndangeredSpot" as a place where some animals are reported as dead or sick, and "CriticalSite" as a location that is both an "EndangeredSpot" and a "PollutedSite". Then, if we feed a semantic reasoner with the ontology and data, it will automatically identify the critical sites.

To identify "EndangeredSpot", the portal needs data for wildlife health events provided monitoring systems such as the Wildlife Health Event Reporter (WHER) and enable the portal to identify "EndangeredSpot" nearby. Then we can get the location "CriticalSite" as a "PollutedSite" having at least one wildlife health are usually monitored at different locations, the constraint usually does not hold. In such cases, we can model "CriticalSite" as a "PollutedSite" having at least one "EndangeredSpot" nearby. Then we can get the location information of "PollutedSite" and "EndangeredSpot" with query snippet (8) and utilize SPARQL filter to identify "CriticalSite" as shown in (9).

\[ \text{:EndangeredSpot rdfs:subClassOf} [\text{rdf:type owl:Class}; \text{owl:intersectionOf} (\text{:SickEventSpot :DeathEventSpot})]. \text{:SickEventSpot owl:equivalentClass} [\text{rdf:type owl:Restriction}; \text{owl:onProperty :hasWildlifeEvent}; \text{owl:someValuesFrom :WildlifeSickEvent}]. \text{:DeathEventSpot owl:equivalentClass} [\text{rdf:type owl:Restriction}; \text{owl:onProperty :hasWildlifeEvent}; \text{owl:someValuesFrom :WildlifeDeathEvent}]. \text{:CriticalSite rdfs:subClassOf} [\text{rdf:type owl:Class}; \text{owl:intersectionOf} (\text{:PollutedSite :EndangeredSpot})]. \]  

(7)

Such modeling and reasoning has a constraint: a monitoring site has records for both environmental observation and wildlife health events. However, as environmental qualities and wildlife health are usually monitored at different locations, the constraint usually does not hold. In such cases, we can model "CriticalSite" as a "PollutedSite" having at least one "EndangeredSpot" nearby. Then we can get the location information of "PollutedSite" and "EndangeredSpot" with query snippet (8) and utilize SPARQL filter to identify "CriticalSite" as shown in (9).

\[ \text{FILTER (} ?\text{siteLat < (?spotLat}+\text{delta+}) \&\& ?\text{siteLat > (?spotLat}+\text{-delta+}) \&\& ?\text{siteLong < (?spotLong}+\text{delta+}) \&\& ?\text{siteLong > (?spotLong}+\text{-delta+}) \text{)}. \]  

(9)

Next, for modeling wildlife habitat, we plan to connect to the Harmonisa project [25] that provides semantic descriptions of land-use and land-cover categories. Furthermore, EPA’s water quality criteria [15] provides multiple types of thresholds and would provide and important facet to complement our existing datasets. These criteria include measures of acute pollution in freshwater, chronic pollution in freshwater, acute pollution in saltwater, and chronic pollution in saltwater. We currently incorporate thresholds for acute pollution in freshwater for two reasons: 1) we mainly focus on inland water bodies; 2) acute pollution can affect both species that live near the polluting water source and pass by the water source occasionally. To support thresholds for chronic pollution, we need to consider some additional factors, e.g. the time that species stay near the polluting water source. We would require species distribution models from animal experts for modeling the chronic pollution. We plan to further enhance our modeling of rationale as provenance and to collect and integrate more data on the health effects of pollution on species.

Lastly, as these datasets continue to grow in size, work in distributed reasoning systems over the web will be critical to providing robust solutions without needing large data centers for performing data integration. Distributed systems would enable the ability for low powered solutions to operate on data streams reported by sensor equipment and provide publishing of pollution events from the field.

V. CONCLUSION

We extended our SemantEco portal based on two driven factors: to facilitate decision support systems for resource managers and to make the portal more broadly reusable. Our extension includes: support for wildlife monitoring; connections to OBOE, integration of wildlife observation data as linked data; enhanced provenance support through the incorporation of rationale; and performance comparison between a standard reasoner and customized rule base reasoner.

ACKNOWLEDGMENT

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, NSF, and USGS. E.W.P. is supported by a National Science Foundation Graduate Research Fellowship.

REFERENCES


25 http://www.whmn.org/wher/


