Ontology Modeling for Oceans Knowledge and Data Management

Samina R. Abidi, Syed SR. Abidi, Mei Kwan, Ali Daniyal

Abstract**—** We present a knowledge modeling solution for representing knowledge about oceans and marine life. Taking a knowledge management approach, we have employed semantic web technologies to model both the domain knowledge and the service descriptions in terms of domain and service ontologies. We present details of the domain and service ontologies, and demonstrate how these ontologies are applied to achieve seamless interoperability between the different data-sets and services. The ontologies are applied to an E-Science platform that is supported by Canada's CANARIE high bandwidth network that enables the transfer of large volumes of data from multiple sites and the execution of the various web services.

Keywords**—**Knowledge Management, Oceans, Ontology, Services-Oriented Architectures, Web Services.

I. Introduction

THE affects of global environment changes are impacting our ecosystem and our oceans are affected in ways that need to be better understood to protect and conserve our coast lines and the marine lives in the oceans. One of the major challenges faced by the scientific community is to retrieve specialized, multi-modal data from global data repositories, then to link the relatively sparse observations on marine life with highly voluminous ocean data, and finally to derive insights from the integrated data through a series of scientific experiments and multi-faceted visualizations. The dispersion of ocean and marine life data across the world, in different formats, demands a collaborative approach for data collection, procurement and operationalization—this approach needs to be supplemented with existing knowledge artifacts (reports, scientific studies, research publications) in order to establish interoperability between the different data sources and also to make sense of the scientific experimental results. So, we need both a collaborative platform to engage researchers from around the world and a scientific platform that provides state-of-the art tools to understand with ocean and marine life data [1, 2].

Raza Abidi is with the Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada. (phone: 1-902-4942129; e-mail: sraza@ cs.dal.ca).

Samina Abidi is a post-doc at the Faculty of Computer Science, Dalhousie

Ali Daniyal is a graduate student at the Faculty of Computer Science, Dalhousie University, Halifax, NS, Canada. (e-mail: daniyal@cs.dal.ca).

Taking a knowledge management approach we have developed an E-science platform—termed as *Platform for Ocean Knowledge Management (POKM)*—to support the oceanographic research community. The functional portfolio of POKM includes a suite of services to (a) *enable* the selection and sharing of multi-modal data collected from different geographic sites, (b) *perform* analytics and simulations, using complex simulation models, to understand various oceanographic phenomenon; (c) *visualize* multiple data layers at a geographic location and simulation results of models; (d) *publish* simulation models for use by the entire community of scientists; (e) *interconnect* two different research communities; and (f) *enable* researchers to design and execute complex experiments by composing specialized experimental workflows. POKM is supported by the CANARIE network (Canada's high bandwidth network) that allows the transfer of high-volumes of ocean data and to facilitate collaboration between eco-scientists across the world to conduct multi-site scientific experiments.

In this paper, we focus on the knowledge management components of POKM. We discuss the knowledge modeling research leading to the definition of two core ontologies—a domain ontology modeling the ocean knowledge and data resources and a service ontology modeling the web service descriptions. We demonstrate how the ontologies are applied to support the E-Science platform—i.e POKM.

II. Functional Design of E-Science Platform

The design of POKM showcases a unique synergy of semantic web, services oriented architectures, web services and visualization technologies. POKM takes a unique knowledge management approach by exploiting semantic web technologies to semantically describe the data, scientific models, knowledge artifacts and web services. POKM pursues a high-level abstraction of ocean and marine science domains to establish a high-level conceptual interoperability between the two domains. This is achieved by developing a rich domain ontology that captures concepts from both domains and interrelates them to establish conceptual, terminological and data interoperability. To define the functional aspects of the e-science services we have developed a services ontology that provides a semantic description of knowledge-centric e-research services. These semantic descriptions of the e-science services are used to both establish correlations between domain and functional concepts that are the basis for data and knowledge sharing, cataloguing and visualization. Fig. 1 shows the layered functional architecture of POKM.

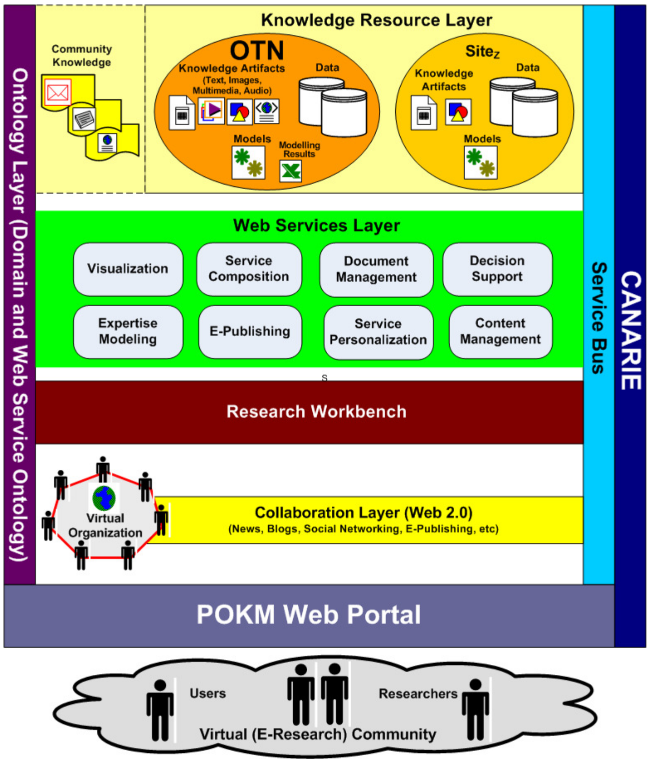

Fig. 1 Functional design of POKM

III. Ontology Based Domain Modeling

The domain ontology in POKM serves as the formal semantic description of the concepts and relationships pertaining to the Marine Biology and the Oceanography domain. POKM provides a core ontology that contains concepts necessary for modelling Marine Animal Detection Data (MADD), Oceanography Data, data transformations and interfaces of the Web Services in POKM.

The taxonomic hierarchy of the domain ontology constitutes 20 highest level classes.; 15 of these classes are further decomposed into sub-classes at the lower levels of hierarchy. The domain ontology is developed in OWL and represents six classes of ocean and marine life concepts.

A. Modeling Marine Sciences

There are six upper level classes related to marine sciences—i.e. Marine Organism, Animal Detail, Taxonomy, Taxon ID, Marine Life Data, Marine Life Data Collection, Data Source, Data Format.

Marine Organism represents all marine animals, plants and plankton via classes Marine Animal, Marine Plant and Plankton respectively. There are four main subclasses of Marine Animal: Fish, Marine Mammal, Reptile and Sea Bird. Marine Plant has two main sub-classes: Algae and Sea grasses. Plankton has three sub-classes representing three functional groups of planktons: Bacterio plankton, Phytoplankton and Zooplankton.

Animal Detail represents all the necessary information to build a marine animal profile. It has five main sub-classes: Age, Life Stage, Movement Behavior, Sex, Tag ID. Age represents the age of the animal. Life Stage represents the current stage of the animal in life, e.g. adult, juvenile, sub-adult etc. Movement Behavior represents various movement behaviors of marine animal that are captured by its sub-classes: Behavioral Switching, Dispersal, Diving, Drift, Foraging, Migrating and Movement Pattern.

Taxonomy represents nine main taxonomic ranks used to categorize marine organisms as follows: Class, Family, Genus, Kingdom, Order, Phylum, Scientific Name, Scientific Name Author and Species.

Taxon ID describes an organism in terms of the above mentioned nine taxonomic ranks.

Marine Life Data represents various aspects of the data about the marine organisms. These include temporal data represented by sub-classes: Day Collected, Month Collected, Year Collected, Date Last Modified and Time Stamp Collected, which has two sub-classes of its own: End Time Stamp Collected and Start Time Stamp Collected. The class Marine Life Data is also used to represent concepts related to the cache of the marine data that is represented using sub-classes such as Cache ID, Record Last Cached, Basis Of Record and Resource ID. This class also represents other aspects of marine life data using sub-classes: Depth, Depth Precision, Temperature and Time Zone.

Marine Life Data Collection is a class the properties of which are used to capture all the data represented by class Marine Life Data.

B. Modeling Ocean Sciences

The classes to model ocean sciences include: Ocean Region, Ocean Parameter, Satellite Information, Instrument, Measure, Movement Model, Model Attribute, File Type, Ocean Region represents all ocean regions categorized by five main sub-classes: Arctic Ocean, Atlantic Ocean, Indian Ocean, Pacific Ocean and Southern Ocean. Each of these classes are further sub-divided into sub-classes representing sub regions of the each ocean region. Ocean Parameter represents all the geophysical parameters used to describe an ocean environment. These are modeled as sub-classes of this class and include parameters about:

- Air, such as: Air Temperature,
- Wind, such as: Wind Gust, Wind Speed, East Ward Wind, North Ward Wind, Up Ward Wind and Wind From Direction
- Water, such as: Water Depth, Water Temperature, Salinity and Density,
- Current, such as: Current To Direction, East Word Current, North Ward Current, Flow Velocity and Up Ward Current.
- Sea Layers, such as: Sea Surface Elevation, Sea Surface Temperature and Thermo cline

SATELLITE INFORMATION represents the satellite used to monitor the oceans, represented in terms of nine sub-classes: SATELLITE ID, ALTITUDE, BEST SIGNAL STRENGTH, FREQUENCY OF TRANSMISSION, ELAPSED TIME, NUM OF MESSAGES RECEIVED, NUM OF SUCCESSFUL PLAUSIBLE CHECKS, QUALITY INDICATOR and SENSOR CHANNEL.

INSTRUMENT represents all the instruments used for the observation of oceans and to measure various parameters, such as: temperature, salinity and density of the ocean water, ocean currents, depth, pressure, etc. These instruments are represented as the following sub-classes: ADCP, ARGOS, ARGO FLOAT, CTD, ELECTRONIC TAG, GLIDER, GLOBAL POSITIONING SYSTEM, SATELLITE and SUBMERSIBLE RADIOMETER.

MEASURE represents all the spatial and temporal measures of the regions used in the domain of Ocean Sciences, and are modelled as two main sub-classes SPATIAL MEASURE and TEMPORAL MEASURE respectively. The sub-class SPATIAL MEASURE has further sub-classes: HEIGHT, LATITUDE, LONGITUDE AND SPATIAL RESOLUTION representing the respective spatial measures of the relevant ocean region. TEMPORAL MEASURE has two sub-classes: TIME INTERVAL and TIME RESOLUTION, representing the respective temporal measures.

MOVEMENT MODEL represents various models used to estimate the migrating and foraging behaviors of marine organisms and their movement parameters such as determining the next positioning estimate of an animal after a period of missing data. These models are represented as sub-classes: FIRST PASSAGE TIME, FRACTAL ANALYSIS, GEO LOCATION MODEL, KERNEL ANALYSIS, STATE SPACE MODEL.

MODEL ATTRIBUTE represents all the attributes of a movement model, represented as sub-classes: HIERARCHY, LINEARITY, OBSERVATION ERROR, OUTPUT, STATISTICAL ESTIMATION ERROR, STATISTICAL FRAMEWORK, STOCHASTICITY and TIME.

UNIT represents all the units used to measure geophysical parameters describing an ocean. It has nine sub-classes: DENSITY UNIT, DEPTH UNIT, LIGHT LEVEL UNIT, SALINITY UNIT, SPATIAL RESOLUTION UNIT, SPATIAL UNIT, TEMPERATURE UNIT, TIME UNIT, VELOCITY UNIT.

C. Modeling Information Sources

The domain ontology contains classes to represent the data used in the two domains.

DATA SOURCE represents the sources of the marine life data, such as: Net CDF, OBIS, OTN and TEST.

DATA FORMAT represents the format in which the marine life data are stored, e.g. List with or without header or String with or without header.

FILE TYPE represents file formats used to display geographic data (such as KML) or for digital storage of data (such as CSV).

TABLE FORMAT describes the tables through two sub-classes COLUMN NAME and TABLE NAME.

USER CREDENTIALS represents the credentials of the users.

D. Relationships Between Classes

The purpose of the domain ontology is to inter-relate the domains of Marine Sciences and Ocean Sciences. There are seventy seven object properties and six data type properties. We describe only the salient properties are described in this section.

The class MARINE ANIMAL (sub-class of MARINE ORGANISM) is related to respective sub-classes of the class MARINE LIFE DATA through properties *has_age, has_sex, has_life_stage, has_movement_behavior* and *has_TagID.* In addition it is also related to class OCEAN REGION through property *has_geographic_area*. Thus, this property relates the domains of marine sciences and ocean sciences.

The class OCEAN PARAMETER is related to class Unit through property *has_unit*. This property is given has Value restriction, to restrict the filler of the property to a specific instance of the class UNIT. For example Air Temperature, which is an Ocean Parameter *has_unit* Degree Celsius, which is an instance to class UNIT.

The class MARINE LIFE DATA COLLECTION is related to respective sub-classes of class MARINE LIFE DATA through properties *has_basis_of_record, has_cache_ID, has_date _last_modified, has_day_collected, has_depth, has_depth_precision, has_latitude, has_longitude, has_month_collected, has_record_last_cached, has_record_ID, has_taxon_ID, has_temperature, has_time_of_display_collected, has_time_zone_collected* and *has_year_collected.* Each one of these properties is a functional property.

The class MOVEMENT MODEL is related to respective sub-classes of class MODEL ATTRIBUTE through properties: *has_hierarchical, has_input_data, has_linearity, has_observation_error, has_output, has_statistical_ estimation_method, has_statistical_framework, has_stochasticity* and *has_time_value*.

Each OCEAN REGION is related to various OCEAN PARAMETERS through properties: *has_density, has_flow_velocity, has_salinity, has_sea_surface_elevation, has_water_depth, has_water_mass* and *water_temperature*. Class OCEAN REGION is also related to respective sub-classes of class MARINE LIFE through sub-classes *has_marine_animal, has_marine_plant* and *has_plankton.* Note that these three sub-classes relate the ocean sciences domain with marine sciences domain.

The class TAXON ID is related with respective sub-classes of class TAXONOMY, in order to capture the identification features of each of the marine species. These properties are: *has_class, has_family, has_genus, has_kingdom, has_order, has_phylum, has_sceintific_name, had_scientific_name_author* and *has_species*. Each one of these properties is a functional property.

The class SATELLITE is related to respective sub-classes of SATELLITE INFORMATION through properties: *has_altitude, has_best_signal_strength, has_elapsed_time, has_frequency_of_transmission, has_num_of_messages_received, has_num_of_successful_plausible_check,* *has_quality_indicator, has_satellite_ID* and *has_sensor_chanel*.

E. Capturing Semantics of Web Services using OWL-S

To enable the system to capture semantic description of Web Services, POKM employs OWL-S ontology in conjunction with the Domain Ontology. OWL-S provides a semantic model for capturing three aspects of a Web Service: namely; Profile, Process Model and Grounding. Since we employ UDDI as a registry for web services, the Profile part of the OWL-S ontology is not instantiated in POKM, to avoid duplication. POKM only employs the Process Model to capture IOPE of the Web Services and Grounding to capture semantics of the concrete WSDL description of web services.

In OWL-S a Service is described by a Process, whereby a Process can be exactly one of Atomic Process, Simple Process or Composite Process. On the other hand a WSDL Service is a collection of operations; and each operation accepts exactly one input message and returns exactly one output message after execution. Therefore, an Atomic Process in OWL-S is used to capture semantics of an operation of a WSDL Service. Based on this correspondence between OWL-S and WSDL, a unique instance of Service is generated in OWL-S for every operation in the WSDL description of a web service.

The Process Model in OWL-S provides classes to capture the notions of Parameters, Inputs, Outputs, Preconditions and un/conditional Effects of a process. The property parameter Type is used to relate a parameter with its semantic type in terms of an OWL class. An Atomic Process in OWL-S may have multiple inputs and multiple outputs, however when modeling semantics of an operation of a web service as an OWL-S Atomic Process, they are mapped to the parts of the input and output messages. This mapping is captured via the WSDL Grounding class in the Grounding part of OWL-S.

Each Atomic Process in OWL-S supports exactly one WSDL Grounding that in turn provides a unique WSDL Atomic Process Grounding. This class facilitates capturing mapping between the message parts of the web service operation and the parameters of the OWL-S Atomic Process. Fig. 4 illustrates how OWL-S is employed in POKM to describe semantics of web services.

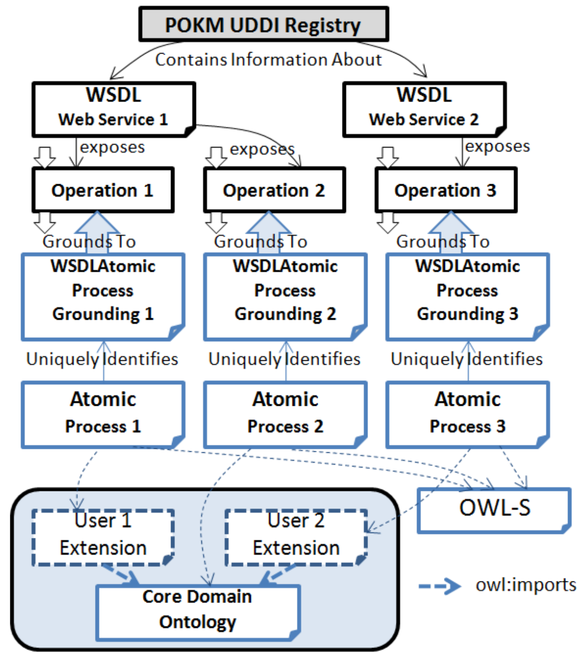

Fig. 2 Semantic Modeling of Web Services in POKM using OWL-S

F. Web Services Deployment and Publishing Life-Cycle

One of the main functionalities of the POKM is to allow users to enrich the pool of web services accessible via POKM. The main intention behind this functionality is to enable scientists to publish their scientific models as web services on the POKM, so that these models can be used in more complex processes and can be shared with other users. To achieve standardization and interoperability all the scientific models are exposed as web services, described by WSDL and compliant with the SOAP messaging protocol. POKM supports three different scenarios of publishing web services on the system: namely, (i) deploying WAR file on the POKM ESB, (ii) publishing scientific models as R-scripts, and (iii) publishing WSDL file on POKM UDDI. Fig. 2 illustrates the process of publishing web services description in the POKM system for these three scenarios.

IV. Concluding Remarks

The semantic modeling of data and knowledge within POKM allows to seamlessly connect multiple data sources. Modeling of the domain concepts in the domain ontology enables the integration of the data sources by mapping both schemas and attributes based on concepts as opposed to labels. Furthermore, the domain ontology allows the use of domain concepts to generate abstract data fetching queries, in terms of standard operations, that can provide access to heterogeneous data sources. Fig. 3 and Fig. 4 provide snapshots of the two portlets in the POKM portal that make use of the oceanographic ontology to request detection data from a selected data source and to visualize the data.

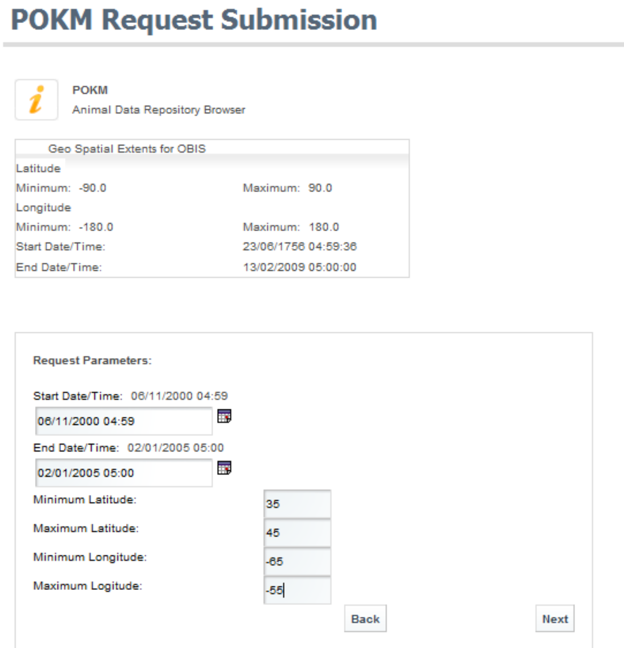

Fig. 3 Portlet for requesting marine animal data on the POKM

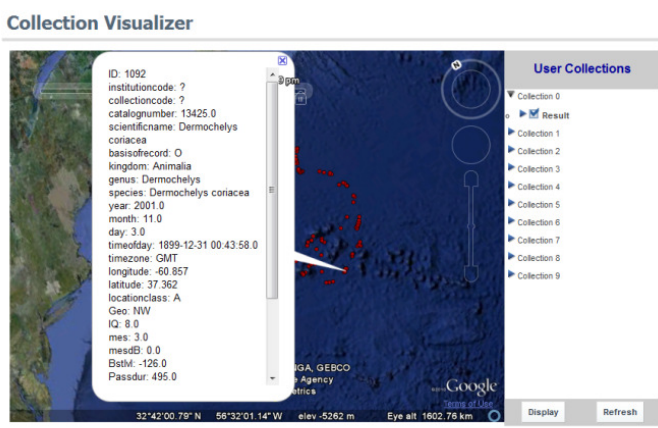

Fig. 4 The Collection Visualizer portlet to obtain requested data in renderable format. The portlet uses GE Plugin to render KML files generated using MADRS.

POKM offers an E-science platform targeting ocean sciences researchers. The key feature of POKM is the underlying knowledge layer—manifested in terms of the semantics of ocean and marine life concepts, captured in the domain ontology and descriptions of the services captured in the service ontology—that enables the seamless integration and interoperability at the data, services and user levels. The domain ontology presented here is scalable to include new concepts and relationships between concepts. The services ontology is coupled with interesting semantic-based methods to discover compatible and relevant services pertaining to a specific user task. In this paper, we have demonstrated the potential of applying knowledge management methods, specifically the use of OWL-based ontologies to semantically describe the domain and service concepts and to operationalize these concepts in terms of e-science services.

The POKM project is currently in operation and used by a community of ocean and marine life scientists in Canada. These scientists are using POKM to study the migration patterns of leatherback turtles across the Nova Scotian shelf in the Atlantic ocean.

Acknowledgment

This project is funded by CANARIE Canada under the Network Enabled Program (NEP). The authors acknowledge the support of CANARIE Canada to develop POKM. The authors acknowledge the support of members of the POKM team, especially the marine biology, ocean data modeling and document management teams.

References

[1] N. Bos, A. Zimmerman, J. Olson, J. Yew, J. Yerkie,, E. Dahl, “From shared databases to communities of practice: A taxonomy of collaboratories”, in *Journal of Computer-Mediated Communication*, vol. 12(2), 2007, pp. 652-672.

[2] J. N. Cummings, S. Kiesler, “Collaborative research across disciplinary and organizational boundaries,” in *Social Studies of Science*, vol. 35(5), 2005, pp. 703-722.