Proceedings of the 11th International Coral Reef Symposium, Ft. Lauderdale, Florida, 7-11 July 2008
Session number 16

The Semantic Reef: A hypothesis-based, eco-informatics platform to support automated knowledge discovery for remotely monitored reef systems.

T. S. Myers[1], I. M. Atkinson[1]

1) James Cook University, School of Mathematics Physics and IT, Townsville, Queensland, Australia.

Abstract. Eco-informatics is the combination of multiple environmental datasets and modeling tools used to test ecological hypotheses and derive information. The Semantic Reef project is developing automated data processing, problem-solving and knowledge discovery systems to better understand and manage reef ecosystems. Three core tools are utilized; Semantic Web, Grid Computing and workflow based e-Research approaches, creating a platform designed to evaluate complex hypothesis queries and provide alerting for unusual events (e.g. spawning, bleaching). Remote environmental monitoring is being widely deployed to gather data in real-time. As the volume of raw data increases, bottlenecks are expected to develop in the data analysis phases - current data processing still involves human intervention and manual manipulation. Semantic Web technologies address this data deluge through using explicit descriptions, called ontologies, of the datasets and structures, making the data machine-understandable, therefore automating data integration and processing. The Semantic Reef project is focused on developing reef ontologies, which, when coupled to datasets, derives inferences from data to 'ask' the system questions for semantic correlation and analysis. Currently, the model is being extended to map dynamic data from reef-based sensor-networks into the ontology in real-time, offering a new approach to solving problems of scale across reefs.

Key words: Eco-informatics, semantic web, ontologies, coral reefs.

Introduction

With the global effects of climate change, and other major issues, eco-informatics is an emerging branch of research where new techniques, tools and infrastructure are being developed to enable far greater scope for problem solving methodologies and analytical ability for the research scientist. This new cross-discipline evolution of global collaborations, with both ecological and computer scientists, is vital in the quest for knowledge and answers when addressing such major issues as climate change.

Problems are arising, however, with the imminent influx of new data and information, appropriately dubbed 'the data deluge'. This increasing flood of data is growing exponentially with the large number of deployed, or soon to be implemented, scientific data collection instruments such as sensors, satellites, scientific experiments and simulations. Hence, as the volume of raw data increases, bottlenecks in the data processing and analysis phases are occurring because current methods still involve human intervention making it progressively more difficult for the scientist to keep up effectively.

The Semantic Reef Project aims to utilize existing coral reef databases, augmented by real time sensor output, to pose hypotheses of the disparate data. The project is a Knowledge Representation (KR) system that employs Semantic Web, scientific workflows and Grid computing technologies to resolve the problems of data integration, synthesis and discovery for coral reef ecosystems (Fig. 1). It will map static and dynamic data to a set of ecosystem ontologies, which through explicit definitions will make the information machine-understandable, enabling the computer to make intelligent inferences, decisions and/or

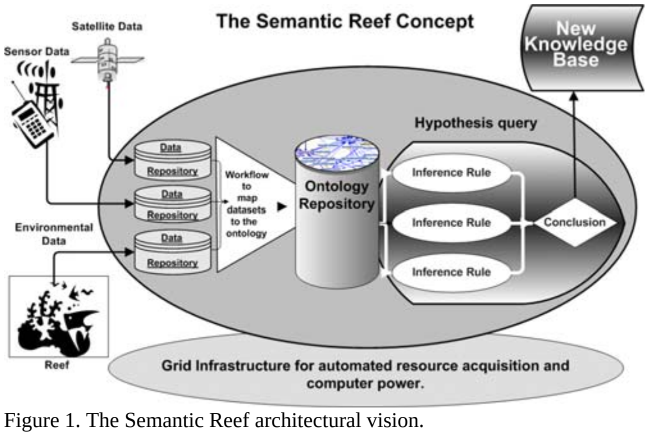


Figure 1. The Semantic Reef architectural vision.

discoveries, using logic systems such as Description Logics (DL) and propositional logic (Myers et al. 2007a).

A brief description of the Semantic Reef project, results to date and the data processing issues marine biologists are facing will be followed by a closer look at Semantic technologies, and how they may address this data deluge problem. Finally, details of the KR system and the reusable design methodologies, to semantically represent a full coral reef ecosystem, will be explained.

Background – The Semantic Reef project

The Semantic Reef project is developing a tool for use in hypothesis-driven research and problem-solving methods. It will allow for automated data processing and analysis of disparate databases and data streams through the technologies enlisted in the architecture. These technologies include Semantic Web, Grid computing and scientific workflows, which together create synergies to address particular issues in data integration. The resulting technology platform is designed to improve our capacity to generate timely warnings of environmental conditions that are conducive to climate change issues or unusual events within the coral reef environment.

At the heart of the Semantic Reef project are the Semantic Web technologies, which are an emerging area of computer science that aims to support automated processing of information. Through ontologies, which are the foundation of Semantic technologies, concepts are explicitly described, giving context and meaning to the data the computer can access. Here, a set of re-usable ontologies have been developed to describe to a computer the concept of, and the relationships within, a coral reef ecosystem.

The model is built on a Grid Computing foundation which provides the tools for secure and reliable access, sharing and management of resources, namely raw data and computational power (Foster et al. 2001). The scientific workflows are developed using the Kepler workflow software (Altintas et al. 2004), to automatically process raw data, and pass the results through the expert system.

To assess the accuracy of the architecture, validation tests were conducted. Initially, with a focus on the coral bleaching phenomena, a reverse hypothesis methodology to ground truth the system was employed and the outcome of the inference propositions were compared with actual historic bleaching events. For the validation, the stress factor most commonly associated with the bleaching event, namely elevated sea temperature (Marshall and Schuttenberg 2006), was used with the available historical data acquired through the Great Barrier Reef Marine Park Authority (GBRMPA) on the 1998 and 2002 mass bleaching events. The thermal stress indices trialed were the Sea Surface Temperature plus (SST+) and Degree Heating Days (DHDs), which are commonly used to describe thermal anomalies (Berkelmans et al. 2004; Maynard et al. 2008). Logical inference rules and DL were used to mimic these metrics then executed using the historic temperature data. The results of the automated inference were found to relate closely to those of previous research on the tolerance of corals to temperature changes, thus verifying a successful test (Myers et al. 2007b).

Since the initial validation exercise, a richer set of ontologies have been developed. The hierarchical design employs re-usable multi-scale ontologies to represent the functionality and complexity of a coral reef ecosystem making the concept computer-understandable for auto-processing objectives. The intention is to allow marine scientists, from around the globe, tools to test hypotheses and probabilities, through enlisting a semantic ‘richness on demand’ environment. The ontologies, ranging in complexity, describe concepts such as the composite population of the reef community, human influence contributors, bathymetry, environmental factors, among others, in computer-understandable semantics. Using this modular approach, reusability is afforded, where the knowledge being sought or the hypothesis being posed will determine the choice of which ontologies to utilize and what data is used to populate them. For example, extending the validation inference rules used for a bleaching occurrence, which are currently based solely on SST, to introduce other premises such as water quality, location, nutrient levels, salinity, etc, would be a trivial task programmatically.

The Data Deluge

Many scientific disciplines are experiencing changes in how research is being performed due to digital technology from the next generation of experiments, simulations, sensors and satellites generating a flood of valuable data for scientists to interpret (Hey and Trefethen 2003). The data is being gathered globally in differing formats and for different agendas and the bridging between all these disparate sets is the real issue that many technological developments, such as the one described here, are attempting to address.

The use of environmental sensor networks to gather data in real-time across widely distributed areas is an expanding field, detailed by projects such as the Integrated Coral Observing Network (ICON) in the USA (NOAA-ICON/CREWS 2008) and the Integrated Marine Observing System (IMOS) in Australia (IMOS 2008). Applications of new technologies and processing systems are being trialed on the Great Barrier Reef (GBR) such as the Great

Barrier Reef Oceans Observing System (GBROOS 2008). As habitat researchers deploy embedded sensor networks, strategies on capturing, organizing, and managing large amounts of streaming live data is becoming a critical issue. It is imperative the processing and analyzing of the data be effective and efficient. However, it is becoming evident that current technologies are not scalable enough to quell the imminent deluge or make proficient methods of analysis and research possible.

The data is being gathered and stored in a myriad of different repositories and although there are data mining efforts to bridge across this disparate data most are still considered stand-alone unconnected data silos. So how do we search and infer new knowledge across the breadth of available data automatically? With the main focus of data accessibility, integration and automated processing, Semantic Web technologies offer a possible solution.

Semantic Web Technologies

Semantics is the study of meaning and dates back millennia in the philosophical arena. The term ‘Semantic Web’ was coined by Sir Tim Berners-Lee in his original proposal to CERN to develop the World Wide Web (Berners-Lee 2000). He described the Semantic Web as being the evolution from a Web of cross platform documents to a ‘Web of data’, where complex decision making by the machine will be possible as the information contained within the web pages will be both human-readable and computer-understandable. Ideally, with the decrease in manual intervention, it will allow search and analysis mechanisms greater autonomy to sift through the massive amounts of knowledge and data available on the Web and automate the process of creating new knowledge.

The Ontology

At the heart of Semantic technologies is the ontology. Ontologies are documents or files that formally define the relations among terms and representing abstract or specific concepts: intentions, beliefs, objects or feelings. These descriptions contain explicit specifications, terms and relationships with formal definitions, axioms and restrictions that constrain the interpretation by the computer, thus making data computer-understandable enabling it to make intelligent decisions based on inference rules and DL (Antoniou et al. 2001). Of course, the computer does not literally ‘understand’ the information, but it can now manipulate the terms much more effectively in ways that are useful and meaningful to the human user, such as being more functional in cross platform environments when dealing with disparate data from diverse sources.

Modern ontologies can be engineered at differing levels of complexity, from shallow (i.e. a domain vocabulary, thesaurus, or taxonomy) to the composite category, which involves applications of logical definitions to automate conclusions, assumptions and subsumptions through classification and inferences. Creating unambiguous definitions of things, concepts and ideas assists in bridging across disparate data sets, for example, well defined synonymous relationships would allow ‘*Acropora formosa*’ in one database schema to be easily bridged to ‘stag horn coral’ in another. In addition, defining homonyms by adding context affords the ability for the computer to interpret contextual information. For example, the word ‘fluke’, to the machine, is five characters of eight binary bits each, it has no meaning to the computer, however, associating the word with a ‘part of’ or an ‘is a’ relationship to another word, such as anchor, dolphin, fish or flatworm, adds enough contextual information so the computer can make meaningful connections automatically.

The Logic Systems

There are number of differing formal logic systems available within the scope of Semantic technologies each orthogonal to the next. The field of DL is a subset of First Order Logic (FOL) and has been used throughout history, it allows for expressing relations between concepts in a generalized fashion for reasoning with logical axioms and quantifications (Baader 2003). To give a simple example, an omnivore could be generally defined as, among other things, any animal that eats both plants and animals. Translated into the Web Ontology Language with DL (OWL-DL) (McGuinness and van Harmelen 2004) would be written as:

```
OmnivoreClass eats SOME
(PrimaryProducerClass AND
(HerbivoreClass OR CarnivoreClass))
```

Upon classifying using a reasoning engine, any instances that are defined with the property ‘eats’ and is linked to both a member of a plant class and animal class (e.g. an instance of the Sea_Grass class and the *Bivalvia* class respectively), will automatically be subsumed to belong to the omnivore class (e.g. all instances of the Loggerhead_Turtle class). This ability of the KR system to automate latent connections and make dynamic inferences of relations (i.e. automatically connecting the hidden dots), is imperative for modeling such an intricate multi-scaled concept as a reef ecosystem.

Propositional logic adds the ability to infer conclusions based on sets of predefined premises. The Semantic Web Rules Language (SWRL) uses ‘horn-like’ rules, composed as syllogisms, to produce logical conclusions (Horrocks et al. 2004), which allow for hypothesizing over the full knowledge base. Posing hypothetical questions with known or best guess factors can allow for conclusions to be drawn by deductive and inductive inference, where observation would prove or disprove the hypothesis.

The Open World Assumption

Marine science, in general, maintains an Open World Assumption (OWA), that is, nothing is false until explicitly proven false. Traditional database implementations such as relational databases are required to maintain a Closed World Assumption (CWA), that is, everything that is known about the world exists within the boundaries of the database and its schema. This creates a mismatch between the researchers need for dynamic multi-scale complexity open to changes upon new discoveries and the current technological capabilities for flexibility.

Semantic Web languages have an OWA where it is assumed knowledge of the world is incomplete. Specifically, 'not true' is not automatically false, it is considered unknown; it assumes the extra information required has not yet been added to the knowledge base. As new discoveries are quite feasible in the marine biology domain, there is a need for flexibility allowing new knowledge and concepts to be assimilated into the system without difficulty; the OWA caters to this need (Horrocks et al. 2003).

In illustrating the differences between OWA and CWA, when describing the makeup of a particular reef the schema includes only these two statements: 'Davies Reef has *Favites*' and 'Davies Reef has *Porites*', a query of whether Davies Reef contains any *Acropora* would return false in a CWA, which would be incorrect. However, with the OWA, unless there is a statement that explicitly declares 'Davies Reef has no *Acropora*', the system will conclude there may be *Acropora*, it simply has not yet been explicitly asserted.

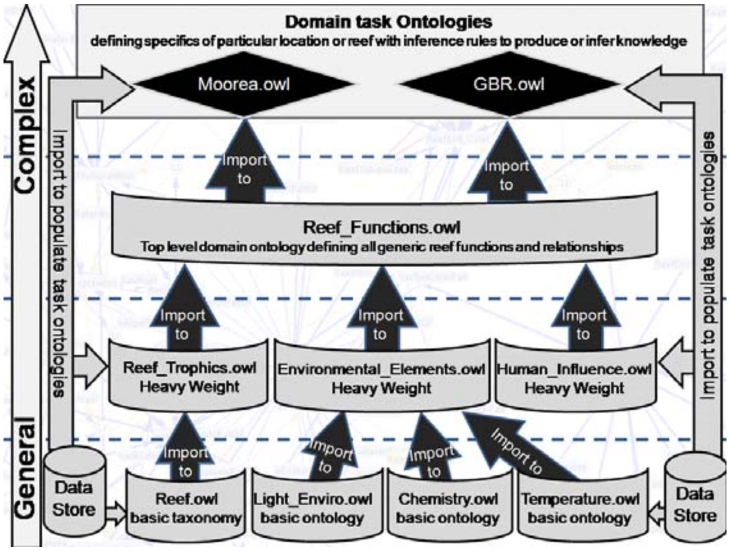
Building the Semantic Reef ecosystem ontologies for multi-scalability

Concepts can be modeled through ontologies in a variety of ways and varying degrees of granularity. The Engineering and design choices on what type of ontology to build is determined purely by the degree of extensibility and expressiveness required to produce the desired information or knowledge. However, the flexibility and reusability becomes more restricted, in ontological design, as complexity increases (Gomez-Perez et al. 2004).

Clearly, coral reefs are highly complex, interdependent ecosystems. It became apparent the use of a singular top-down, large ecosystem ontology would not be simple to create, implement nor maintain, as the number of variables and multi-scale relationships are immense. Therefore, a 'mix and match' bottom-up approach in the ontological design was adopted. Specifically, a range of separate ontologies were created to describe the many diverse concepts that make up a coral reef, such as, community and environmental composition, hydrodynamics, human influence and trophic layers, among others.

The current set of ontologies consist of reef and environmental taxonomies at the lowest layers, which import to heavier-weight DL ontologies for concepts such as trophic layers that require richer relational descriptions (Fig. 2). To illustrate, the reef community composition is an uncomplicated ontology, where only synonymous and hyponymous (i.e. 'same as' and 'is a') relations are defined. When populating the ontology, this method allows for a less complicated bridging mechanism across disparate databases, which may list either scientific names or common names, but not both. This and other simplified ontologies, such as one that describes and maintains temperature information, can then be ported to the heavier-weight ontologies. Separating the logical complexities from the instance data at the lower layers, makes ontological structure reusable.

'Domain task specific' ontologies lie at the highest level of granularity. It is at this level the finely detailed rules are introduced as propositions, written in SWRL, to infer conclusions from the available data within the knowledge base. For example, to hypothesize about coral bleaching on the GBR would have all lower ontologies populated with available or relevant data (e.g. salinity, chlorophyll, real-time temporal and spatial SST data, etc), then imported into a domain task GBR ontology, where inference



Domain task Ontologies
defining specifics of particular location or reef with inference rules to produce or infer knowledge
Moorea.owl
GBR.owl
Complex
General
Import to populate task ontologies
Import to populate task ontologies
Import to
Import to
Reef_Functions.owl
Top level domain ontology defining all generic reef functions and relationships
Import to
Import to
Import to
Reef_Trophics.owl Heavy Weight
Environmental_Elements.owl Heavy Weight
Human_Influence.owl Heavy Weight
Import to
Import to
Import to
Import to
Data Store
Reef.owl basic taxonomy
Light_Enviro.owl basic ontology
Chemistry.owl basic ontology
Temperature.owl basic ontology
Data Store

Figure 2. The bottom-up, hierarchical ontology design.

rules can be structured to infer a bleach alert. As the infrastructure is designed to be generically reusable, the same ontologies, populated with different location specific data, could be used for a different hypothesis on any reef, for example, a sensitivity analyses for coral spawning on Moorea Reef (Fig.2).

Future Works and Conclusions

We are drowning in a sea of data, which occasionally is generously referred to as 'information' and almost all of it must be interpreted by humans to be of any use. The growth and availability of new data sources and, therefore, our need to consider it in research, decision-making and planning is growing exponentially, and our systems, rather than helping with this, are predominantly contributing to the problem. Hence, the demand for automated data analysis and/or hypothesis-testing systems is becoming increasingly imperative as the escalating range of data gathering devices and instruments are deployed.

The Semantic Reef is a new approach to such data analysis and interpretation issues. The modular ontology design allows for inclusion of both scientifically known factors as well as phenomena yet to be discovered. Therefore, as user driven propositions change due to new findings, information or queries, and as new data sources become available, the open world nature of Semantic technologies will make uncomplicated additions and eliminations to the KR system possible.

Here, highly diverse backgrounds and expertise have combined effectively in collaboration to structure a semantically driven architecture to assist the marine biology domain in confronting the data deluge challenge. Currently, the processing capabilities of the Semantic Reef system are being tested to infer a coral bleach warning using satellite data from the US National Oceanic and Atmospheric Administration (NOAA 2008) and real time SST data streamed directly from the Davies Reef microwave site, part of the GBROOS project (GBROOS 2008). Further additions to the hypothesis will see the addition of other causal factors, as the data becomes available, such as chlorophyll provided by AIMS and salinity levels from IMOS.

Acknowledgement

We thank Scott Bainbridge and Stuart Kininmonth from AIMS and Sally Holbrook and Russell Schmitt from the Coral Reef Environmental Observatory Network (CREON) group for their interest, helpful discussions and feedback. We also thank Ron Johnstone from the CMS, UQ, M Palaniswami from the ARC Research Network on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP) and the team at the GBRMPA's Climate Change Response group for their integral role with content expertise and design advice.

References

Altintas I, Berkley C, Jaeger E, Jones M, Ludäscher B, Mock S (2004) Kepler: An Extensible System for Design and Execution of Scientific Workflows. 16th Intl Conf on Scientific and Statistical Database Management (SSDBM'04):21-23

Antoniou G, Billington D, Governatori G, Maher MJ (2001) Representation results for defeasible logic. ACM Trans Comput Logic 2:255-287

Baader F (2003) The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press

Berkelmans R, De'ath G, Kininmonth S, Skirving WJ (2004) A comparison of the 1998 and 2002 coral bleaching events on the Great Barrier Reef: spatial correlation, patterns, and predictions. Coral Reefs 23:74-83

Berners-Lee T (2000) Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web. Collins

Foster I, Kesselman C, Tuecke S (2001) The Anatomy of the Grid: enabling scalable virtual organizations. Int J Supercomputer Appl 15:200-222

GBROOS (2008) Great Barrier Reef Ocean Observing System. http://www.imos.org.au/nodes/great-barrier-reef-observing-system.html

Gomez-Perez A, Corcho O, Fernandez-Lopez M (2004) Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition (Advanced Information and Knowledge Processing). Springer, London

Hey AJG, Trefethen AE (2003) The Data Deluge: An e-Science Perspective. In: Berman F, Fox GC, Hey AJG (eds) Grid Computing - Making the Global Infrastructure a Reality. Wiley and Sons, pp 809-824

Horrocks I, Patel-Schneider PF, van Harmelen F (2003) From SHIQ and RDF to OWL: The making of a web ontology language. J Web Semantics 1:7-26

Horrocks I, Patel-Schneider PF, Boley H, Tabet S, Grosof B, Dean M (2004) SWRL: A Semantic Web Rule Language - Combining OWL and RuleML. W3C. http://www.w3.org/Submission/SWRL/

IMOS (2008) Integrated Marine Observing System (IMOS). http://imos.org.au/

Marshall P, Schuttenberg H (2006) A Reef Manager's Guide to Coral Bleaching. Great Barrier Reef Marine Park Authority, Townsville, Australia

Maynard JA, Turner PJ, Anthony KRN, Baird AH, Berkelmans R, Eakin CM, Johnson J, Marshall PA, Packer GR, Rea A, Willis BL (2008) ReefTemp: an interactive monitoring system for coral bleaching using high-resolution SST and improved stress predictors. Geophys Res Lett 35:1-5

McGuinness D, van Harmelen F (2004) OWL Web Ontology Language overview. W3C. http://www.w3.org/TR/owl-features/

Myers TS, Atkinson IM, Lavery WJ (2007a) The Semantic Reef: Managing Complex Knowledge to Predict Coral Bleaching on the Great Barrier Reef. 5th Australasian Symposium on Grid Computing and e-Research 68:59-67

Myers TS, Atkinson IM, Maynard J (2007b) The Semantic Reef: An eco-informatics approach for modelling coral bleaching within the Great Barrier Reef. ERE Environmental Research Event

NOAA-ICON/CREWS (2008) Integrated Coral Observing Network/Coral Reef Early Warning System. National Oceanic and Atmospheric Administration. http://www.coral.noaa.gov/crews/

NOAA (2008) National Oceanic and Atmospheric Administration. http://www.noaa.gov/