Eco-informatics modelling via semantic inference

Trina Myers [a, b] and Ian Atkinson [b].

[a] *James Cook University, School of Business (IT), Townsville, Queensland,4811, Australia.*
[b] *James Cook University, e-Research Centre, Townsville, Queensland,4811, Australia.*

Email: (trina.myers, ian.atkinson)@jcu.edu.au

Corresponding Author:
Dr Trina Myers
telephone: + 61 7 4781 6908
Facsimile:+ 61 7 4781 5880

ABSTRACT

There is a demand for new and evolved research practices resulting from the so called "data deluge" emerging from high volume digital collection methods. As the volume of raw data increases traditional data processing methodologies, especially those involving manual manipulation are becoming increasingly difficult to manage. This paper presents the "Semantic Reef" architecture that offers an alternative approach to the development, application and execution of observational hypotheses involving studies of coral reef ecosystems. The Semantic Reef Knowledge Representation system is an eco-informatics application designed to assist in the integration of remotely sensed data streams and historic data sets supporting flexible hypothesis design and knowledge extraction. The system is an ontology-based architecture built to allow researchers to combine disjoint data sets into a single Knowledge Base for modelling the impact of climate change on coral reef ecosystems. The Knowledge Base consists of a hierarchy of ontologies developed to maximise usability and reusability by separating data instances from the concept descriptions. The model can be effectively reused to extract or disclose phenomena of any coral reef. This paper both demonstrates and describes a performance analysis of the Semantic Reef knowledge system.

KEYWORDS: Semantic Web, ontology engineering, scientific workflows, knowledge systems

1. INTRODUCTION

The development of an alternative approach to extrapolate information from diverse data sources is the focus of this paper. The collection of real-time remotely sensed data across widely distributed locations is rapidly being developed and adopted through remote environmental monitoring (including sensor networks) [1-4]. Researchers are arguably finding it increasingly difficult to create timely and well managed information and knowledge and take advantage of all available data sets to inform their studies [1]. The integration of the growing amount of sensed data with other forms of data (e.g., satellite, models and/or historic data sets) is of immense value in the creation of new knowledge and is increasingly a requirement for environmental managers and researchers. However, to take advantage of all available data in a data deluge to create timely and well managed information is becoming progressively more difficult for researchers [1].

A feature of environmental research activity is that raw sensor and observational data is often collected by independent agencies and is often heterogeneous. Notably, similar information is often collected by different organisations but maintained in non-interoperable forms. For example, both the National Oceanic and Atmospheric Administration's (NOAA) and Australian Bureau of Meteorology (BoM) both gather weather data. However, the data is heterogeneous in terms of data standards, temporal/spatial resolution, etc. This heterogeneity works to impede data integration by individual researchers, and consequently the discovery of new knowledge, which could come from merging the independent data sources.

Currently, research efforts such as the Semantic Web focus on the development of automated data synthesis technologies. Despite the possibly overwhelming data deluge in some disciplines, modern research practices are evolving with the adoption of enabling e-Research/e-Science technologies to improve previous research methodologies and/or research processes. Such research efforts include adopting developments made by the Semantic Web and Knowledge Representation (KR) communities to resolve "data deluge" or big-data problems [5, 6]. Implementations of semantic technologies within e-Research are prevalent in the Medical and Life Science disciplines, but are still emerging in the Earth and Ecological Sciences [5].

Coral reef ecosystems worldwide are facing many pressures from both natural and human-induced stresses. Because of the scale and complexity of studies into coral reefs research, there are propositions that research will be enhanced if data from unrelated studies could be merged to build a more complete understanding of the systems in question. Improved analysis, better reuse of data and access to data repositories would all benefit from a more integrated approach. Many of the most influential papers in coral reef science of the past few years have been "synthesis" papers aggregating long-term observations into new hypotheses and conclusions, for example the Inter-governmental Panel on Climate Change reports (IPCC) [7].

This paper explores the use of semantic technologies to provide a basis for the efficient reuse of data through data integration methods and flexibility in the hypothesis design that may assist in the future research of climate change effects on our coral reef ecosystems. The Semantic Reef project is a platform that consists of a semantic Knowledge Base (KB) and scientific workflows so researchers can combine and question scientific data in an integrated hypothesis-based research tool. Semantic Web technologies are inherently machine-centric. They focus on software, data and application layers that connect disconnected data and integrate it in ways that can be manipulated by a computer. In contrast, workflow technologies are both human and machine centric. They enable human actors to make the connections between the different technologies, software and hardware from diverse domains. Developed as a KR platform, the Semantic Reef allows researchers to combine disjoint data into a single KB and flexibly pose observational hypotheses of the data and/or provide alerting for unusual events (e.g., coral spawning or bleaching).

The Semantic Reef architecture (Figure 1) offers an alternative approach to the development, application and execution of observational hypotheses in the coral reef ecosystem domain [8]. The scientific workflows retrieve remote sensor data and data available via the Web and integrate the data into the existing KB for further synthesis and analysis. Throughout this process, the automated workflow performs any necessary calculations, reformats the data and then routes it into the KB. The semantic KB consists of a hierarchy of ontologies to describe a coral reef ecosystem that can enable ontology-based data integration. The data can be reasoned over and inferences can be made once the ontologies have been populated by the workflow. For example, a domain expert, either a marine scientist of reef manager, can query the KB to extract information of interest, pose observational hypotheses, or construct an alerting system by inferring events. The Semantic Reef model is a research case study that combines semantic technologies, scientific workflows, first order logic (FOL) and propositional logic systems. The architecture has been tested and demonstrated to handle disparate data for propositional suppositions to extract or disclose phenomena from the data. The approach is extendable to different hypotheses over many environmental monitoring and climate change concerns.

To design the Semantic Reef, knowledge in semantic systems as well as an understanding of ecological science was required. This requirement highlights the importance of a collaborative approach where different expertise and skills need to be combined. Clearly, better systems and example implementations will be required before the widespread adoption of semantic tools, of the type described here, become commonplace. Even so, collaboration, end-user involvement, and potentially even community engagement, all work to increase uptake and adoption of what would otherwise seem inaccessible technologies.

A background and review of the technologies and similar initiatives is presented in section 2. Section 3 describes the development and design methodologies employed to create the hierarchy of ontologies that makeup the KB. Section 4 illustrates the potential the system offers in flexible hypothesis-driven research environments. Section 5 describes the validation of the system and demonstrates advantages offered by applying the specific technologies. Section 6 reports the performance analysis of the architecture. Section 7 concludes with a brief summary and discussion.

2. BACKGROUND

Some of the recent significant advances in science have been achieved through sharing complex interdisciplinary skills, data and analysis [1]. In fact, the connections between disconnected ideas, domains, people and data contributes significantly to the creation of new knowledge and its reuse [5].–A central requirement of the new generation of research software is the capability to automatically search, access, move, manipulate, and mine data stored in vast distributed digital repositories and/or discreet data silos [5]. To accomplish the rearrangement and juxtaposition of inter-disciplinary data in interesting, efficient and exploratory ways requires the application of automated data integration methods and technologies, such as semantic and scientific workflow technologies.

2.1. Semantic Web technologies

The Semantic Web links data so it can be accessed, reused and/or manipulated more readily by the machine [9]. The concept creates links between data, rather than simply inputting data on the Web. The technologies make available contextual information about the data, and thus make the data understandable to the computer and therefore automatically processable by the computer [9]. This form of machine processing enables the automation of tasks such as data fusion and data integration.

Data integration is a primary motivation for the development of Semantic Web technologies and is at the heart of the Semantic Reef project. Ontologies lay the foundation of Semantic Web technologies to support automated processing of information. An ontology gives context and meaning to the data available to the computer by describing “things” that exist within a domain, whether they are abstract or specific [9, 10]. A concept is modelled and its interpretation constrained by specifying the domain's vocabulary and the terms, axioms and restrictions to describe the entities and the relationships that exist between entities [11].

Ontologies can bridge disparate data held in data silos or available via Web [9]. Competing approaches to amalgamate heterogeneous data sources include data warehousing and data mining. However, the application of ontologies for these integration tasks is potentially more flexible as they can resolve the semantic conflicts in definitions that invariably arise from the application of diverse schematic sources [12]. Herein, a set of reusable ontologies have been developed to describe to a computer the concept of, and the relationships within, a coral reef ecosystem.

The Semantic Web technologies are currently being implemented and/or developed by many diverse efforts. Many of these efforts focus predominantly on the knowledge found in the documentation on web pages, while others focus on data produced by a variety of instruments. The Linking Open Data project [13] and the Creative Commons (CC) [14] organisation are examples of

document-centric Semantic Web activities. In contrast, the Semantic Sensor Web (SSW) project [15], which aims to provide an environment for improved query and reasoning in a sensor domain, is an example of a data-centric Semantic Web initiative. Whether the undertaking focuses on document-centric or device-centric data, the development of the Semantic Web is driven to improve communication and bridge web-accessible disparate data. The Semantic Reef project aims to ingest both forms of semantically available data through a different approach to hypothesis design to automate the data analysis process.

2.2. Scientific work flows

Software systems such as Kepler [16], Taverna [17], Triana [18] are tools that allow scientists to capture scientific workflows. The software chosen for the data flow implementation of the Semantic Reef architecture is Kepler, which is an open-source scientific workflow tool [16, 19]. The choice of Kepler as the workflow system was motivated predominantly due to the flexibility in workflow design and manipulation. As shown in a taxonomic study of workflow systems by Yu [19], Kepler is a user directed system that supports flexible data movement methods. These methods include: A centralised approach where data is transferred between resources via a central point; a mediated approach where the locations of the data are managed by a distributed data management system; and a peer-to-peer approach where data is transferred between processing resources [19]. The flexible data movement supported by Kepler workflows enables access to a diverse range of data resources, such as the distributed data repositories and streaming sensor data required to populate the ontologies within the KB.

Prominent examples in the implementation of scientific workflows in eco-informatics are the Science Environment for Ecological Knowledge (SEEK) project [20] and the myExperiment project [21]. Previously, the large and disparate ecological and biodiversity data has been impossible to coordinate into one workflow. However, SEEK can streamline data acquisition and archive tasks through data integration, transformation, analysis, and synthesisation [20]. myExperiment is a virtual research environment for the social curation and sharing of scientific objects, such as research investigative designs, questions, results, publications, and in particular, scientific workflows and *in silico* experiments [22].

Workflows are employed here to process data automatically and pass the results to the Semantic Reef KB. Specifically, the workflows initiate Web services to collect both near real-time data from remote sensors and existing web available data from archives and repositories. The KB could be filled with relevant data available from diverse sources, such as remotely sensed data, satellite data or web-accessible statistical data (Figure 1). Hypothesis questions or alerts can then be posed of the data by questioning semantic correlation and analysis with Description Logics (DL) and inference rules.

2.3. Related work

Two highly relevant initiatives that apply semantic technologies and/or scientific workflows to manage data are the Science Environment for Ecological Knowledge (SEEK) project and the Semantic Sensor Web (SSW).

As discussed above, the SEEK project is a National Science Foundation (NSF) funded eco-informatics initiative. The system is designed to support data acquisition and management of ecological and biodiversity data. The project encompasses many cyber-infrastructure tools that are necessary to integrate complex ecological data and enable rapid development and reuse of complex scientific analyses [20].

SEEK encompasses three integrated systems: a Grid computing infrastructure for data storage, sharing and access; a semantic mediation system that reasons over data to determine whether it is relevant to a designated workflow; and a modelling system, for use by ecologists to

design, modify and incorporate analyses when composing new workflows. The primary goal of the SEEK project is the production of an efficient tool for ecologists to capture, organise and search for data, and apply analytical processes from their desktops.

The Semantic Reef project can benefit from the resources made available through the SEEK facilities such as the data sources and the semantic mediation system. For example, a hypothetical proposition that is run in the Semantic Reef system can adopt the ecological data, which are available via the SEEK EarthGrid portal[1], as resources. The ontology-based services, provided by the semantic mediation layer, support the Kepler workflow system in data discovery and integration and offer a knowledge-based query system for the integration of disparate data resources.

Also available through SEEK, for use by systems such as the Semantic Reef, are a range of top-level formal and informal ecological ontologies. These external ontologies can be mapped to the KB because ontology design supports interoperability, scalability and reuse and enables mapping capabilities for both internal and external ontologies. Once imported, the ontologies can be modified or added to, depending on the purpose of the system. The ontologies cover unit and measurement systems and temporal/spatial concepts, among others, and can be imported to the Semantic Reef KB and adapted to suit a purpose (e.g., domain specific terms, parameters, etc.).

As discussed in Section 2.1, the SSW project aims to provide an environment for enhanced query and reasoning within a sensor network and effectively connect sensors to the web. The SSW annotates sensor data with spatial, temporal, and thematic semantic metadata to increase interoperability of that data and provide enhanced descriptions and information essential for data discovery and analysis [23]. This proposed technique builds on current standardisation efforts within the W3C and Open Geospatial Consortium (OGC) by extending them with Semantic Web technologies [15].

The SSW, for example, can apply complex queries about weather data collected from the urban Geographic Information System (GIS) systems and weather services. A prime motivation of the SSW is to merge the data gathering instruments (e.g., remote sensors, video and other cameras devices, etc.) with the collection and analysis process. This merger is important because there is otherwise a lack of integration and communication between multi-layer sensor nodes, such as high-level and low-level sensor networks. The information, once integrated to the SSW, is valuable to query or inference applications suitable for end users (e.g., traffic control, weather alerts, etc.). The data can be queried, reasoned over and/or have inference rules applied, including rules to automate alerts[15].

There are three main differentiations between these projects and the Semantic Reef project: the level of applied semantics, the data scope and support for workflows:

Firstly, the level of applied semantics: The SEEK initiative incorporates the semantic mediation layer as a component to its architecture. The mediation layer reasons over data to determine data relevancy and analytical components for automatic transformation and use in a selected workflow. SEEK also applies the higher levels, such as description logics, within some of the environmental ontologies. The ontologies are maintained as a repository for public access and use (e.g., the food web or biodiversity ontologies) and are part of the infrastructure offered by the SEEK program. Notably, rules based inference is not supported as a component of the SEEK implementation.

The Semantic Reef and the SSW projects incorporate all the logic and inference systems available in the Semantic Web stack. The agenda of the Semantic Reef project was to explore the possible benefits these technologies offer to hypothesis-driven research in the marine science domain. In contrast, the SSW focuses predominantly on the annotation and quality control of sensor data. The SSW aims to explore higher semantic functionality within the sensor technology standards and proposes new additions to the current sensor standards. This proposal includes the addition of semantic annotation to the sensor layers as metadata of sensor data for access to sensor

[1] http://ecogrid.ecoinformatics.org/ecogrid/

data streams. Accordingly, when data managed by the SSW is relevant to marine research, the data in the storage level of the SSW architecture will be a valuable source of quality assured sensed data for import to the Semantic Reef system.

Secondly, data scope: The limitations, flexibility and scalability of information outcomes depend on the source of data. These dependencies include whether the data must be from a quality assured source or completely open source; whether the project can only use data from a preset number of sources (data silos, distributed data, etc.); and/or whether the data has temporal limitations (historical data versus real-time streamed data). The Semantic Reef architecture is a scalable general-purpose ontology-based model that permits any digitalised data from any openly available source. SEEK is also designed to incorporate open data. SEEK is a support infrastructure with a holistic view of the eco-informatics domain and permits scientists to structure their own experiments. In contrast, the Semantic Reef model is an atomistic application that focuses predominantly in the subset of coral reef ecosystems.

Thirdly, workflows: Related fields such as model-driven architectures and semantic modelling are developing possible solutions to streamline data analysis. However, scientific workflows have not been widely applied to these techniques [24]. Tools are required that let domain scientists effectively harness the functionality of an e-Research infrastructure without the need to become computer scientists themselves. Similar to the Semantic Reef the SEEK project employs scientific workflows (the Kepler framework is part of SEEK). The SSW does not employ independent scientific workflow tools.

3. OVERVIEW OF THE KNOWLEDGE BASE

The Semantic Reef KB consists of a hierarchy of ontologies written in the Web Ontology Language (OWL) [25] that describe coral reef ecosystems. Essential to the development of the Semantic Reef KB was the combination of the perspectives of a discipline specialist (Coral Reef Ecology) with human/computer translation functionality [26]. Holmes [27], the coral reef domain expert here, developed a functional model of a generic coral reef (Figure 2) that describes the basic functionality of a coral reef system and includes components such as coral reef community composition, nutrient dynamics and environmental and anthropogenic influences. Figure 2 shows at a broad level the processes functioning in any coral reef independent of the reef type, where it is situated globally and what is contained in the community "mix" (i.e., the categorization of species) [27].

Each principal component of the Holmes's model is a concept that can be defined independently as a composite of its sub-nodes. The model is a hierarchy of concepts based on a holistic view of any coral reef that begin with the main functional nodes, hydrodynamics, human influence, light environment, etc. In turn, each node contains a hierarchical composite of features and conceptual terms at an atomic level.

The hierarchical components of the expert's model were used as semantic "building blocks" for translation into a modular ontological form in the design of the KB. Myers [26] describes the modelling of the reef system in a hierarchy of ontologies to support flexibility and reuse of the KB that aligns with the holistic and atomic levels of Holmes's model.

3.1. Reusable and usable ontologies to describe coral reefs

A hybrid ontology-design methodology was adopted in the creation of the KB to maximise reusability and usability [8, 26]. To support reuse, a modular design was required as opposed to a single monolithic coral reef ontology. Ontology design methodologies support modularity for more effective conceptual modelling [28]. To describe a whole reef system modelled in a single ontology would be too complex, bespoke and not easily reusable because of the intricate relationships of each ecosystem. The ontologies were designed as "reusable" domain ontologies, to

describe coral reefs generically, and as "usable" domain-specific and application ontologies, to describe individual coral reefs and the rules of the hypotheses about that reef.

The relationship definitions between concepts in the ontologies need to be flexible for future modifications in hypotheses or the introduction of new domain-specific information. Although well researched, many (possibly most!) linkages and connections between the functions of a coral reef within the ecosystem as a whole, remain unstudied, or poorly understood, and new modifications may lead to logical defeasibility [29]. The design methodologies used in this hybrid model include the *seven step knowledge engineering methodology* [30], Uschold and King's [31] three strategies to indentify concepts and the *Developing Ontology-Grounded Methods and Applications* (DOGMA) approach [32]. The first and second are a generic set of guidelines to construct ontologies of any concept and were employed for the "usable" domain specific and application ontological development. The third offers a strategy that effectively separates the domain knowledge from the application or domain tasks to focus on ontology reusability verses ontology usability [26].

The base-level light-weight reusable ontologies are imported to the more complex DL ontologies and so forth to form a "ground-up" physical hierarchy within the KB (Figure 3). The lightweight ontologies, written in OWL Lite for simplicity, can be populated using a Kepler workflow with historic or real-time data. The more complex concepts of Holmes's model, trophic layers and human influence components (Figure 3), were created with the more expressive OWL DL. DL constructs offer functions for reasoning, such as existential and universal quantification, cardinality and Boolean combinations, to impose explicit restrictions on properties and infer connections in a DL KB [33]. The task-specific and application ontologies, at the higher levels, employ the reusable coral reef ontology base beneath. The DOGMA approach distinguishes ontologies for use and reuse by effectively separating the classes and properties that are contained in the domain ontologies from the instance data (domain-specific ontologies) and rules (application ontologies).

The application ontologies are the highest level of the KB where the hypothesis-specific data and inference rules are introduced. Detailed inference rules can be written to the system as propositions to infer conclusions about a specific problem on a particular reef, regardless of location. The application ontologies import a populated domain-specific ontology (e.g., Reef 1, Reef 2, etc.), which contains the instance data and the lower "reusable" domain ontologies. Then, propositional testing is implemented through Semantic Web Rules Language (SWRL) inference rules to perform tasks such as posing a hypothesis of the KB or to query the KB [34]. The SWRL rules are written to represent the hypotheses posed by a marine researcher as many hypotheses have a syllogistic format that can be fashioned in a Horn clause form.

The capacity for reuse relies on the ability to introduce new or different data into the system dependent on the hypothesis and the reef in question. Each new hypothesis may be a different line of enquiry and require different data from a preceding one. The KB can be refilled, depending on the line of enquiry, by separating the "reusable" domain KR from the "usable" instance data (i.e., the DOGMA technique). Specifically, the domain (reusable) ontology (i.e., the "Coral Reef" ontology) is separated from the higher application (usable) domain-task ontologies (i.e., a specific reef and hypotheses) (Figure 3). The reusable components of the ontology base are contained within the "Coral Reef" generic ontology, which imports all lower ontologies to describe any coral reef. The usable component of the KB lies in the domain-task ontologies. After the "Coral Reef" ontology is imported, the domain-task ontology is populated with instance data pertinent to the specific reef system and hypothesis (e.g., Reef 1, Reef 2, etc.).

An example use-case may be a researcher who is searching for the cause of coral bleaching at Reef 1 and would import the "Coral Reef" ontology to the "Reef 1" domain-specific application ontology (Figure 3). The domain specific ontology is populated with data relevant to Reef 1. Alternatively, research conducted on Reef 2 would employ the same generic "Coral Reef" ontology for the research hypothesis but populate the KB with data pertinent to Reef 2. The elements and

classes of the generic ontologies will be the same for either reef locations, but the instance data and rules that represent the hypothesis will differ.

4. THE SEMANTIC APPLICATION - BENEFITS AND DISTINCTIONS

The following sections illustrate the capabilities and potential of the Semantic Reef system. The specific benefits discussed include flexible hypothesis design, data integration capabilities, automation, modularity and reuse in the application of semantic technologies to hypothesis-driven research.

4.1. Versatile hypothesis design

A researcher using the Semantic Reef system is not required to predetermine precise hypothesis prior to data collection and the population of the KB. Rather, the questions can be as flexible, and may evolve as new data becomes available and/or as ideas emerge [35]. For example, a researcher may initially propose a bleaching event with two factors: Sea Surface Temperature (SST) and salinity. Then, decide to also include some unorthodox or seemingly unconnected factor to the hypothesis such as sales of a brand of fertiliser, documented catches of a fish species or scheduled dredging for that region. If information is available, it can be imported to the KB and added as a factor in any hypothesis (Figure 4).

The hypothesis statements within the system are axioms that give the information required for the system to logically infer results. The axioms may or may not be true of the real world, but would be the monotonic suppositions in a specific hypothesis that would be stipulated in the research methodologies and assumptions. Because they can be modified based on the researcher's assumptions or models, if a proposition requires conjecture on the part of a researcher the axioms in the KB can be arbitrarily changed to depict the proposed environment.

4.2. Data integration and the open world assumption

The unstructured nature of the Open World Assumption (OWA) allows the KB to have a flexibility which can easily modify or adapt to new or additional concepts [9]. Data from research institutions, governments, non-profit organisations and commercial companies is commonly stored in unconnected data repositories and ontology-based data integration can be employed to bridge these data silos [12]. Additional information and unstructured data is expected under the OWA because the system assumes it never has a complete view of its world and there are always unknown facts to be added [36]. Hence, the OWA allows new information to be easily added to the Semantic Reef KB based on changes in the researcher's line of query or as new data or information evolves.

4.3. Query and inference

Semantic technologies offer both query functionality and extensive inference capabilities. Query capability is possible in semantic-based systems at either, or both, the RDF or OWL levels [37]. Currently, these levels require different query paradigms, SPARQL is the query language used to query RDF triplestores and SQWRL is used to query at the OWL DL level [38, 39]. Both semantic query levels can be applied in the Semantic Reef system.

4.4. Semantic modularity

The hierarchical modular design of the Semantic Reef KB is an example of component architecture that makes repopulation and reuse of the KB possible. The ontology hierarchy is independent of any particular coral reef and its environment or human influential factors.

Adaptable lines of enquiry are possible through modularity (Figure 4). To illustrate, coral bleaching was the theme in the validation process, with a specific focus on a small sample of coral reefs within the GBR. To infer a bleach-alert for coral reefs in different locations would simply require the repopulation of the KB (if the data were available). The inference rules would remain generally the same. Alternatively, if the line of enquiry were different, modifications would only be required at the higher usable layers of the KB. If the theme was not coral bleaching, but instead regeneration rates or coral spawning, for example, the underlying reusable ontology modules would remain the same but the instance data and hypothesis rules would differ. A domain-specific ontology would be created for the new line of enquiry and import the lower ontologies, repopulate the KB with relevant domain-specific data, and then new proposals as inference rules can be posed.

5. HYPOTHESIS VALIDATION AND DEMONSTRATIONS

This section presents the validation of the system and exemplars to illustrate data integration, flexibility in hypothesis design, and the benefits of automated classification. The examples are intentionally simple. The purpose is to demonstrate the potential advantages and applications of the Semantic Reef architecture.

5.1. Validation

The KB was substantiated via a reverse-hypothesis or ground-truth approach [8, 26]. The methodology involved a comparison between historic events with the ensuing observational research and the inferred outcome from the KB. The mass coral bleaching episodes that occurred on the Great Barrier Reef (GBR) in 1998 and 2002 were the validation subject. The KB was populated with the historic SST data and the outcomes of the system's rules were evaluated against the historic data analyses and *in situ* field observations of the mass bleaching events.

Corals live in a symbiotic relationship with single-celled algae called zooxanthellae that live within the coral's tissue to provide an essential food source [29, 40]. Coral bleaching results from a breakdown of this symbiotic relationship caused by a stress conditions such as higher-than-normal sea temperatures. Elevated temperatures of 1^0 C above the long term monthly summer averages are sufficient to cause the stress factors that result in coral bleaching in many coral species [29]. The algae give corals their characteristic colours and when they are expelled due to stress from high temperatures, what remains is the white skeleton. If stressful conditions continue, the corals bleach and die [40].

Two major coral bleaching events occurred in the GBR during February and March of 1998 and 2002. Mild bleaching began in late January of each summer and intensified throughout February after hotter than normal temperatures.[41]. The bleaching severity during each event was assessed by underwater video survey at fourteen sites on the central GBR by the Australian Institute of Marine Science (AIMS) [41]. Four of the initial fourteen sites were re-surveyed in 2002 to evaluate changes, namely, Kelso Reef, John Brewer Reef, Faraday Reef (via the Myrmidon Reef monitoring station) and Florence Bay at Magnetic Island (Figure 5). The reefs showed significant levels of bleaching in both the 1998 and 2002 bleaching events and have been used to estimate the relationship between accumulated thermal stress and bleaching severity [41, 42].

The historical data for the validation was supplied by AIMS's large-scale temperature monitoring program [43]. The raw data used in the surveys consisted of SST taken for the summer periods 1995 through to 2003 and contained minimum, maximum and mean daily SST. Maynard's study [42] of the thermal tolerance of major coral genera was the benchmark for the validation of the Semantic Reef KB.

The temperature logger data was stored in tabular Comma-Separated Value (CSV) format and ported to the KB via a Kepler workflow. The workflow was created to physically manipulate the data in preparation for the KB, to test the system's ability to provide a coral bleaching "alert".

The workflow imports the data in XML format and, through the XPATH actors in Kepler, extracts each date with its corresponding mean, minimum and maximum temperature data value. All values from the workflow are passed to the KB at the domain-specific ontology level to populate the environmental, temporal and the community mix properties of each reef.

Coral bleaching risk is estimated by calculating thermal stress indices that measure bleaching severity based on temperature characteristics. Four indices of temperature elevation are in common and include [44, 45]:

- The magnitude of SST anomaly (SST+), which calculates the temperature anomaly as the number of °C above the Long-term Mean Summer Temperature (LMST) observed for that month ;
- The maximum summer temperature (MaxSST), in contrast, is based on the Local Mean Summer Maximum (LMSM) temperature ;
- The "HotSpot" anomaly is also an anomaly metric but differs from the previous two because it is not based on the average of all SSTs. Instead it is based on the climatological mean SST of the hottest month for the region, referred to as the Maximum Monthly Mean (MMM); and
- The Degree Heating Days (DHD), which describes the accumulation of thermal stress as opposed to the SST anomaly metrics.

The anomaly indices (SST+, $_{Max}$SST and HotSpot metrics) and the accumulation index (DHD) were the focus in the validation of the KB where logical inference rules, DL and queries were used to mimic the metrics [8, 26].

Sets of axioms and inference rules were defined in the domain-specific and application ontologies to characterise the concept of coral bleaching. Relevant characteristics such as rising SST or whether coral is ahermatypic or hermatypic, meaning it depends upon zooxanthellae for nutrients and is thus susceptible to bleaching [29, 43]. SWRL rules were created to mimic the anomaly metrics (SST+, $_{Max}$SST and HotSpots) to infer a bleaching event and the SWRL query language (SQWRL) was used to query the KB to mimic the DHD accumulation metrics [34, 39]. Only instances in the KB that matched the inferred rules were classified to designated "bleach-watch" classes. The inferred outcome of the validation exercise correlated with the observed bleaching occurrences for both periods when signs of bleaching began to show and the subsequent mortality rose [42].

5.2. Thermal indices with live data flows

The ability to automate coral bleaching alerts is extended here to portray the real-time prediction potential of the system. The near real-time SST data in this exercise was streamed from the Cleveland Bay (an inner shelf reef), Davies Reef (a mid shelf reef) and Myrmidon Reef (an outer shelf reef) monitoring sites (Figure 5) [46]. The data was made available for the three reefs from the weather observing system via the AIMS data access portal.

The temperature logger data was imported to the KB via a Kepler workflow. A workflow was created to manipulate the data in preparation for the KB and test the system's ability to provide a real-time coral bleaching "alert". The workflow imports the data in an XML format and, via the XPATH actors[2] in Kepler, extracts each date with its corresponding mean, minimum and maximum temperature data value. The workflow also extracts the LMST, LMSM and the MMM, which are components of the bleaching metrics. The validation inference rules were applied to detect a problematic area and only instances of a particular reef that proved true were automatically inferred to belong to a categorised bleach-watch class.

[2] Each workflow step is represented by "actors," which are individual processing components that can be manipulated through a "drag and drop" method into a workflow, via Kepler's visual interface.

A bleach watch warning for Davies Reef was the result of the inference rules. The NOAA Coral Reef Watch's Satellite Bleaching Alert system issued a bleaching watch alert for Davies Reef, which coincided with the inferred bleach risk instances from the Semantic Reef system [8]. Cleveland Bay was not inferred to the urgent bleach watch class even though it experienced moderately high temperatures it remained below the LMST threshold for the summer.

5.3. Applying disparate data to theorise the coral bleaching tipping-point

To demonstrate ontology-based data integration disparate data were mapped to the KB for inclusion in a sample hypothesis. The independent data sources included AIMS, NOAA's Integrated Coral Observing Network/Coral Reef Early Warning System (NOAA ICON/CREWS) [2], the Australian Bureau of Meteorology (BoM) and the Australian Bureau of Statistics (ABS).

Dangers to Australia's coral reefs fall into three categories:

- Natural stresses of which corals have evolved to cope with;
- Direct anthropogenic pressures that include sediment and nutrient pollution from land run-offs, fishing practices that overexploit fish populations; engineering and modification of shorelines; and
- Global climate change and ocean acidification.

Many of these threats are closely linked and exacerbate each other as shown in studies of global climate change in coral reefs, such as increased coral bleaching and coral disease [43].

Coral bleaching is not uniform, but instead occurs in discrete regions within a reef system. Active research is attempting to find the "tipping point" that leads to coral death from bleaching because at present there is still only a limited understanding of the physical and biological causal factors [29]. Studies entail the cumulative combination of ecological factors and stressors that contribute to the tipping point. The Semantic Reef system is a tool to help pose hypotheses and can be employed to theorise about the cumulative factors of bleaching. Once phenomena in the data are disclosed, *in situ* observations can be performed to confirm or negate the theory.

A range of environmental and anthropogenic information was mapped to the KB. The three reefs from the previous example (i.e., Cleveland Bay, Davies Reef and Myrmidon Reef) were used plus an additional second inner reef system: Bowling Green Bay (Figure 5). The choice of reef systems depended on the public availability of the data. The environmental factors incorporated in this test consist of average and maximum SST, Photosynthetically Active Radiation (PAR), Chlorophyll concentration and rainfall. Where appropriate, the gaps in the data were supplemented with representative data from a proxy location, which is common practice in marine research. PAR is a common proxy for Chlorophyll-a, which measures the abundance of coral food sources. Rainfall is commonly used as a proxy for the water salinity.

SST and PAR were extracted from the AIMS data centre, while rainfall and PAR data were supplied by the NOAA ICON/CREWS site. The anthropogenic factors mapped to the KB consisted of human population density and quantity for the coastal transect from Townsville to the lower Burdekin (Figure 5). The population data extracted from the ABS online database included the geographic figures and demographic breakdowns (age and gender) of both regions. Questions that theorise about the effects on coral reefs as a result of the human coastal population density may now be posed with the additional anthropogenic information.

The two locations, Townsville Ross and Burdekin Rivers (Figure 5), were appropriate for this exercise because they both have river outlets. The quality characteristics of these local rivers differ. The Townsville Ross River opens into Cleveland Bay and is influenced by the dense population and industries of that region. In contrast, the Burdekin River sustains a low rural population density and agriculture and opens onto Cape Bowling Green. Hypotheses can now be explored that examine water quality and coral health about the two inner shelf reefs, Cleveland Bay and Bowling Green Bay.

A Kepler workflow imports and transforms the disparate data and prepares the KB by populating the ontologies (Figure 6). The data from the four disparate data sources is manipulated via the Kepler XPATH and Python actors. The XPATH expressions and queries extract the specific data values from the data streams and then converted to an array of values to be sent to the Python scripting actor. The Python actors implemented simple scripts written to tag each value with a unique URI and sent to the KB to populate the appropriate ontology modules. On completion, the KB was populated with 360 instances, 90 for each of the four reefs. One temporal instance was created for each day per reef in the summer period. Property assertions to describe the environmental information and reef community composition were included for each instance and linked to the population quantity and density human influences for that location.

The rules were fashioned as observational hypotheses. Therefore, if any phenomena in the data were uncovered the location could be observed for *in situ* confirmation of the hypothesis. If there were a change in the hypothesis due to new information or an epiphany, the rules could be modified to express the new hypothesis simply by adding or removing antecedents to the rules. An inference rule to question the variations in human populace in correlation with other prescribed factors as exemplified in the following SWRL rule:

```
Coral_Reef:Coral_Reef(?x)  ∧
Coral_Reef:has_Human_Influence(?x, ?y)  ∧
Human_Influence:Influence(?y)  ∧
Human_Influence:hasPopulationDensity(?y, ?pop)  ∧
swrlb:greaterThan(?pop, 5000)  ∧
Coral_Reef:hasLightEinsteinsOf(?x, ?par)  ∧
swrlb:greaterThanOrEqual(?par, 500)∧ swrlb:lessThan(?par,750)
∧  Coral_Reef:hasDailyAverageSSTof(?x, ?meanTemp)  ∧
Coral_Reef:hasAverageLongTermSeaSurfaceTemperatureOf(?x,?LMST)
∧  swrlb:greaterThanOrEqual(?meanTemp, ?LMST)  ∧
Reef_Stock:Coral(?partCoral)  ∧
Coral_Reef:hasPart(?x, ?partCoral)  ∧
Trophic:hasGrowth(?partCoral, Trophic:fast )
Trophic:is_Hermatypic(?partCoral, true)
-> Coral_Reef:Observe_Reef(?x)
```

The inferred instances of this rule were Townsville regional locations. The temporal reef instances that fit this combination of influence and environmental values could be observed *in situ* for signs of bleaching. The results from the rule exposed instances from the 1st to the 16th and the 21st to the 28th of December at Cleveland Bay (Townsville), which aligned to actual summer bleaching signs.

5.4. Reasoning and classifying community makeup and location

Data gathered from monitoring stations are commonly used as representative data for surrounding reefs in current marine research methods. To have a sensor deployed at every reef is unfeasible due to the cost and/or the ecological interruptions involved. Data from one sensed location can represent other surrounding reefs, and in special cases as surrogate data for reefs that are similar by their type and not just their location.

Models that describe or represent a reef by type include community make-up, thermal sensitivity, and/or by nutrient concentrations [42]. The important concept here is that the method of characteristic models rather than simple proximity implies that data from one reef may be indicative or representative of others. Accordingly, automatic reasoning and classification can be applied to show how these “reef-type” models may be integrated into the Semantic Reef system. Logical

axioms are explicitly expressed to describe the reef types by the thermal sensitivity of the community composition and by location.

5.4.1. Classifying reef-type by the community mix

The classification of reef-type by its community composition and sensitivity to heat stress in our model can be described in the following statements [43]:

- Type A reef - is a reef with a high percentage of slow growing coral and is thermally tolerant; whereas,
- Type B reef - is a reef with a high percentage of fast growing coral and is thermally sensitive.

A query to select all sensitive reefs would not return any results unless specific reefs were manually asserted to belong to a Type A or B reef class. The thermal tolerance assumptions can be defined in an ontology as "necessary and sufficient" axioms of a reef class type. The property restriction axioms were added to a selection of reefs to illustrate the automated inference capability of the system. The selection included different reef types: fringing and barrier reefs, and different locations: inner, mid and outer shelf areas.

The reasoner classification of the ontologies resulted in the reef classes being correctly inferred to the various reef-type classes. All reefs that have a greater percentage of fast growing coral were subsumed to automatically belong to the "Fast Growth Reef" class. The reefs that had been appointed with a higher mix of slow growing corals were subsumed to belong to the "Slow Growth Reef" class (Figure 7). Then, once all reef classes were subsumed to belong to the functional types, inference rules were posed based on the data in correlation to the reef's characteristics as well as environmental factors.

5.4.2. Classifying reef-type by location

There is a consensus that the location of a coral reef in proximity to other reefs has environmental commonalties [43]. This methodology, which is predominately related to environmental factors, assumes that the factors of a specific geospatial transect will be similar and indicative of each reef in that transect. A similar technique to the previous example was applied but instead of the community mix, the reef classes were inferred to belong to a reef-type class based on its specific geospatial grid location [8].

The GBR is divided into gridded areas for both management and research purposes. This GBR grid is based on longitude and latitude values and divided into a matrix from North to South and inner-shore to outer-shelf. To define each gridded area of the marine park, geospatial property values were declared for each coral reef class. The longitude and latitude property restriction axioms test the asserted values to subsume a coral reef to also belong to the grid regions. On reasoning over the KB the reefs of the GBR are subsumed to the different reef-types and questions can be posed of the system.

6. PERFORMANCE ANALYSIS – METHODOLOGY AND RESULTS

To test the Semantic Reef system as a desktop tool for use in hypothesis-based research, instances of the reasoning and inference functionality were implemented. The performance analyses consisted of a series of exercises administered in a simulated computing environment indicative of a researcher's *in silico* environment. The test scenarios focused on the quantity of data (i.e., triples) introduced to the system versus the time to load the KB and then reason and infer over the KB. The evaluation methodology incorporated strategies implemented in other similar knowledge-based tests [47, 48].

Marine researchers typically use standard desktop machines to run *in silico* analyses that must run in reasonable periods of time. Long execution times lead to the disengagement of researchers.

6.1. The Knowledge Base software

The Knowledge Base Software consisted of Protégé[3], which is an open source ontology editor and KB framework. Protégé offers wide developer and user support through an active development community and importantly has a range of direct and indirect support for a variety of reasoning engines, such as Pellet[4], FaCT++[5], and RacerPRO[6].

At the time of writing, there are two means to initialise the reasoning engine through Protégé; indirectly, through the DIG interface [49] and directly through an inline memory connection. The DIG interface provides a communication connection to any DIG compliant reasoner (e.g., Pellet, FaCT++, RacerPRO, etc.) but the primary disadvantage of the indirect access is the lack of support DIG 1.1 has for data-type properties. Protégé 3.4 has reasoning support via the DIG interface and, in this case, RacerPRO was the reasoner chosen for the trials. Alternately, through the direct in-memory connection of the KB framework, the FaCT++ reasoner is available in Protégé 4 and the Pellet reasoning engine is available to both Protégé 3.4 and Protégé 4.

Propositional logic is the basis of observational hypotheses, so SWRL functionality was a crucial requirement of the system. Protégé 3.4 was chosen as the main infrastructure for the Semantic Reef architecture due to the extensive support for SWRL inference and SQRWL queries.

6.2. The performance analysis methodology

The performance analysis centred on the scenarios from the previous section and compared the quantity of data versus the processing time. The tests were run over a three-month period of data (a single summer) with datasets of daily and hourly observations for four reefs. The tests were varied using a matrix of attributes (Table 1), which could be changed by factor and the outcome then compared for each run of the system. The limitations and performance of the desktop computing environment was tested by processing reasoning and inference functions over a growing range of triples.

The metrics used in the experimental performance runs were as follows:

- Time to load all triples to the KB and the time to load the complete KB;
- Time to reason over the KB with Pellet, FaCT++ and RacerPRO. This test was completed for two ontology levels, both at the "usable" level of the hierarchy: the domain-specific ontology ("GBR.owl") and the application ontology ("GBR_Rules.owl"). The Pellet and RacerPRO (via Dig 1.1) reasoning engines were run at both ontology levels using Protégé 3.4, and the Pellet and FaCT++ reasoning engines were run in Protégé 4 for the domain-specific ontology level (i.e., GBR.owl); and
- Performance of the inference rules was tested at the application rules ontology (GBR_Rules.owl) with the Jess Inference engine via the SWRL Tab in Protégé 3.4 and the results are presented in Myers [8].

The scenario variables that can be changed for each test consisted of the data-type and object properties asserted to each instance. The data-type properties available for assertion were: SST (average, maximum and minimum), date, time, LMST, LMSM, MMM, longitude and latitude,

[3] The Protégé project currently has two framework versions available, Protégé 3.4 and Protégé 4, which are being developed concurrently (http://protege.stanford.edu/). An important factor in selecting the KB framework was the reasoning support and both versions offer adequate availability to this functionality.

[4] Pellet: the open source OWL DL reasoner. http://clarkparsia.com/pellet

[5] FACT++: Fast Classification of Terminologies Description Logic classifier. http://owl.man.ac.uk/factplusplus/

[6] RacerPRO: Renamed ABox and concept expression reasoner. http://www.racer-systems.com/

PAR, precipitation, carbon dioxide (CO_2) concentration, acidity (pH), alkalinity, salinity, water depth, turbidity, light quanta, cloud cover, spatial resolution, spatial instrument ID, sensor ID, percent of coral coverage, percent of algal coverage and the fast growth check. The asserted object properties were "type of human influence" and "has part" which represented the population information and the community composition of the reef. As shown in Table 1 the scenario variables available for manipulation were the number of reefs, the collection intervals, the number of asserted property values and the number of atoms in each SWRL inference rule.

The focus of the scenario parameters in the first assessment was the performance versus the scaling of triples. The scenarios increased the number of triples in two ways:

- Additional reef instances – Either the number of reefs (3 or 4) with daily versus half-hourly collection intervals; and
- Additional property assertions to each individual – SST only, which required 13 property assertions versus all environmental values (26 assertions). The community composition assertions are common in all examples.

The time to run the inference rules with a growing number of triples was the focus of the second set of assessments. To compare triple quantity versus inference time performance the inference rules for the coral bleach indices were applied with the following attributes (Table 1):

- A growth in triples via additional reef instances (i.e. the number of reefs) or additional properties (i.e. 13 property assertions (SST only) versus all values from the 26 asserted properties; and
- An increase in the number of atoms that comprise the SWRL rules (5, 9 or 16 atoms).

6.3. Loading and reasoning functionality results

Seven versions of the KB were trialled with three runs each. Table 2 depicts the averaged results. For the purpose of a comparison benchmark, the first KB version, labelled A, is empty of any reef instances. The other six versions, labelled B through to G, changed the composition of the KB by the number of reef instances, properties asserted and the temporal intervals of each reef instance. The number of total instances and triples in the KB and the resulting time in seconds taken to run the reasoning engines were the key factors in the performance tests. Due to memory allocation errors in handling larger numbers of triples there is a lack of data for the RacerPRO reasoner and incomplete data for the FaCT++ reasoner.

Table 3 shows a comparison of the KB versions driven by the scenario parameters, and the statistical results of four scenarios. The KB version legend from Table 2 is indicated in Table 3 to depict the comparison operands of the four scenarios and the correlation coefficient and marginal percentage are shown for each.

The correlation coefficient analysis looks at bivariate sets of data that compare the test outcome of the KB versions (e.g., A, B, C, etc.) and the change in processing time versus either the number of instances or the number of triples in the KB. Figure 8 shows a scatter plot diagram of one correlation comparison: the relationship of KB B and C in scenario 2 of Table 3, firstly for instances versus time to process and then triples versus time. This graphic is indicative of the correlation coefficients for all scenarios. There is a strong correlation relationship between the duration and the rise in triple quantity. The correlation was weaker for the number of instances asserted to the KB versus processing time. The increase or decrease in marginal percentage showed a linear progression in scale.

6.4. Inference rules atomic quantity functionality results

This set of performance analyses focuses on the inference rules and inference engine where the actual SWRL rule versus the processing time was examined. Because each rule is a series of atoms and each atom relates to an asserted property, a class member or SWRL built-in, they require

processing time to port to the Jess inference engine via the SWRL Bridge. For the purpose of this paper one rule was chosen from the original study to illustrate the outcome [8]. The rule contains sixteen atoms that refer to all possible environmental property values in the current KB (PAR, pH, salinity, etc.).

Table 4 logs and compares the change in processing times. The comparison is between the loading and running the inference engine versus the number of triples and the number of reef instances. The correlation coefficient was distinctly positive for both time versus triples or reef instances and a linear scale relationship of the processing time and the growth in triples or reef instances was observed.

The linear relationship is also independent of the number of triples or the number of instances. The antecedents in the inference rules are constant for each test and port only the relevant property or instant to the Jess inference engine. Therefore, as opposed to the dominant correlation of only time versus quantity of triples from the previous analysis, this outcome would be similar, independent of the number of triples or reef instances.

The time involved in running the reasoning engine and the inference engine were directly relative to the quantity of triples in the Knowledge Base. At 600,000 triples the system was taking longer periods of time to process; however, it was successfully completing the task. The system could not be scaled to billions of triples due to memory limitations (e.g., the Java Virtual Machine finite memory allocation). However, Protégé has been tested up to two million triples, which would make the Semantic Reef system an efficient desktop hypothesis-driven tool for posing questions of small to medium scale datasets.

Although currently limited to two million triples, the KB can conceivably handle years of data for a reef or number of reefs. Proportionally, this amount of triples can equate to changes in the scenarios, such as adding reefs (up to approximately 250 reefs), extending the durations (annual data versus summer), changing data logging intervals (e.g., hourly to daily), adding previous years data for long term analysis, or additional asserted properties for environmental parameters. Further, to alleviate replication and problems of scale, as discussed in the previous section, numerous models of reef-types are represented concurrently by proxy reef data. Reef instances can be automatically classified to simultaneously belong to a reef type “by proximity”, “by climate factors” or “by community composition”, among others.

The addition of quality assurance functionality is a component of future work. Currently the KB has no internal quality checks, but instead assumes the incoming data is already quality assured. However, this functionality would add considerably to execution times. Instead source data quality assurance is better processed by the Kepler workflow where checking for gaps in data sequences, adding a scale of belief as provenance annotations, etc. are more easily coded. The Semantic Reef system will need to incorporate a quality assurance mechanism for the resulting data if it is to achieve a complete solution to automate certain data processing tasks and alleviate manual intervention.

7. DISCUSSION AND CONCLUSION

New data collection methods that scale-up for single instruments and/or scale-out across many sensors/locations are resulting in a data deluge that demands new and/or evolved research practices.

The Semantic Reef architecture was designed to explore some of the data bottlenecks that are the result of the data deluge. The demand for automatic data analysis and hypothesis testing is emerging. They will be even more important as current and future data production and collection infrastructures are deployed (such as sensor networks). The system’s development involved merging technologies to explore their potential synergies and observe how they may help solve current research problems.

The architecture of the Semantic Reef is an exemplar of the evolving methods for managing rich data sources in ways that can be more effective. The architecture employs semantic inference

includes methods for modularity, reusability and data integration. Together these methods offer benefits in flexible hypothesis design to foster knowledge discovery. The modular ontology design within the KB aims to simultaneously maximise both reuse and usability of data for different hypotheses through a new hybrid of current methods to achieve a separation of data instances from the concept descriptions [26]. The KB can be easily reused for different investigations by simply repopulating it with data and information relevant to a specific study

The functionality of the Semantic Reef on a desktop platform typical of a researcher's *in silico* environment was tested. The quantity of data (i.e., triples) introduced to the system versus the time to load the KB and then reason and infer over the KB were the focus of the performance analyses [48, 50]. Although restricted by the limitations of a desktop environment, the tests proved the system to be an efficient hypothesis-driven tool for posing questions of small to medium scale datasets. That is, the number of triples stored in the KB does not need to be extensive because only instances necessary to a specific hypothesis need to be imported to the system.

The Semantic Reef use-case is a subset of broader eco-informatics applications, which can help process the large volumes of data to create knowledge. The architecture offers an alternative approach to the development, application and execution of observational hypotheses in the coral reef ecosystem domain and may be extended to other research applications. The system has been tested and proven to handle limited quantities of disparate data for a range of propositional suppositions and can extract or disclose phenomena within the data. Concrete examples of the capabilities have been demonstrated by inferring a coral bleaching alert and posing hypotheses to explore the causal factors of the bleaching phenomena.

Finally, the Semantic Reef system is capable of disclosing or extracting anomalies, phenomena, and knowledge in data from disparate sources. Most propositions could be processed on a desktop computer with sample data imported to develop the rules and hypotheses for *in situ* observations. Research tools of this type are important given the emergence of diverse new data sources and the complexity of environmental issues globally.

ACKNOWLEDGEMENTS

The authors wish to thank Dr. Ron Johnstone, Dr. Glen Holmes from the University of Queensland, Jeff Maynard from University of Melbourne, Scott Bainbridge and Stuart Kininmonth from the Australian Institute of Marine Science for their expertise in the marine domain. We also thank Professor Marimuthu Palaniswami from University of Melbourne, Jarrod Trevathan from Griffith University and Richard Monypenny from James Cook University for their interest, helpful discussions and feedback.

BIBLIOGRAPHY

[1] T. Hey and A. E. Trefethen, The Data Deluge: an e-Science perspective, in: Grid Computing - Making the Global Infrastructure a Reality, Berman F, Fox GC, and Hey AJG, Eds. West Sussex, England: John Wiley and Sons Ltd., 2003, pp. 809-824.

[2] NOAA-ICON/CREWS, Integrated Coral Observing Network/Coral Reef Early Warning System, http://www.coral.noaa.gov/crews/ [February, 2012].

[3] GBROOS, Great Barrier Reef Ocean Observing System, http://www.imos.org.au/nodes/great-barrier-reef-observing-system.html [January, 2012].

[4] J. Trevathan, I. Atkinson, W. Read, N. Bajema, Y. J. Lee, R. Johnstone, and A. Scarr, Developing Low-Cost Intelligent Wireless Sensor Networks for Aquatic Environments, Presented at the Proceedings of the International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP'10), Brisbane, Australia, 2010.

[5] C. Goble, O. Corcho, P. Alper, and D. De Roure, e-Science and the Semantic Web: a symbiotic relationship, Proceedings of the Proceedings from the 9th International Conference in Discovery Science (DS 2006), Barcelona, Spain, (2006).

[6] W. Hall, D. D. Roure, and N. Shadbolt, The evolution of the Web and implications for eResearch Phil. Trans. R. Soc. A 2009; **367**(1890): 991-1001.
[7] IPCC, Climate change 2007: the physical basis, Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change, Cambridge Univ. Press, Cambridge, U. K. 2007.
[8] T. Myers, *Applying semantic technologies and artificial intelligence to eco-informatic modelling of coral reef systems*. PhD Thesis, James Cook University: Townsville, QLD, 2010.
[9] G. Antoniou and F. van Harmelen, *A Semantic Web primer (2nd edition)*, 2nd ed. The MIT Press: Cambridge, MA, USA, 2008.
[10] T. Gruber, A translation approach to portable ontology specifications. Knowledge Acquisition 1993; **5**(2): 199-220.
[11] N. Guarino, Understanding, building and using ontologies. International Journal of Human-Computer Studies 1997; **46**(2-3): 293-310.
[12] H. Wache, T. Vogele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann, and S. Hubner, Ontology-based integration of information - a survey of existing approaches, Presented at the 17th International Joint Conference on Artificial Intelligence (IJCAI 01) Workshop: Ontologies and Information Sharing, Seattle, WA, USA, 2001.
[13] T. Heath and C. Bizer, Linked Data: Evolving the Web into a Global Data Space, in: Synthesis Lectures on the Semantic Web: Theory and Technology, vol. 1, J. Hendler and F. vanHarmelen, Eds.: Morgan & Claypool Publishers, 2011, pp. 1-136.
[14] CC, Creative Commons, http://creativecommons.org/ [November, 2011].
[15] A. Sheth, C. Henson, and S. S. Sahoo, Semantic Sensor Web. IEEE Internet Comput 2008; **12**(4): 78-83.
[16] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, et al., Scientific workflow management and the Kepler system. Concurrency and Computation: Practice and Experience 2006; **18**(10): 1039-1065.
[17] T. Oinn, M. Addis, J. Ferris, D. Marvin, M. Senger, M. Greenwood, T. Carver, et al., Taverna: a tool for the composition and enactment of bioinformatics workflows. Bioinformatics 2004; **20**(17): 3045-3054.
[18] I. Taylor, M. Shields, I. Wang, and A. Harrison, The Triana workflow environment: architecture and applications, in: Workflows for e-Science. London, England: Springer, 2007, pp. 320-339.
[19] J. Yu and R. Buyya, A taxonomy of scientific workflow systems for grid computing. SIGMOD Rec. 2005; **34**(3): 44-49.
[20] W. Michener, J. Beach, M. Jones, B. Ludäscher, D. Pennington, R. Pereira, A. Rajasekar, et al., A knowledge environment for the biodiversity and ecological sciences. Journal of Intelligent Information Systems 2007; **29**(1): 111-126.
[21] D. De Roure, C. Goble, and R. Stevens, The design and realisation of the myExperiment virtual research environment for social sharing of workflows. Future Generation Computer Systems 2009; **25**(561-567.
[22] C. A. Goble, J. Bhagat, S. Aleksejevs, D. Cruickshank, D. Michaelides, D. Newman, M. Borkum, et al., myExperiment: a repository and social network for the sharing of bioinformatics workflows. Nucleic Acids Research 2010; **38**(suppl 2): W677-W682.
[23] C. A. Henson, H. Neuhaus, A. P. Sheth, K. Thirunarayan, and R. Buyya, An ontological representation of time series observations on the Semantic Sensor Web, Proceedings of the 1st International Workshop on the Semantic Sensor Web (SemSensWeb 2009), Crete, Greece, (2009).
[24] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, et al., Examining the challenges of scientific workflows. IEEE Computer 2007; **40**(12): 24-32.
[25] D. L. McGuinness and F. van Harmelen, OWL: Web Ontology Language overview http://www.w3.org/TR/owl-features/ [August, 2011].
[26] T. S. Myers, I. M. Atkinson, and R. Johnstone, Supporting coral reef ecosystems research through modelling a re-usable ontology framework. J. Appl. Artif. Intell 2010; **24**(1): 77-101.
[27] G. R. Holmes, Estimating three-dimensional surface areas on coral reefs. J. Exp. Mar. Biol. Ecol 2008; **365**(1): 67-73.
[28] B. C. Grau, I. Horrocks, Y. Kazakov, and U. Sattler, Modular reuse of ontologies: theory and practice. J. Artif. Intell. Res 2008; **31**(1): 273-318.
[29] T. P. Hughes, A. H. Baird, D. R. Bellwood, M. Card, S. R. Connolly, C. Folke, R. Grosberg, et al., Climate change, human impacts, and the resilience of coral reefs. Science 2003; **301**(5635): 929-933.

[30] N. F. Noy and D. L. McGuinness, Ontology development 101: a guide to creating your first ontology, Knowledge Systems Laboratory, Stanford Medical Informatics, Stanford, CA, USA, Technical Report SMI-2001-0880, 2001.
[31] M. Uschold and M. King, Towards a methodology for building ontologies, Presented at the Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, International Joint Conference on Artificial Intelligence. 14th International Joint Conference on Artificial Intelligence (IJCAI 1995), Montreal, Canada, 1995.
[32] M. Jarrar and R. Meersman, Ontology engineering - the DOGMA approach, in: Advances in Web Semantics I, vol. 4891, Lecture Notes in Computer Science. Berlin/Heidelberg, Germany: Springer, 2008, pp. 7-34.
[33] F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, The Description Logic handbook theory, implementation and applications (2nd ed.), F. Baader, D. Calvanese, D. L. McGuinness, D. Nardi, and P. F. Patel-Schneider, Eds., 2nd ed. ed. Cambridge, USA: Cambridge University Press, 2007, pp. 545.
[34] M. J. O'Connor, S. W. Tu, C. I. Nyulas, A. K. Das, and M. A. Musen, Querying the Semantic Web with SWRL, Proceedings of the The International RuleML Symposium on Rule Interchange and Applications (RuleML2007), Orlando, FL, USA, (2007).
[35] T. Myers, I. Atkinson, and R. Johnstone, Semantically Enabling the SEMAT Project: Extending Marine Sensor Networks for Decision Support and Hypothesis Testing, Presented at the the 4th International Conference on Complex, Intelligent and Software Intensive Systems (CISIS'10) 3rd IEEE International Workshop on Ontology Alignment and Visualization (OnAV'10), Krakow, Poland, 2010.
[36] I. Horrocks, P. F. Patel-Schneider, and F. van Harmelen, From SHIQ and RDF to OWL: the making of a web ontology language. J. Web. Semant 2003; **1**(1): 7-26.
[37] D. L. McGuinness, Question answering on the semantic Web. Intelligent Systems, IEEE 2004; **19**(1): 82-85.
[38] R. J. Brachman and H. J. Levesque, *Knowledge representation and reasoning*. Morgan Kaufmann: San Francisco, CA, USA, 2004.
[39] M. O'Connor and A. Das, SQWRL: a query language for OWL, Presented at the OWL: Experiences and Directions (OWLED 2009), Virginia, USA, 2009.
[40] R. Jones, O. Hoegh-Guldberg, and A. Larkum, Temperature-induced bleaching of corals begins with impairment of the CO2 fixation mechanism in zooxanthellae. Plant Cell and Environment 1998; **21**(12): 1219-1230.
[41] R. Berkelmans, G. De'ath, S. Kininmonth, and W. J. Skirving, A comparison of the 1998 and 2002 coral bleaching events on the Great Barrier Reef: spatial correlation, patterns, and predictions. Coral Reefs 2004; **23**(1): 74-83.
[42] J. Maynard, K. Anthony, P. Marshall, and I. Masiri, Major bleaching events can lead to increased thermal tolerance in corals. Marine Biology 2008; **155**(2): 173-182.
[43] AIMS, Marine blueprint - climate change and the fate of the Great Barrier Reef, http://www.aims.gov.au/docs/research/climate-change/position-paper.html [Novermber, 2011].
[44] NOAA, Coral Reef Watch - methodology and description, http://coralreefwatch.noaa.gov/satellite/methodology/methodology.html [October, 2011].
[45] J. A. Maynard, P. J. Turner, K. R. N. Anthony, A. H. Baird, R. Berkelmans, C. M. Eakin, J. Johnson, et al., ReefTemp: an interactive monitoring system for coral bleaching using high-resolution SST and improved stress predictors. Geophys Res Lett. 2008; **35**(L05603): 1-5.
[46] S. Kininmonth, S. Bainbridge, I. Atkinson, E. Gilla, L. Barrald, and R. Vidaude, Sensor networking the Great Barrier Reef. Spatial Sciences Qld. Journal 2004; **Spring 2004**(1): 34-38.
[47] S. Liang, P. Fodor, H. Wan, and M. Kifer, OpenRuleBench: an analysis of the performance of rule engines, Presented at the 18th international conference on World Wide Web (WWW'09), Madrid, Spain, 2009.
[48] B. Bishop and F. Fischer, Iris-integrated rule inference system, Presented at the Proceedings of the Workshop on Advancing Reasoning on the Web: Scalability and Commonsense (ARea2008), the 5th Annual European Semantic Web Conference (ESWC 2008), Tenerife, Spain., 2008.
[49] S. Bechhofer, R. Möller, and P. Crowther, The DIG Description Logic interface, Presented at the Proceedings of the International Workshop on Description Logics (DL2003), Rome, Italy, 2003.
[50] S. Liang, P. Fodor, H. Wan, and M. Kifer, Openrulebench: An analysis of the performance of rule engines, Presented at, 2009.

FIGURES

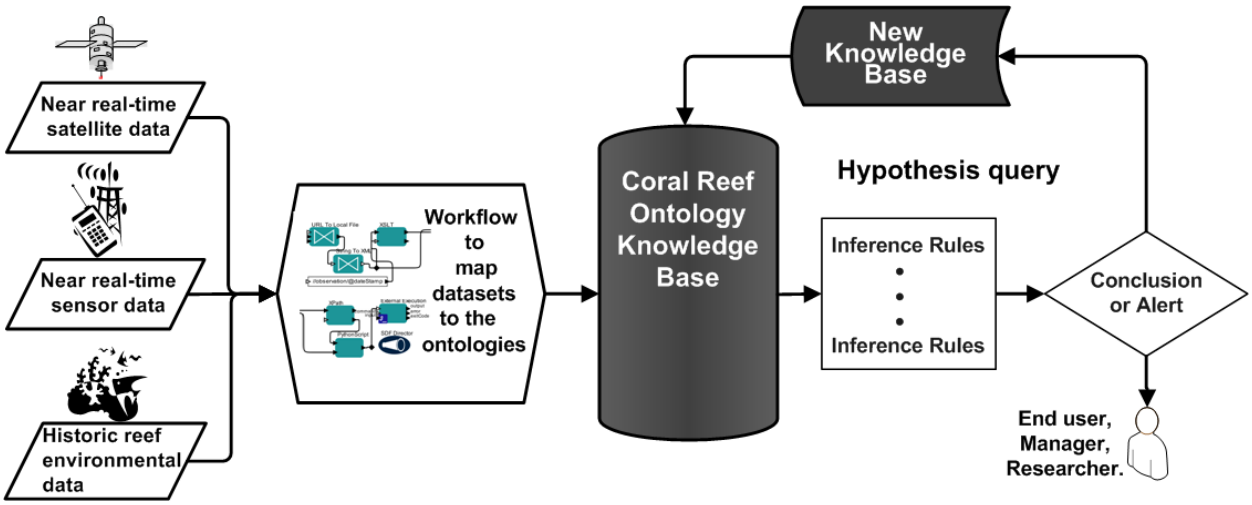


Figure 1 – The end-to-end Semantic Reef workflow.

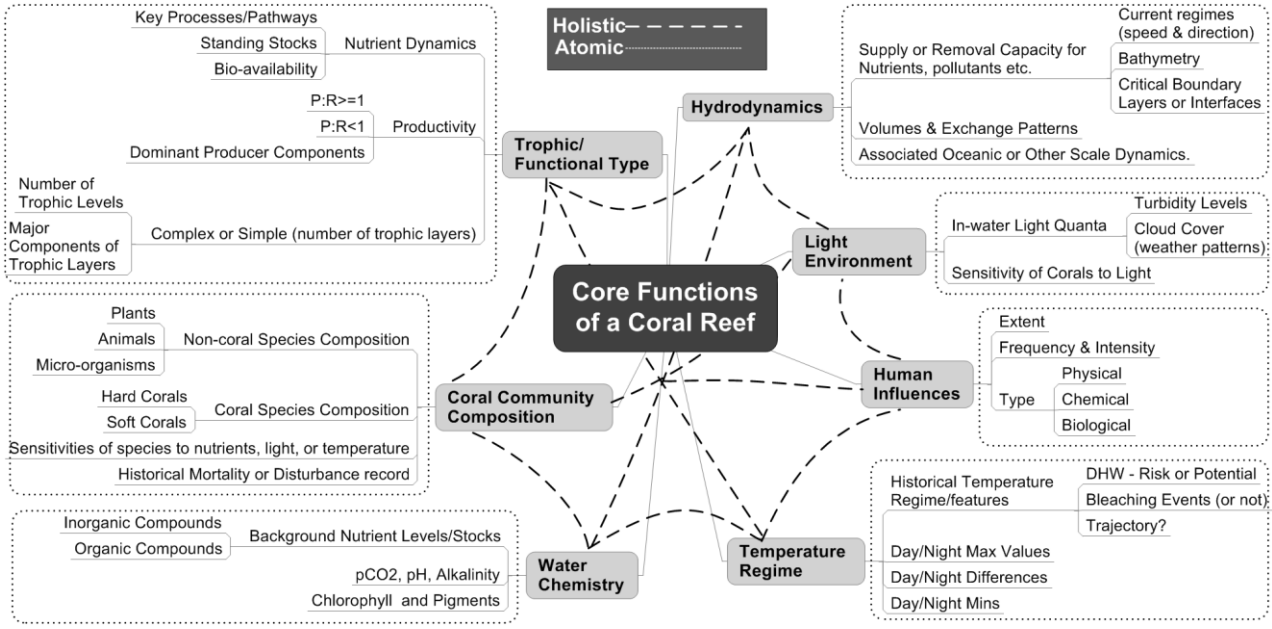


Figure 2 - Coral Reef functional concepts supplied from a marine expert – Each function has a natural hierarchy of sub-functions or related factors.

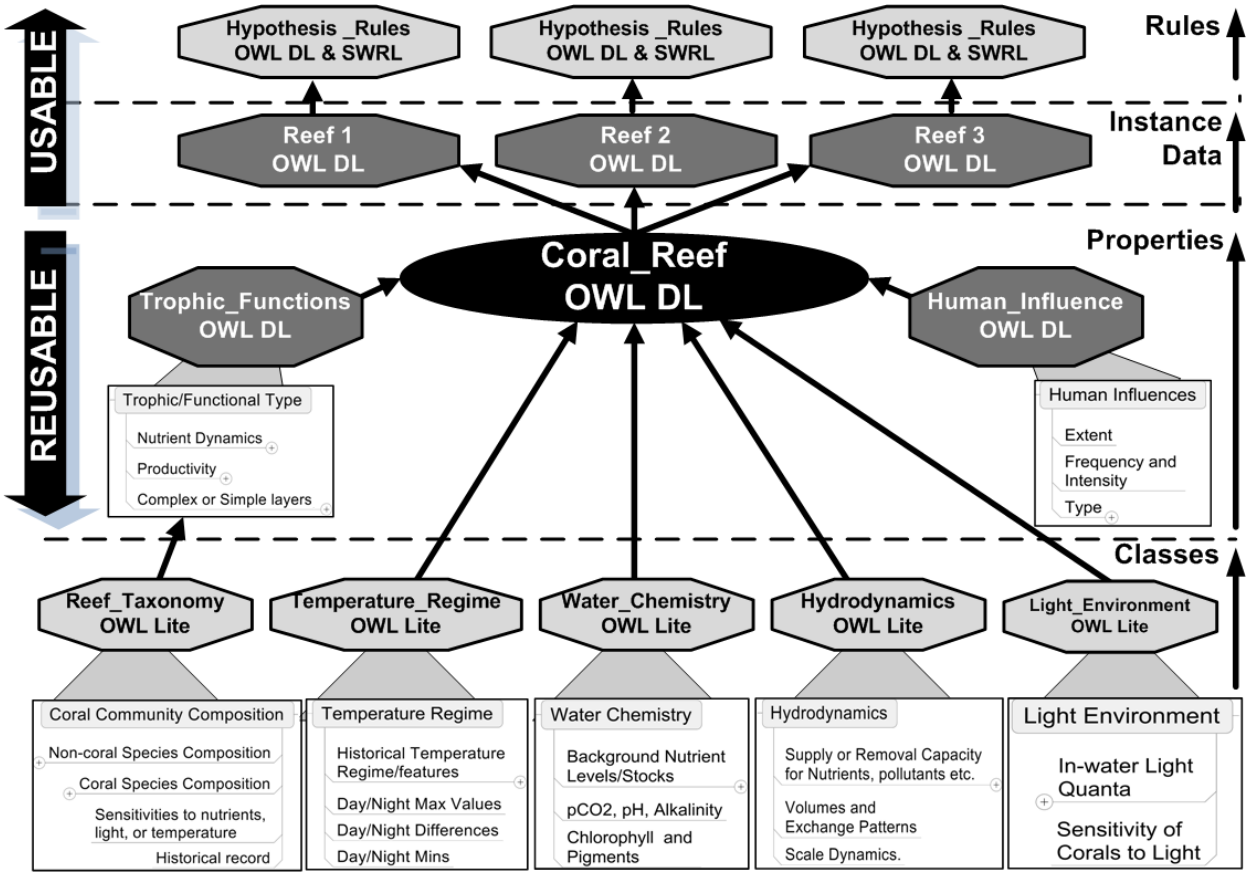


Figure 3 – A domain experts coral reef model segmented into a hierarchy of informal to formal ontologies. The inter-ontology methodology supports simultaneous reusability and usability by separating the domain ontologies from the applications ontologies.

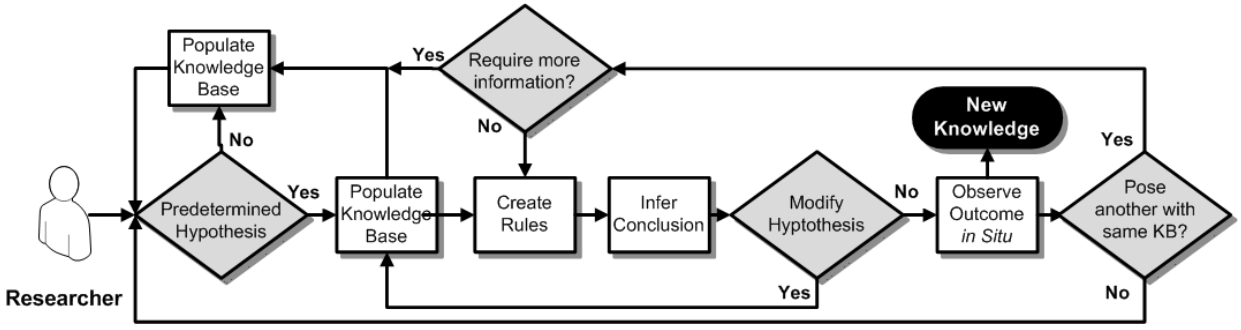


Figure 4 – A flowchart of the hypothesis design process. The propositions are fully flexible in light of new ideas or additional interesting data.

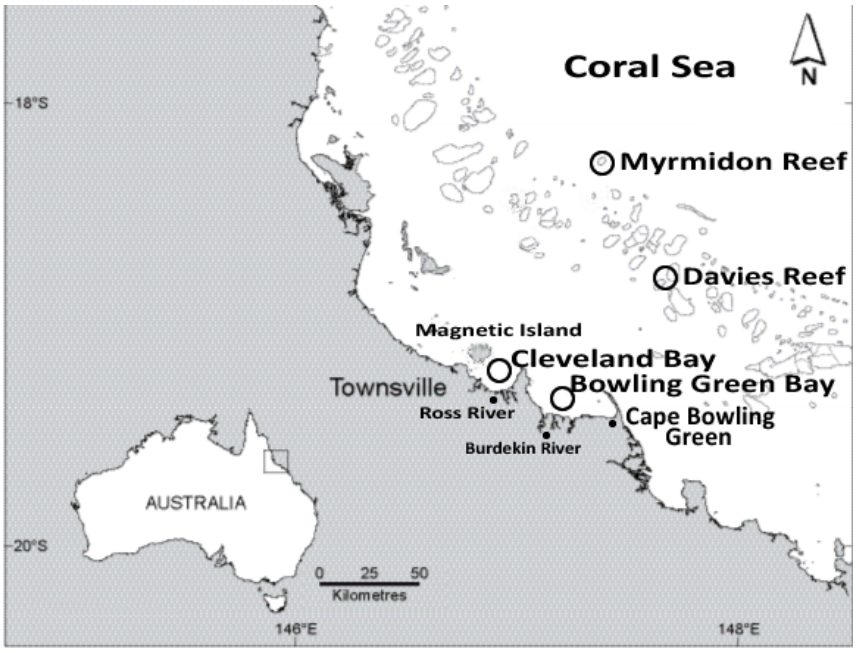


Figure 5 – Sitemap of the targeted reefs (central section of the GBR).

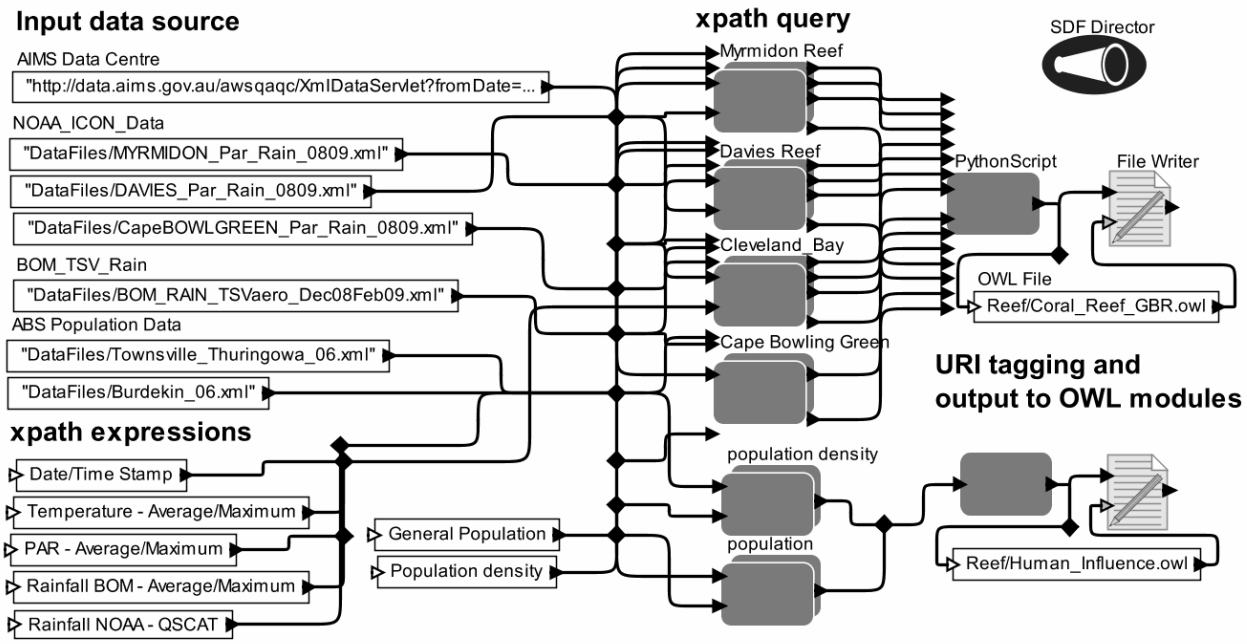
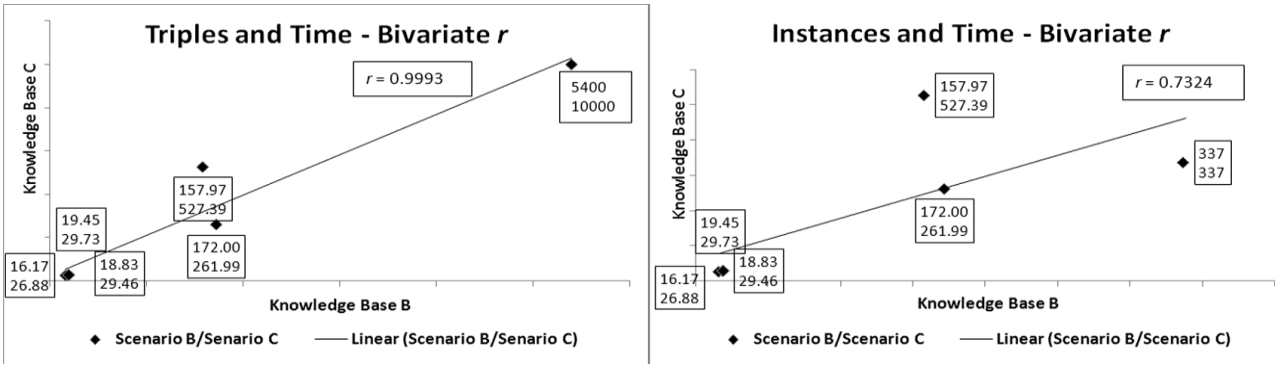


Figure 6 – A Kepler workflow to populate the KB with PAR, rain, salinity and SST data from AIMS, NOAA and BOM and human population quantity and density from the ABS. XPATH and Python actors were initiated to achieve the data transform and population of the KB.

Figure 7 – After classification with the Pellet reasoner the reefs were subsumed to belong to the correct reef type.



(a) **(b)**

Figure 8 – Correlation Coefficient example depicts the comparative relationship of Scenario 2 between KB version B (3 reefs, SST only) and KB version C (3 reefs, all environment values asserted):
(a) Instances and time (r =.73), (b) Triples and time (r =1)

TABLES

Table 1 – A matrix of the testing attributes – the variations in the growth of triple and reef instance quantity

Attributes	Comparison Matrix																			
Number of reefs	3										4									
Temporal intervals	Daily					Half hourly					Daily					Half hourly				
Number of property assertions	13		26			13		26			13		26			13		26		
Number of inference rule atoms	5	9	5	9	16	5	9	5	9	16	5	9	5	9	16	5	9	5	9	16

Table 2 – KB versions and legend – The test results for quantity of triples versus time to load the KB and run the reasoning engine.

REASONER TEST 3 MONTHS	Legend	Instances	Triples	Load Triples(s)	Load KB(s)	Protégé 3.4 Pellet(s)	Protégé 4 Pellet(s)	Protégé 4 FaCT++(s)
N0 ASSERTED INSTANCES	A	67	160	8.59	10.78	11.53	2.70	83.65
3 REEFS/ DAILY / SST ONLY	B	337	5400	16.17	18.83	19.45	172.00	157.97
3 REEFS/ DAILY / ALL VALUES	C	337	10000	26.88	29.46	29.73	261.99	527.39
3 REEFS/ HALF-HOURLY/ SST ONLY	D	12886	250000	344.53	397.81	734.14	7746.09	
3 REEFS/ HALF HOURLY/ ALL VALS	E	12886	440000	772.03	831.17	1347.45	15993.75	
4 REEFS/ HALF HOURLY/ SST ONLY	F	17159	330000	467.11	543.13	1003.90	10754.92	
4 REEFS/ HALF HOURLY/ ALL VALS	G	17159	590000	1061.02	1157.42	3755.30	27372.18	

Table 3 - The marginal percentage and correlation coefficients for the four comparison scenarios. The results show a correlation between the number of triples versus the time to load and reason over the KB (*an example graph of the Correlation Coefficient for the B&C comparison is depicted in Figure 7).

Legend	Marginal Percentage increase/ decrease (%)							Correlation Coefficient Triples vs. Time	Correlation Coefficient Instances vs. Time
	Instances	*Triples*	*Loading Triples(s)*	*Load KB(s)*	*Protégé 3.4 Pellet(s)*	*Protégé 4 Pellet(s)*	*Protégé 4 FaCT++(s)*		
Scenario 1 - Compare KB (no Reef Instances) growth in triples via quantity and property assertion									
A&B	80.12	97.04	46.9	42.7	40.7	98.4	47.0	0.886	0.651
A&C	80.12	98.40	68.0	63.4	61.2	99.0	84.1	0.898	0.848
A&D	99.48	99.94	97.5	97.3	98.4	100.0	99.7	0.997	0.758
A&E	99.48	99.96	98.9	98.7	99.1	100.0	99.9	0.997	0.377
A&F	99.61	99.95	98.2	98.0	98.9	100.0	99.9	0.997	0.742
A&G	99.61	99.97	99.2	99.1	99.7	100.0	99.9	0.996	0.220
Scenario 2 – Amount of property assertions – Average SST versus All property values									
*B&C	0.00	46.0	39.8	36.1	34.6	34.3	70.0	*1.00	*0.730
D&E	0.00	43.2	55.4	52.1	45.5	51.6	45.1	1.00	0.890
F&G	0.00	44.1	56.0	53.1	73.3	60.7	44.6	1.00	0.815
Scenario 3 – Amount of reef instances – Temporal intervals - Daily versus Half hourly									
B&D	97.38	97.8	95.3	95.3	97.4	97.8	99.5	1.00	0.526
C&E	97.38	97.7	96.5	96.5	97.8	98.4	99.1	1.00	0.947
Scenario 4 – Amount of triples –SST only and All property values- 3 reefs versus 4 reefs									
D&F	24.90	24.2	26.2	26.8	26.9	28.0	44.1	1.00	1.00
E&G	24.90	25.4	27.2	28.2	64.1	41.6	43.6	1.00	0.984

Table 4 - The marginal percentage and correlation coefficients for a rule with 16 atoms. The number of triples and asserted, or inferred, instances versus the time to load the rules to the Jess inference engine

INFERENCE RULES TEST (GBR_Rules.owl)	Legend	Instances	Triples	Property Assertion	Load Rule to Jess	Inferred Instances	Correlation Coefficient Triples vs. Time	Correlation Coefficient Instances vs. Time
3 REEFS/ DAILY	A	270	25000	2160	3.67	15		
3 REEFS/HLF-HOURLY	B	12819	455000	102552	147.30	2915		
4 REEFS/HLF-HOURLY	C	17092	605000	136736	198.00	2915		
Marginal percentage increase/decrease	A&B	97.9	94.5	97.9	97.5	99.5	0.9903	0.9998
	B&C	25.0	24.8	25.0	25.6	0.0	1.0000	1.0000
	A&C	98.4	95.9	98.4	98.1	99.5	0.9901	0.9999