World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:7, No:8, 2013

Semantic Support for Hypothesis-Based Research from Smart Environment Monitoring and Analysis Technologies

T. S. Myers, J. Trevathan

***Abstract*—**Improvements in the data fusion and data analysis phase of research are imperative due to the exponential growth of sensed data. Currently, there are developments in the Semantic Sensor Web community to explore efficient methods for reuse, correlation and integration of web-based data sets and live data streams. This paper describes the integration of remotely sensed data with web-available static data for use in observational hypothesis testing and the analysis phase of research. The Semantic Reef system combines semantic technologies (e.g., well-defined ontologies and logic systems) with scientific workflows to enable hypothesis-based research. A framework is presented for how the data fusion concepts from the Semantic Reef architecture map to the Smart Environment Monitoring and Analysis Technologies (SEMAT) intelligent sensor network initiative. The data collected via SEMAT and the inferred knowledge from the Semantic Reef system are ingested to the Tropical Data Hub for data discovery, reuse, curation and publication.

***Keywords*—**Information architecture, Semantic technologies Sensor networks, Ontologies.

I. Introduction

SENSOR data is prolific and growing. There is extensive development in sensor technologies due to the economic and productivity advantages they offer data acquisition (such as the *Smart Environment Monitoring and Analysis Technologies* (SEMAT) project [2]). Sensor networks are being increasingly used to gather data in real-time, across widely distributed areas for environmental monitoring. While the amount of data being collected by sensor networks is increasing, there are constraints in the data integration phase. Whether data is acquired via event-driven or timed methods, the flow from the initial collection of sensed data to the output of new knowledge has many phases where problems or constraints occur.

To alleviate the bottlenecks that arise from the unequal balance of input versus output, there are ongoing research activities around sensor data and its integration with other data sources [3]-[5]. The requirements for data synthesis, correlation, simulation and visualization pose equally dramatic challenges in data analysis and the representation of the results. Traditionally, powerful supercomputers providing computational and storage resources have met these needs. However, workstations equipped with sophisticated analysis tools and special purpose visualization hardware can alleviate the bottlenecks found in the data processing phase of research. Personal computing resources can manage smaller scale flexible hypothesis-testing through systems such as the *Semantic Reef* (SR) project [1] without significant support from additional high-end resources. The SR system combines semantic technologies (e.g., well-defined ontologies and logic systems) with scientific workflows to enable hypothesis-based research.

This paper presents a framework for integrating remotely sensed data with web-available static data for use in observational hypothesis testing and the analysis phase of research. We describe how the data fusion concepts from the SR architecture map to the SEMAT sensor network project. An example is provided of how the framework operates using an algae study that collects data via SEMAT and from other external sources. The data and the inferred knowledge from the SR system are then ingested to the *Tropical Data Hub* (TDH) [6] online repository for data discovery, reuse, curation and publication. The integration of these tools, automation of the workflow, and the simplicity and flexibility of the processing power required, can significantly reduce the burdens of deriving knowledge from voluminous amounts disparate data sources collected via sensor networks.

This paper is structured as follows. Section II discusses related work and the motivation for our proposed approach. Section II describes the SR knowledge base components and architecture in the context of marine research. Section III illustrates the sensor data streams and layers of the SEMAT project. Section IV describes the TDH, which is used to maximize data fusion, discovery and reuse. Section V illustrates the benefits of linking the SR and TDH and SEMAT to provide data fusion, hypothesis testing and data reuse. Section VI gives some concluding remarks.

II. Related Work

Recent significant advances in science have been achieved through the sharing of complex interdisciplinary skills, data and analysis [7]. The creation of new knowledge and its reuse is contributed to by making connections between disparate ideas, domains, people and data. The capability to automatically process, manipulate and mine data that is

T.S. Myers is with the School of Business (Information Technology), James Cook University, Townsville, Queensland, 4811, Australia (phone: 6107 4781 6908; e-mail: Trina.Myers@jcu.edu.au).

J. Trevathan is with the School of Information and Communication Technology, Griffith University, Brisbane, Queensland, 4111, Australia (phone: 6107 3735 5046; e-mail: j.trevathan@griffith.edu.au).

streamed from sensed networks or stored in vast distributed digital repositories and/or discreet data silos is becoming increasingly imperative [8].

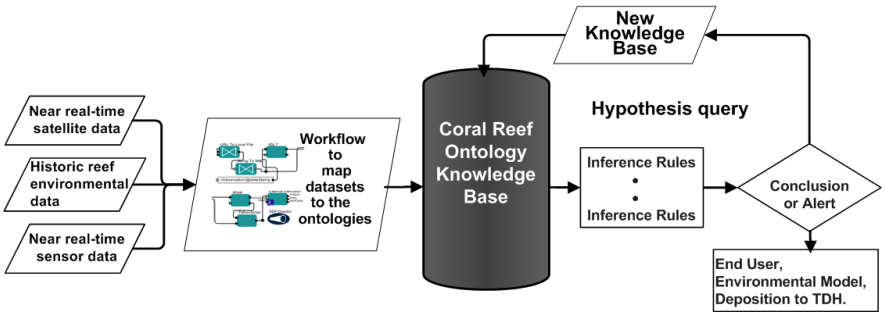

Fig. 1 The end-to-end Semantic Reef framework [1]

A highly relevant initiative to apply semantic technologies to manage sensed data is the Semantic Sensor Web (SSW) [4]. The SSW project aims to provide an environment for enhanced query and reasoning within a sensor network and effectively connects 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 [9]. This proposed technique builds on current Sensor Web standardization efforts within the W3C and Open Geospatial Consortium (OGC) by extending them with Semantic Web technologies [4]. The SSW can apply complex queries about weather data collected from the urban Geographic Information System (GIS) systems and weather services. The data can be queried, reasoned over and/or have inference rules applied, including rules to automate alerts[4].

As similar initiative is the Linked Stream Middleware (LSM), which is a platform that brings together the live real world sensed data and the Semantic Web in an integrated model. The LSM provides functionalities to access stream sources and transform the raw data into semantically enabled "Linked Stream Data". Once in linked data format, data annotation and visualization is possible via a web interface and live querying is possible via a SPARQL endpoint [5].

The main differentiations between these projects and the framework proposed in this paper are the level of applied semantics and the data scope. Firstly, 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. In contrast, the LSM focuses at the RDF and SPARQL query levels of the semantic web by adding semantic annotation to the sensor layers as metadata of sensor data for access to sensor data streams.

Secondly is the 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 and LSM framework both permit any digitalized data from any openly available linked data source. LSM is focused predominantly with sensed data sources that can be integrated with open linked data for query processing.

In contrast, the Semantic Reef model is an atomistic application that focuses predominantly on hypothesis-based research in the subset of coral reef ecosystems. The scope of the SSW's data is limited to urban terrestrial data. The SSW merges data gathered from urban instruments (e.g., remote sensors, video and other cameras devices, etc.) with the collection and analysis process. 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 Semantic Reef project can benefit from the resources made available through the LSM 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 sensor data, which are available via the LSM portal[1], as resources. 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.

III. Semantic Reef

The SR project aims to utilise existing marine databases, augmented by real-time sensor output, to pose hypotheses of the disparate data. This knowledge representation system employs Semantic Web technologies and scientific workflows to resolve the problems of data integration, synthesis and discovery for marine ecosystems (Fig. 1).

[1] http://lsm.deri.ie/

A. The Semantic Technologies

Semantic Web technologies allow for a flexible scalable environment to model abstract and concrete concepts in a way that is "understandable" to the machine [10]. Ontologies are the foundation of these technologies that explicitly define concepts and information to be "computer-understandable". The vocabulary and terms that describe the entities within domains, and the relationships they have with each other, are defined using axioms and restrictions that constrain the interpretation for use by the computer, making the concept computer-processable (i.e., "computer-understandable") [11]. These definitions enable the computer to make intelligent inferences, decisions and/or discoveries, using logic systems such as *Description Logics* (DL) and propositional logic.

B. Ontology Design to Support Hypothesis-Based Research

A set of re-usable ontologies have been developed with a modular design to describe the concept of a marine environment [12]. The hierarchical modular design is an example of component architecture that makes repopulation and reuse of the knowledge base possible.

The complex coral reef eco-system relationships were conveyable to the constructs of the ontology languages. For example, the consequences and connections between "human influence", such as an oil spill, and other principal components within the KB, such as light intensity levels, can be modeled with well-defined axioms. In the case of an oil spill, the information introduced to the "human influence" ontology (i.e., impact factor, intensity and extent) changes properties within the "light environment" ontology, particularly the in-water light intensities level, as they are interconnected. In contrast, the "human influence" component can be defined at the atomic level by the composite of factors and properties (i.e., the type, the degree, the frequency and the extent of the influence). These factors can all be categorically described from low through too high and have interactive relationships with the each other.

The scope of complexity in ontological design spans a range of data models with varying degrees of granularity to serve a distinct purpose. The hierarchy of ontologies range from "informal" taxonomies and vocabularies of domain concepts, such as animal and plant species or environmental elements, to the complex "formal" ontologies that incorporate logic systems for autonomically inferring new knowledge. The choice of ontology was determined by the extensibility and expressiveness required; that is, by the information or knowledge it is designed to produce.

The informal ontologies are imported to the more complex formal DL ontologies and so forth to form a "ground-up" physical hierarchy within the KB. Except for the highest-level "domain-specific" ontology, all ontologies in the hierarchy are independent of any particular coral reef and its environment or human influential factors.

This modularity enables greater adaptability in hypothesis design and testing. To illustrate, one line of enquiry may be the discovery of casual factors of coral bleaching while another is to investigate algal blooms. The hypotheses are very diverse and require different data from separate locations. The re-usable ontology modules of the knowledge base are simply populated with instances from the pertinent location and relevant source and then repopulated for the next line of enquiry. The only modification to the knowledge base is at the highest level - the "domain-specific" ontologies that hold the instance data and the hypothesis rules. In fact, the researcher is not required to predetermine precise hypothesis prior to data collection and the population of the knowledge base. The questions can be flexible, and may evolve as new data becomes available and/or as ideas emerge.

The level of granularity required, or the depth of the analysis, is multi-scalable and independent of location, reef type or standing stock. The functionality of entities within the reef ontology is maintained as applied logic systems, such as description logics and propositional logics. Where reasoning engines, such as Pellet [13] and inference engines such as Jess [14], are used to automatically infer latent connections or additional information, that is not explicitly asserted, from the data (Fig. 1).

C. The Workflow

The workflows within the SR map static and dynamic data to the hierarchy of ontologies (Fig. 1). Scientific workflow technologies and tools are adaptive software programs to capture complex analyses in a flow of which the data is taken through one analytical step after another [15].

The scientific workflow software chosen for the data flow implementation is the open-source Kepler system. Kepler is a user directed system that supports flexible data movement methods that include:

- A centralized 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 [16].

Each workflow step in the Kepler system is represented by "actors" that can provide access to a diverse range of data resources. For instance, the geographically distributed data repositories and sensor data required to populate the ontologies within the knowledge base [17].

Here, workflows are employed to automatically process the data and pass the results to the SR knowledge base via a series of workflow steps. The SR uses Protégé, which is a free open source ontology editor and knowledge base framework that offers extensive capabilities for reasoning and inference [18]. The workflows initiate Web services to collect both near real-time data from SEMAT and existing web available data from archives and repositories. Ultimately, the knowledge base could be filled with relevant data available from diverse sources, such as ocean temperature, salinity, nitrogen, pH and bathymetry information, as well as biological data such as coral, algae and fish stocks. Hypothesis questions or alerts can be posed of the data by questioning semantic correlation and

analysis with description logics and inference rules, such as finding the tipping point between a healthy reef and a dying reef or alerts such as coral spawning or algal blooms.

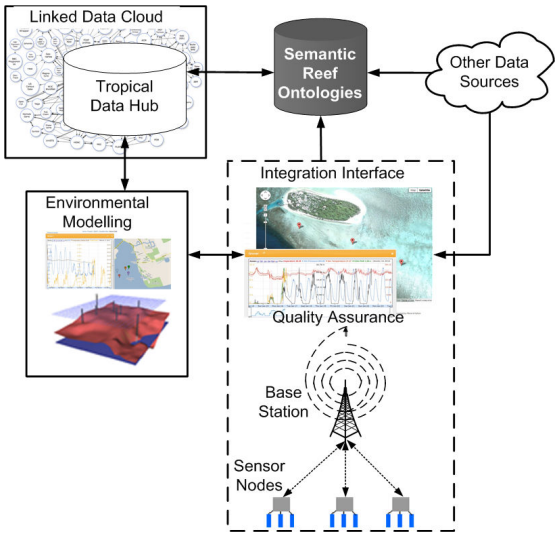

Fig. 2 The end-to-end framework for data collection, integration, analysis, curation and discovery

IV. SEMAT

Sensor networks for environmental monitoring are becoming an increasingly essential tool for planning and management of sensitive ecosystems. Fig. 2 presents the architecture for a SEMAT wireless sensor network. At the lowest level of the network is a sensor device. Each sensor device collects information (in the form of electrical stimuli) about some aspect of world. This can include features such as light, temperature, pressure, and video streams. Data is sent over a network to a base station, which then relays it back to an end user via the Internet.

Sensor networks can be deployed in extremely hostile environments (e.g., in the ocean, on a mountain, down a mineshaft, etc.). Sensors face significant challenges that are less of a concern in traditional network settings. Some of these challenges include limited power, prohibited storage and computational capacity, and restrictive communications. Much of this is dictated by the environment in which the network is deployed, the cost of the equipment, the level of performance and reliability required, and the technical skill of the personnel overseeing/maintaining the deployment.

SEMAT is a multidisciplinary project aimed towards making sensor networks a more accessible option for environmental planners and scientists [2]. The goal is to use a low-cost, commodity-based approach to design sensor networks. As the system is cheaper, more pervasive and comprehensive systems can be deployed that can dramatically increase the amount of data sensed. SEMAT also aims at removing the amount of technical expertise required to create, deploy, and maintain a sensor network. Essentially, the user can purchase the required sensors (from any vendor), plug them in, deploy the system, and then have the data seamlessly arrive on a laptop.

The upper SEMAT infrastructure adheres to *Sensor Web Enablement* (SWE) standards [19]. SWE is an initiative by the Open Geospatial Consortium [19] to make sensor networks interoperable with each other and to have data available online via the "Sensor Web". The Sensor Web facilitates data reuse in that the data is accessible to any interested party. Data also persists beyond the life of the deployment.

SEMAT's data acquisition method is based on *rsync* [20], which is a software application and network protocol for Unix-like and Windows system that synchronizes files and directories from one location to another. The software on the sensor nodes operates each sensor, reads the data, and transmits the data to the base station (via rsync). The sensor node manager is a *cron* job that transfers the data to the server. The cron process is alerted by the rsync that new data has arrived. The base station transmits the data back to the server via the Internet. The server then processes the data for web-visualization.

At the back-end of SEMAT is an interface for an environmental model. The environmental model can be for any scientific application. For example, a hydrodynamic model for predicting algal blooms, or a bush fire model showing how a fire will move through an area. One of the novel features of SEMAT is that the environmental model has two-way communication with the sensor network. That is, sensed data can update the model, and in return, the model can re-task or "tweak" the sensors. For example, if several sensors are indicating a phenomenon of interest, then the environmental can instruct the relevant sensors to increase their sensing intervals to capture more information. This can be achieved using the services provided by SWE.

While SEMAT is making progress on addressing sensor network hardware and technical challenges, the aforementioned discussion highlights the increasing importance of how to handle the data and the data fusion process. A cheaper, more pervasive system means that there is significantly more data being collected. Likewise, interaction with an environmental model and communication back to the sensors implies that not only more data will be collected, but there must be a smart fusion process in order to interpret the data. Extraneous and less important data has the potential to make the system ineffective and overloaded. Furthermore, SWE only defines a standardized structure for the data and metadata about sensors and readings. There is no method for a computer to comprehend the meaning behind, or the relationships between, different datasets. These factors exacerbate the "data deluge" problem – too much data, but not enough information.

Recent progress is being made in the areas of the *Semantic Sensor Web* (SSW) [4]. The SSW standards are in development and they define how sensor data is annotated with the contextual information required to enable semantic awareness [21]. SWE is then extended by the addition of Semantic Web [10] technologies to provide enhanced

descriptions and access to sensor data. Once semantically aware, logical paradigms can be applied via reasoning or inference engines to create knowledge.

A semantic inference engine can be used for two purposes in the SEMAT architecture:

1. Data filtering; and
2. Model refinement.

As the environmental model is being faced with a data deluge, semantically annotated data can be interpreted more quickly by an inference engine. The engine can then work out what data may be the most relevant and alert the environmental model. The environmental model can then focus on this data and determine whether it really is the most important to the study. If it is, then the SEMAT architecture allows for the environmental model to communicate back to the sensors to either re-task them, or give priority network resources to those currently sensing the most important data. The second role of the inference engine is in refining the environmental model. This can be done as an offline process or dynamically over time.

The issue addressed by this paper is how to enhance SEMAT with semantic technologies from the Semantic Reef framework to alleviate the data analysis phase. The following sections describe how existing tools and technologies can be used in unison to create a framework for a "smart" sensor network system.

V. The Tropical Data Hub - Exposure for Discovery and Cross-Disciplinary Data Fusion

The *Tropical Data Hub* (TDH) [6] is a platform to serve data sets related to tropical research from a single virtual location (Fig. 3). This *virtual research environment* is a platform for distributed data sharing and processing that enables researchers, managers and decision-makers to collaborate around the data. The TDH is to be a single entry point to hosted and remotely stored tropical data for researchers and managers and their data and information. The architecture facilitates the collection, discovery and curation of tropical data.

Biodiversity is not distributed evenly on Earth. A high proportion of the world's biodiversity is located in the tropics, including up to 80% of animal and plant species and 92% of world's coral reefs. Limitations exist for informed decision making on sustainable management and conservation of sensitive ecosystems. These limitations are often due to a lack of the correct data or lack of awareness of its existence, expertise, knowledge and/or research. Therefore, a cross-disciplinary approach is the only way to address the challenges of population growth, urbanisation, climate change, and biodiversity loss in the tropical regions.

The TDH philosophy is to span traditional 'vertical' research disciplines and enable 'horizontal' research. Specifically, vertical research is the traditional discipline and data-specific research paradigms are conceptual silos of concentrated research efforts. In contrast, horizontal research spans a cross-connect through disciplines, research methods, data resources and experimental techniques to enable synthesis of a diverse range of disciplines and data. The semantic layer of the TDH will enable data linkage ability.

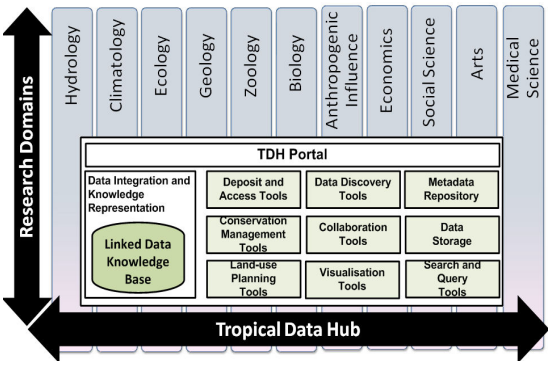

Fig. 3 The Tropical Data Hub portal is an open data collaborative model to support cross-disciplinary research in tropical regions

The TDH automates the data capture process by providing a standardized interface for sensor web accessible systems. This means that data can be obtained or indexed in near real-time from environmental sensor networks that comply with SWE standards (such as those offered by SEMAT). The TDH can be updated without user intervention, which will therefore reflect the latest in varied environmental data across regions. These environmental data combined with data collections from many other disciplines (e.g., economics, social science, biology, anthropology, etc.) are publicized by the TDH. The tropical data and metadata is then available and accessible for reuse and integration to enable the exploration of phenomena from a diversity of perspectives.

VI. Linking the Semantic Reef, SEMAT and the TDH for End-to-End Data Collection, Analysis, Curation and Discovery

This section describes how the SR, the SEMAT data portal and the TDH can be used to provide data fusion, hypothesis testing, and data discovery for SEMAT.

Fig. 2 shows the overall linkages of different systems to achieve end-to-end data collection, integration, Semantic analysis, curation and discovery. Here an instance of desktop data fusion is applied for the synthesis of disparate data. To exemplify, the knowledge base is populated with data from a real world SEMAT deployment, the *Australian Bureau of Meteorology* (BOM) and anthropogenic data from the *Australian Bureau of Statistics* (ABS). Inferences in the form of observational hypotheses about marine phenomena can then be conducted. The outcome of these inferences, which is the conversion of data to knowledge, will be maintained in the TDH.

During February 2011, the SEMAT system was deployed in Deception Bay, Queensland, Australia [2]. The deployment's goal was to study *Lymbya* algae outbreaks in the region and to examine how sediment from recent flood activity had affected

the water quality in the bay. Five locations were selected based on a hydrodynamic model (i.e., the environmental model) [22]. Samples were taken for pressure, light and temperature every 15 minutes. The following sections describe the proposed framework within the context of the Deception Bay study.

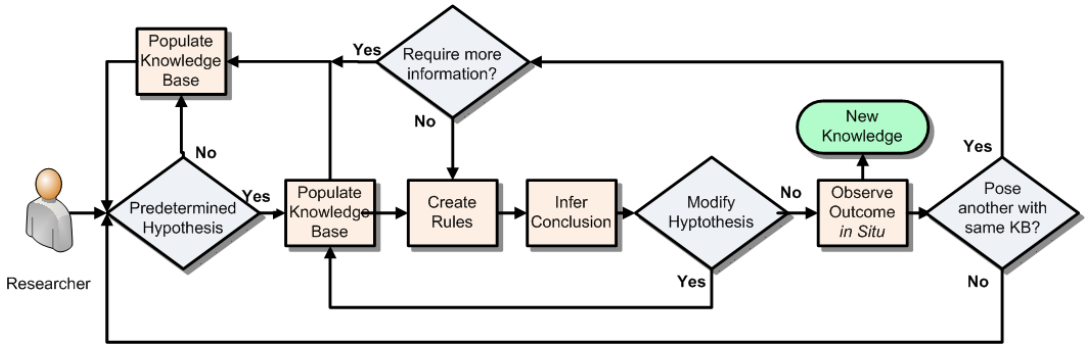

Fig. 4 Flexible observational hypothesis design flowchart

A. Data Fusion - The Semantic Reef and SEMAT Trial

The SEMAT data streams and/or the outcomes from the SEMAT environmental model are mapped to the SR knowledge base for inference and hypothesis testing. Disparate data was mapped to the knowledge base for inclusion in a sample hypothesis for the Deception Bay study to demonstrate ontology-based data fusion. The independent data sources included the SEMAT Deception Bay data, and data from BOM and the ABS.

In terms of the marine science goal of the Deception Bay deployment, questions were posed about algal blooms. The goal was to find the "tipping point" that leads to increased growth of algae or the areas where it is most likely to be present. Algae is a concern in Deception Bay as it is toxic, kills vital seagrass (the main source of food for the Bay's dugong population), scars the sea floor, and pollutes the beach and shore line once it matures and breaks free of its roots.

The hypotheses posed entail the cumulative combination of ecological factors that contribute to the tipping point from a balanced environment to an algal infestation. The SR is a tool to pose such hypotheses and automate inferences of the available data and, therefore, is an appropriate method to theorize about the cumulative factors of algal blooms. Once phenomena in the data are disclosed, *in situ* observations can be performed to confirm or negate the theory.

Meteorological and anthropogenic information of Deception Bay was mapped to the knowledge base. Specifically, sea temperature, photosynthetic active radiation (light), pressure (turbidity), rainfall and anthropogenic influences were included. The sea temperature, light and pressure data were extracted from the SEMAT portal. The rainfall data was supplied from the BOM Website. The anthropogenic factors from the ABS online database were mapped to the knowledge base and consisted of human population quantity and density for the Deception Bay region. The population data included the geographic figures and demographic breakdowns (age and gender) of the coastal regions. This allowed the scientist to pose questions theorising about the effects on algae blooms and seagrass as a result of the human coastal population density.

The SR's Kepler workflow imports and transforms the disparate data and prepares the knowledge base by populating the ontologies. The data from the three data sources are manipulated via the Kepler XPATH and Python actors. The XPATH expressions and queries extract the specific data values from the data streams and convert them to an array of values to be sent to the Python scripting actor. The Python actors implement simple scripts written to tag each value with a unique URI and send to the knowledge base to populate the appropriate ontology modules. On completion, the knowledge base is populated with one temporal instance for each measurement and linked to the rainfall and population quantity and density human influences for that location. Hypotheses could then be asked to examine water quality and observed algal growth for specified dates.

Inference rules were fashioned as observational hypotheses. Both logical inference and hypothetical based research entail syllogistic statements to draw a conclusion. Therefore, if any phenomena in the data were uncovered, *in situ* observations from the locations could confirm the hypothesis. Notably, 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 (Fig. 4) [1], [12].

Here inference rules were created to predict an algal bloom based on the temperature, PAR and turbidity for each location. The results could alert to a potential outbreak and validated by the predictions from the environmental model hydrodynamic model. The SR inference rules could be easily modified and the outcome ingested back to the environmental model to inform a different trajectory model.

B. SEMAT Web Portal for Integration to the Semantic Reef and TDH

This section describes the stages in which SEMAT data is collected from the sensor network deployed in the field, the visualization of the data, and the ability for the data to be

exported to other applications.

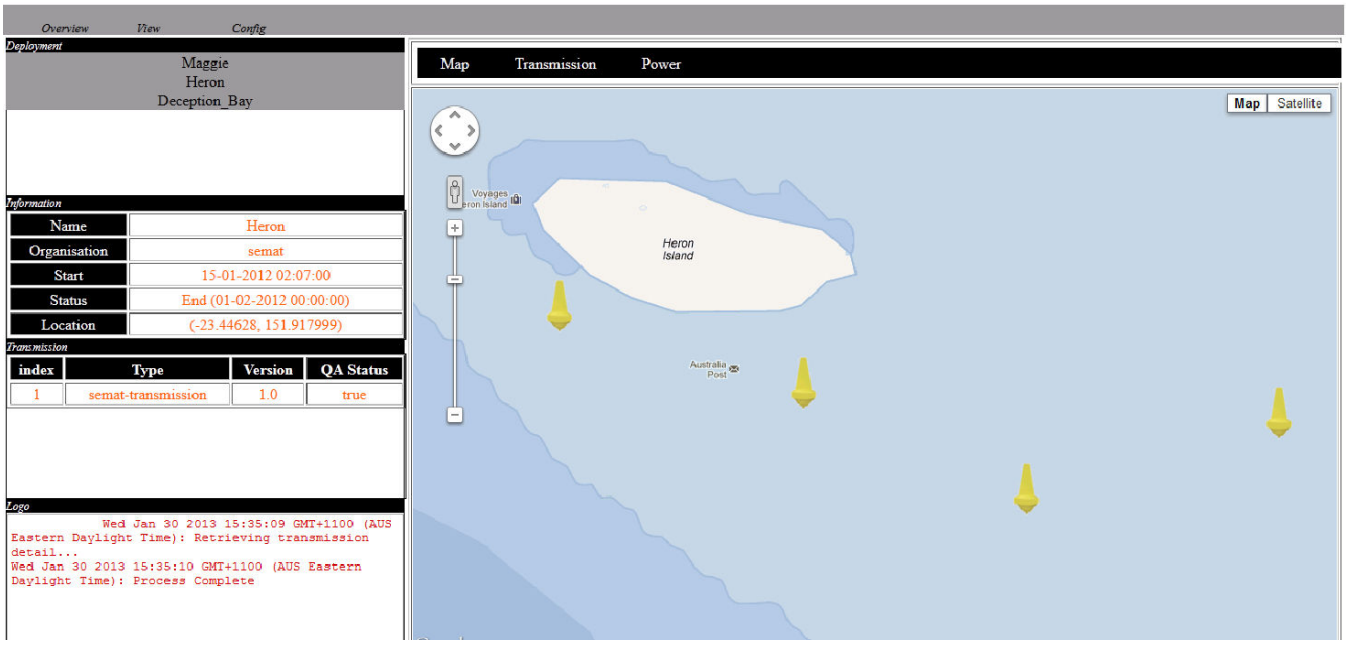

Fig. 5 The SEMAT data integration interface showing the Deception Bay sensor deployment

Fig. 5 shows the user interface for the SEMAT sensor network management system [23]. The core of the user interface is a Google Map, which shows the positions at which the buoys (sensor nodes) were deployed in the field. A user can click on a sensor node in the map and all of the associated information for that node is displayed (e.g., location/coordinates, sensor types, power status, alert status). A user can then graph any of the data from the sensor node by clicking on the appropriate data stream for a particular sensor. Data from several sensors can also be displayed as a mash-up. Calibration data are automatically combined with the sensor data if a calibration file is present. The user interface also provides other tools for monitoring the deployment's status, such as quality assurance techniques, alerts for offline sensor nodes, abnormal gaps in sensor data, and a real-time power management graph.

In addition to the graphical interface for visualizing the collected data, the system provides functionality for exporting the data to other applications. The simplest methods for exporting the data are allowing it to be saved as a text file or in a Microsoft Excel format. However, the system also has the ability to export data automatically in defined XML and RDF formats. This creates an interface for pushing marine data to the SR system for hypothesis testing and/or alerts.

The SEMAT portal's framework is a step towards efficient data availability for automating data extraction and data analysis. The portal combines data collection with accessibility, integration and automated formatting functionality. The formatting functionality handles much of the SR's Kepler workflow tasks, such as transforming data to triple form for ingestion to the knowledge base. Applications such as the SR or the TDH benefit from this open cross-platform approach to data accessibility as it minimizes manual manipulation and/or excessive processing.

C. Data Fusion - SEMAT with the Tropical Data Hub

The outcomes of the data integration efforts from the environmental model the SR, or a combination of both, will produce new datasets and collections. These datasets and their metadata can be deposited to the TDH to expose them for publication and ultimately further integration and/or reuse.

The TDH facilitates the collection, discovery and curation of tropical data. This data includes the remotely sensed data from initiatives such as SEMAT, knowledge generated by tools such as the SR and the varied qualitative and quantitative outcomes from other disciplines (e.g., economics, biology, the arts, social science, etc.). Once the SEMAT data is gathered and made available for discovery and analysis within the TDH it is open for synthesis with other linked cross-disciplinary data.

The data capture process includes the harvesting of metadata from the SEMAT portal [23] and the pointers to the physical location of the raw data. Then analysis tools available within the TDH for visualisation, statistical examination and collaboration can be applied. This process was undertaken with the Deception Bay deployment data [2].

D. Towards SEMAT and the Semantic Sensor Web

The ultimate goal is to have SEMAT form part of the SSW. This section briefly describes how the proposed framework for SEMAT will be altered to facilitate research and further development of the SSW.

As the SSW is still a new concept, it is envisaged that the

workflow phase of the SR will be simplified, or even made redundant, when the data is available in linked data form. That is, when the data is streamed from the sensors to a SOS server that supports semantic annotations through SemSOS [21] it will automatically be tagged with URIs at the source. This would take the place of the workflow process to prepare data for population to the knowledge base. The data is then fully prepared for integration to systems such as the SR.

The TDH is currently being enhanced to include a semantic layer for data linkage ability. The horizontal linkages across domains will only be possible using technologies being developed by the linked data movement (i.e., the Semantic Web) [24]. The internally and externally stored data will be published via the TDH in a form that is conducive to the linked data initiative. The data is then open for publication and connection with all other data sources exposed on the Semantic Web. Then, tools such as the SR have access to a greater range of cross-disciplinary data for hypothesis research and can contribute back to the TDH with the inferred outcomes.

Future work involves the implementation of this architecture to access sensor data from sensed terrestrial data. The system accesses data for flexible query and inference made possible through the hierarchy of light to heavy-weight ontologies. This system would be adaptable to other disciplinary fields such as terrestrial ecosystems. Simply by changing the ontologies in the knowledge base and mapping relevant data via the scientific workflows, new lines of enquiry can be investigated. For example, hypotheses based on data from the Daintree Rainforest at Cape Tribulation for research in environmental biodiversity loss and climate change effects.

VII. Conclusions

This paper has described how applied data fusion methods can take place at the desktop level through hypothesis-enabling tools such as the SR, at the data streaming level such as the environmental model or at the repository level such as the TDH. The result is a semantically-enhanced framework for SEMAT. We presented an example of how the framework operates using the SEMAT Deception Bay algae study.

The SR demonstrated that the linking of both sensed and static web available data can assist in the knowledge discovery phase. The architecture is flexible and can be modified to explore other domains. The system can infer knowledge from other disciplinary regimes such as terrestrial science simply by importing domain-specific ontologies to the knowledge base. Once the ontologies are populated with relevant data, for example, via the TDH, directly from the source by means of the workflows for external Web available data or via automated data feeds such as that form the SEMAT portal, questions can be posed specific to the line of enquiry.

The environmental model in SEMAT has two-way communication with the sensor network where the sensed data can update the model, and in return the model can re-task or alter parameters on the sensors. The models created from the environmental model can be used to inform a new hypothesis to be explored via the SR. Alternately, the inferred outcome from the SR can be used to inform a new environmental model. The results from the data collected and semantic inferences can then be seamlessly stored in the TDH and become publicly accessible for future reuse, curation, and knowledge discovery.

Future work involves the full incorporation of SEMAT with the SSW initiative. Then, the SR workflows can be easily modified to access the semantically annotated sensor data for integration in the knowledge base. Finally, once the data and metadata are added to the TDH repository, a wider scope of data integration is possible from the exposure and linkage to the Semantic Web.

Acknowledgment

This work is funded in part by the Australian National Data Service, Queensland Cyber Infrastructure Foundation, the University of Melbourne and the Queensland Government National and International Research Alliances Program. The authors would like to thank Professor Ron Johnstone from the University of Queensland, Professor Ian Atkinson from James Cook University and Yong Jin Lee for their assistance, advice and feedback.

References

[1] T. Myers and I. Atkinson, "Eco-informatics modelling via semantic inference," Information Systems, vol. 04, Available online 28 April 2012 2012.

[2] J. Trevathan, R. Johnstone, T. Chiffings, I. Atkinson, N. Bergmann, W. Read, S. Theiss, T. Myers, and T. Stevens, "SEMAT – The next generation of inexpensive marine environmental monitoring and measurement systems," Sensors, vol. 12, pp. 9711-9748, 2012.

[3] Gray, R. García-Castro, K. Kyzirakos, M. Karpathiotakis, J.-P. Calbimonte, K. Page, J. Sadler, A. Frazer, I. Galpin, A. Fernandes, N. Paton, O. Corcho, M. Koubarakis, D. De Roure, K. Martinez, and A. Gómez-Pérez, "A Semantically Enabled Service Architecture for Mashups over Streaming and Stored Data," in The Semanic Web: Research and Applications. vol. 6644, G. Antoniou, M. Grobelnik, E. Simperl, B. Parsia, D. Plexousakis, P. De Leenheer, and J. Pan, Eds.: Springer Berlin / Heidelberg, 2011, pp. 300-314-314.

[4] Sheth, C. Henson, and S. S. Sahoo, "Semantic Sensor Web," IEEE Internet Computing, vol. 12, pp. 78-83, 2008.

[5] D. Le-Phuoc, H. N. M. Quoc, J. X. Parreira, and M. Hauswirth, "The Linked Sensor Middleware: Connecting the real world and the Semantic Web," in Presented in the Semantic Web Challenge 2011, 10th International Semantic Web Conference (ISWC 2011), Bonn, Germany, 2011.

[6] T. Myers, J. Trevathan, and I. Atkinson, "The Tropical Data Hub: A Virtual Research Environment for tropical science knowledge and discovery," International Journal of Environmental, Cultural, Economic and Social Sustainability, January 2012, to be published.

[7] 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.

[8] Goble, O. Corcho, P. Alper, and D. De Roure, "e-Science and the Semantic Web: a symbiotic relationship," in Proceedings from the 9th International Conference in Discovery Science (DS 2006), Barcelona, Spain, 2006, pp. 1-12.

[9] A. Henson, H. Neuhaus, A. P. Sheth, K. Thirunarayan, and R. Buyya, "An ontological representation of time series observations on the Semantic Sensor Web," in 1st International Workshop on the Semantic Sensor Web (SemSensWeb 2009), Crete, Greece, 2009, pp. 79-94.

[10] G. Antoniou and F. van Harmelen, A Semantic Web primer Cambridge, MA, USA: The MIT Press, 2004.
[11] N. Guarino, "Understanding, building and using ontologies," International Journal of Human-Computer Studies, vol. 46, pp. 293-310, 1997 1997.
[12] 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, vol. 24, pp. 77-101, 2010.
[13] Mindswap. (2007) Pellet: the open source OWL DL reasoner [Online]. Available: http://clarkparsia.com/pellet
[14] (2009) Jess - The rule engine for the Java platform [Online]. Available: http://herzberg.ca.sandia.gov/jess/
[15] Altintas, C. Berkley, E. Jaeger, M. Jones, B. Ludäscher, and S. Mock, "Kepler: an extensible system for design and execution of scientific workflows," in Proceedings of the 16th International Conference on Scientific and Statistical Database Management (SSDBM 04) Santorini Island, Greece: IEEE, 2004, pp. pp. 423- 424.
[16] Yu and R. Buyya, "A taxonomy of scientific workflow systems for grid computing," SIGMOD Rec., vol. 34, pp. 44-49, 2005.
[17] B. Ludäscher, I. Altintas, C. Berkley, D. Higgins, E. Jaeger, M. Jones, E. A. Lee, J. Tao, and Y. Zhao, "Scientific workflow management and the Kepler system," Concurrency and Computation: Practice and Experience, vol. 18, pp. 1039-1065, 2006.
[18] Protégé. (2009) The ontology editor and knowledge acquisition system [Online]. Available: http://protege.stanford.edu/
[19] M. Botts, G. Percivall, C. Reed, and J. Davidson, "OGC® Sensor Web Enablement: Overview and High Level Architecture," in GeoSensor Networks. vol. 4540/2008 Berlin: Springer Berlin / Heidelberg, 2008, pp. 175-190.
[20] Samba. (2012) rsync [Online]. Available: http://www.samba.org/ftp/rsync/rsync.html
[21] A. Henson, J. K. Pschorr, A. P. Sheth, and K. Thirunarayan, "SemSOS: Semantic sensor observation service," in Proceedings of the 2009 International Symposium on Collaborative Technologies and Systems (CTS '09) Washington, DC, USA: IEEE Computer Society, 2009.
[22] G. S. Hamilton, F. Fielding, A. W. Chiffings, B. T. Hart, R. W. Johnstone, and K. Mengersen, "Investigating the Use of a Bayesian Network to Model the Risk of Lyngbya majuscula Bloom Initiation in Deception Bay, Queensland, Australia," Human and Ecological Risk Assessment, vol. 13, pp. 1271-1287, 2012/06/04 2007.
[23] Y. J. Lee, J. Trevathan, I. Atkinson, W. Read, T. Myers, and R. Johnstone, "The Evolution of the SEMAT Sensor Network Management System," in Proceedings of the 7th International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP 2011), Adelaide, Australia, 2011, pp. 229-234.
[24] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," Int. J. Semantic Web Inf. Syst., vol. 5, pp. 1-22, 2009.