Deriving an Australian Marine Ontology from Existing Ontological Models: a practical evaluation

ABSTRACT

Domain and purposive ontologies provide explicit and formal definitions of real-world concepts. They are indispensable components of any system whose objective is to support the exchange and semantic integration of data and information. The reuse of existing formalised ontologies is often encouraged to reduce development overheads and increase semantic interoperability between data service providers. This paper presents some selection criteria that are deemed important for choosing potentially reusable ontologies when dealing with marine science data. An Australian marine science case study involving the Integrated Marine Observing System (IMOS) was used to investigate ontological requirements. Climate Science Modelling Language (CSML) and Observation and Measurement (O&M) were the two main ontologies under consideration for re-use. By adapting an ontology construction methodology, for use instead as an evaluation tool and by using sample data from the IMOS project, CSML and O&M were evaluated for their suitability to support the exchange of data within the IMOS information infrastructure. As a result of the evaluation exercise a merged CSML and O&M ontological model was deemed preferable to using either ontology on its own.

Keywords: Marine, Ontology, Geography Mark-up Language (GML), CSML, O&M, Web Services, Selection, Evaluation.

1   INTRODUCTION

An ontology is an explicit and agreed formal specification of a conceptualisation (Gruber, 1993). In simple terms it is a method of describing the domain of discourse for a particular community, or field of endeavour. A domain ontology makes explicit the type of concepts, or real-world objects that are encountered in a discipline and will generally include information about concept properties, concept relationships and value restrictions on both properties and concepts (Gomez-Perez *et al*, 1996). As well as offering a mechanism for a community of interest to define a shared terminology, ontologies and machine accessible metadata (data about data) are fundamental to facilitating machine mediated transactions and processing on the World Wide Web (WWW). But building ontologies is a highly resource intensive task (Lozana-Tello & Gomez-Perez, 2004) and so many authors are advocating re-using and extending existing ontologies to reduce the development overhead and as a corollary increase the level of interoperability between ontologies and the systems that rely upon them (Uschold, 2005). To re-use ontologies, however, we must be able to evaluate those that exist and then have criteria for selecting the ones that are most appropriate for our needs and operating contexts.

Several formal methodologies have emerged in recent times designed to assist users to make comparisons between ontologies (Sabou *et al*, 2006). However, many of these techniques do not appear particularly user-friendly to non-academics, are often labour intensive to apply (Hartmann *et al*, 2005) and it is doubtful that they are practical to use in most real-world scenarios. Kalfoglou *et al* (2004) have argued that ontologies not built and vetted by domain experts, but by academic computer scientists, are usually rejected. As a result, Australia's Integrated Marine Observing System (IMOS) information infrastructure was used in this research to case study and explore ontology selection and evaluation issues from a practical and domain-centric standpoint. The IMOS data management community had already determined that they would be sharing data using a service-oriented-architecture (SOA) and would be heavily reliant on web mapping technologies, particularly those conforming to the Open GIS Consortium (OGC) standards (http://www.opengeospatial.org/standards). Amongst other constraints this implied that the primary language for conveying IMOS data exchange ontologies would have to be the Geography Mark-up Language [GML] (Lake *et al*, 2004), the standard mandated by the OGC for web mapping services.

An evaluation exercise was undertaken of two existing ontological models, Climate Science Modelling Language [CSML] (Woolf *et al*, 2005) and Observation and Measurement Schema [O&M] (Cox, 2006). Both models are internationally well known, but not yet extensively exercised through implementation. Using the development of Australia's Integrated Marine Observing System information infrastructure as a case study for the operational environment in which the existing ontology is to be used, it became apparent that a combination of both models perhaps better suited IMOS requirements, than either model could individually. A merged model was developed with specialisations and extensions made to suit the IMOS context. Because CSML was originally developed for application within the physical sciences, its treatment of biological phenomena had to be supplemented by incorporating ontological concepts from emerging biological data exchange standards. The merged model's general utility in accommodating the encapsulation of IMOS datasets was then tested using sample datasets from biological and oceanographic applications.

This latter exercise lead to the development of encoding "patterns" which it is anticipated could readily be applied to other data types that will be exchanged and processed within the IMOS infrastructure.

The remainder of this paper first describes the methodology used to evaluate CSML and O&M. The strengths and weaknesses of CSML and O&M are then presented and a merged model is postulated which includes specialised and extended concepts to cater for IMOS data requirements. Two typical IMOS datasets are then encoded in GML instance documents and the rationale for the patterns chosen are discussed. In conclusion, further and more rigorously conducted research is suggested that is designed to test the broader applicability of the evaluation criteria considered practically important by the IMOS community. Additional exploration and fine-tuning of the GML encoding patterns is also warranted using different types of sample data.

2   METHODOLOGY

The IMOS network, which provides the case study in this paper, comprises remote and in-situ measurements of parameters such as: water column temperature, salinity and nutrients; currents and ocean productivity. Biotic sampling will involve the use of continuous plankton recorders, acoustic tagging of fish and various reef monitoring activities. Sample datasets used for testing ontologies are those typically collected by this network.

2.1   Ontology Evaluation Process

An analysis of the literature on ontological evaluation methods showed that this is still very much an emerging field of enquiry. Many of the approaches being advocated (Guarino & Welty, 2004; Corcho *et al*, 2004; Lozano-Tello & Gomez-Perez, 2004; Hartmann *et al*, 2005; Spyns, 2005) did not seem entirely applicable to the practical evaluation task faced in this case study and often dealt with one or two evaluation dimensions (across categories of function, structure and usability, but not all three), rather than providing for a comprehensive and balanced treatment. Methods such as those of Gangemi *et al* (2005), which are multi-dimensional in their approach, still rely on the use of quite complex measures and if anything, the measurement burden is compounded in this approach. What the IMOS community required was a practical evaluation process that: (a) could be conducted by a domain expert rather than requiring ontological engineering skills, (b) had a predominantly qualitative rather than a quantitative evaluation model to provide for rapid assessment, and (c) tested only matters of perceived significance. The steps therefore pursued by the author to evaluate some candidate ontologies followed an ontology building method outlined by Annamalai & Sterling (2003), but with small modifications made to better suit an assessment process context. These steps include:

   a) Specifying the purpose and uses of the ontology through the construction of use cases and consultation with the IMOS data management community,
   b) Selecting sample data streams from within IMOS to help develop a purposive conceptual model and devise focussed competency questions,
   c) Sketching an outline of the general concepts and their properties (for the purposive model and the selected sample data) rather than trying to exhaustively model all of the concepts and their relations,
   d) Developing competency questions as a benchmark to test the ontology,
   e) Identifying potential reusable domain ontologies,
   f) Comparing the existing domain ontologies with the general purposive model, noting any deficiencies,
   g) Constructing unsupported portions of the purposive model,
   h) Testing the model against competency questions until satisfied (repeat g & h),
   i) Extracting reusable, generalisable concepts and/or noting fruitful modifications that could be made to existing ontologies, and
   j) Evaluating which ontologies best meet requirements.

2.2   Ontology Evaluation Measures

In this case study, evaluation is performed at various stages in the process and measures are qualitative. The process described is relatively rapid. In most cases the measures are subjective but the intent of the exercise is simply to make an informed judgement regarding an ontology's fitness for use. The process and measures used are designed to provide a structured approach to the assessment so that the virtues and detractions of an assessed ontology can be reasonably argued, explained and then shared with others. The measures selected are significant to the IMOS community and were derived through various consultations with IMOS data management experts and were then dimensionalised according to Gangemi *et al* (2005) but with two dimensions added, i.e "maintenance" and "governance". These two additional dimensions could have been included under the "usability" category described by Gangemi *et al*, but they were

considered to be of sufficient importance to warrant being labelled a dimension in their own right. The evaluation measures used in this study are explained in Table 1 of the Appendix.

2.3   Use Cases, Competency Questions and GML

The broad requirements of the IMOS information infrastructure were scoped by the IMOS data management community through a series of face-to-face meetings, a commissioned data scoping exercise (Bainbridge, 2007) and a standards workshop in March 2007. Informed by these expert consultations, the author constructed three high-level use-cases that characterised the type of functionality that the IMOS infrastructure should support. Only the first two use-cases are presented in this paper and were used to guide the evaluation exercise (see Table 2 in the Appendix). The third use-case involving data selection and manipulation using inference techniques was considered beyond the scope of the first IMOS infrastructure release and is not achievable using GML alone as the ontology encoding language. GML lacks the constructs required for formal logic processing which are necessary to support inferencing. A sample of typical competency questions constructed to fit with these high level use-cases is shown in Table 3 of the Appendix. These competency questions were used primarily to inform what information must be present in the ontology and how this information must be encoded in order for it to be readily accessed.

Note that because IMOS is anticipating using GML as the main ontology encoding language, a "concept" in GML is generally referred to as a 'feature", which like a concept in traditional ontology description languages such as RDF (http://www.w3.org/RDF/) or Owl (http://www.w3.org/TR/owl-features/), is an abstraction of a real-world phenomena. While features are the focal elements of most GML documents, the language is actually comprised of "objects" which include features, geometries, coordinate reference systems and styles. In GML, a feature instance is an identifiable object in the world, or the digital representation of it and features are classified into feature types on the basis of common sets of characteristics or properties (e.g. attributes, associations and relationships, operations and behaviours). A GML "application schema" describes the logical structure and semantic content of a dataset using a feature-based model (for further details see the Generalised Feature Model as described by ISO 19109). This model has also been named the object-property model (similar to the RDF subject-property model). Most GML application schema are augmented by the use of "dictionaries", i.e. external instance documents that define things like coordinate reference systems, units of measure, value types and temporal reference systems.

2.4   Tools and Test Data

Throughout this paper conceptual modelling is presented using the Unified Modelling Language (UML - http://www.uml.org/). GML instance documents and schemas were evaluated and developed using XMLSPY V.2006(sp2). The two sample datasets chosen to test the ontological models were data captured through the use of Conductivity Temperature Depth (CTD) and Continuous Plankton Recorder instruments. A typical database schema for CTD data was acquired from the CSIRO Marine and Atmospheric Research Data Centre and a CPR database schema was supplied by the Australian Antarctic Data Centre (see ctd_cluster.jpg and CPRTows.doc at http://aadc-maps.aad.gov.au/imos/). These schemas provided a good benchmark with which to test the ontological models in terms of their ability to model the concepts and attributes required for defining these specific data-streams.

Aspects of "usability" were explored using DEEGREE V2.0 (see http://www.deegree.org/) open source spatial web service middleware and XSLT scripting (see http://www.w3.org/TR/xslt).

3   EVALUATION OF CSML & O&M

3.1   CSML Overview

CSML (Woolf *et al*, 2005) is an integral component of the UK NERC Data Grid Project (Lawrence *et al*, 2003). This ontology defines its high level concepts (or feature types) primarily on the basis of geometric and topologic structure and not on the semantics of the observable or the measured property, as do other observation-centric ontologies. If two features are structurally identical, i.e. say they both record values at certain depths vertically through the water column at a fixed location in space, even though the physical 'phenomenon' recorded at these depths is different, then they are modelled in CSML as the same feature type. In this particular example the feature type would be a "ProfileFeature". CSML (V2) uses 13 Feature Types and all of them are expressed as Coverages – which are logical implementations of the ISO 19123 CV_DiscreteCoverage coverage class. Every coverage has a spatiotemporal domain called a "domainSet" and an associated parameterised value range called a "rangeSet" – see Figure 7 in Woolf (2007). More simply put, a coverage can be thought of as a set of (geometry, value) pairs (Lake *et al*, 2004).

CSML is a GML application schema with the twist that feature types and their instances can either be encoded and accessed in-line within a CSML "Feature Collection", or the feature types can be encoded in-line but the actual instance data is stored in a legacy (non GML) file-format. Data stored in legacy formats are then accessed from the GML instance document using an X-Link mechanism - see Figure 8 in Woolf (2007). X-Link is an XML language for creating and describing links between resources –http://www.w3.org/TR/xlink/). The rationale for this approach is that the GML document functions to provide the requisite "feature" metadata for interoperability purposes, whilst allowing data custodians who have already invested heavily in using alternate formats and encodings to exploit their existing data using the GML (and ultimately web services) framework. Many of these legacy files use compact (binary) encoding formats that streamline the transport of highly voluminous data (e.g. satellite and model derived data) and are bound to disciplinary-specific, data manipulation software and tools. While this is a desirable characteristic of CSML, its limited capacity to encode supporting information such as how a feature was captured or how it may have been processed before being transmitted is a drawback given IMOS requirements.

Because CSML was devised primarily for the physical ocean and atmsopheric disciplines its application for encoding biological observations is somewhat untested. Bennet *et al* (2006) have summarised a range of limitations CSML has in appropriately modeling ecological data. Some of these limitations included inadequate support for species names in the current CSML phenomena dictionaries, inability to encode important sampling related metadata and difficulty in modelling quadrat style data efficiently.

3.2   O & M Overview

In contrast to CSML, the Observation and Measurement Schema models its universe of discourse from a metadata perspective – allowing more scope for inclusion of information of interest to the IMOS data community users. Under this model (Figure 1), an observation is an event whose result is an estimate of the value of some property of a "feature-of-interest" obtained using a specified procedure (Cox, 2006). While Cox describes a range of utility feature types for the "feature-of-interest" these types don't currently offer the same degree of encoding concordance, as CSML does, with how many physical scientists are currently modelling their observational data – which is particularly "grid" or "coverage" centric. Woolf and Cox are now actively working to ensure harmonisation of the "sampling features" of the two ontological models (A. Woolf, personal communication, 4 August, 2007) but evaluation performed in this paper was undertaken using the existing published specifications for CSML and O&M
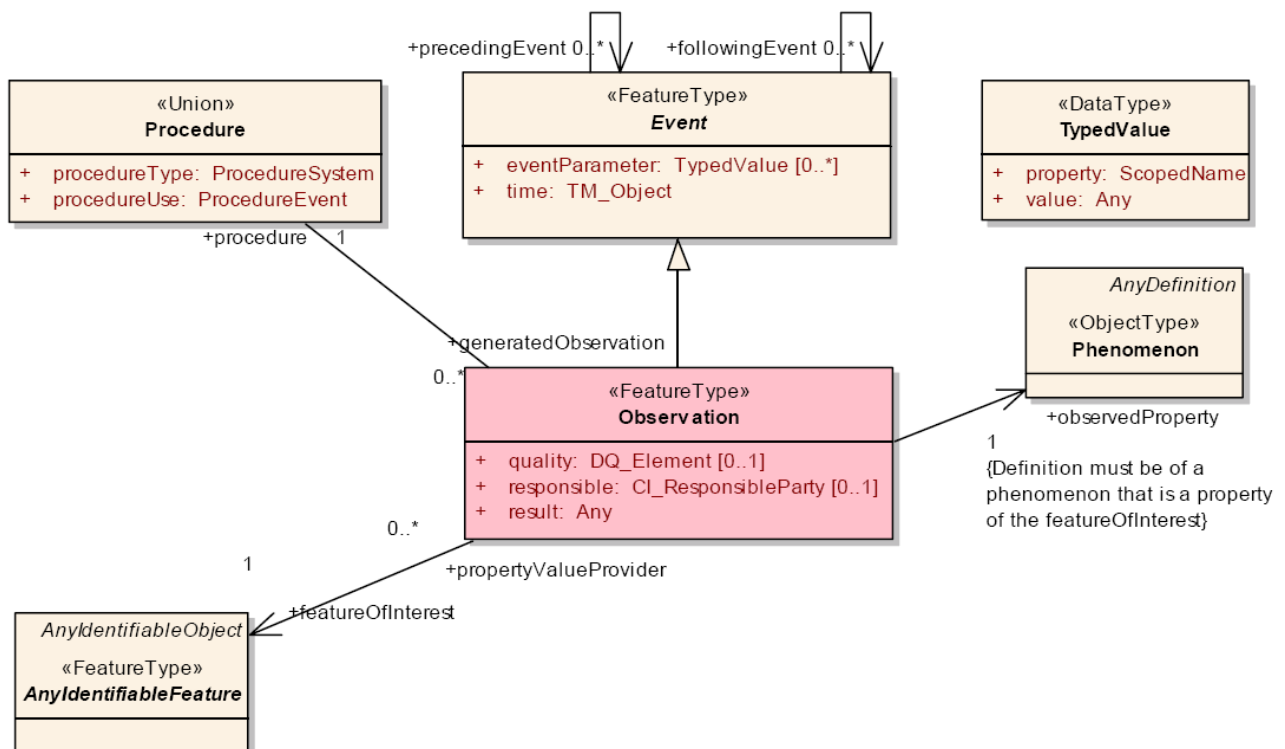

Fig. 1. O&M Feature Relationships (Cox, 2006)

3.3   IMOS Merged Purposive Model

In the current version of CSML, a CSML "Feature Type" can already be mapped directly to an O&M "FeatureOfInterest" and the CSML "Coverage" can be considered to represent an O&M "result". Both O&M and CSML share a phenomenon object class but from a CSML perspective this object represents a "parameter" and from an O&M perspective it represents an "observedProperty" (see Figure 2).
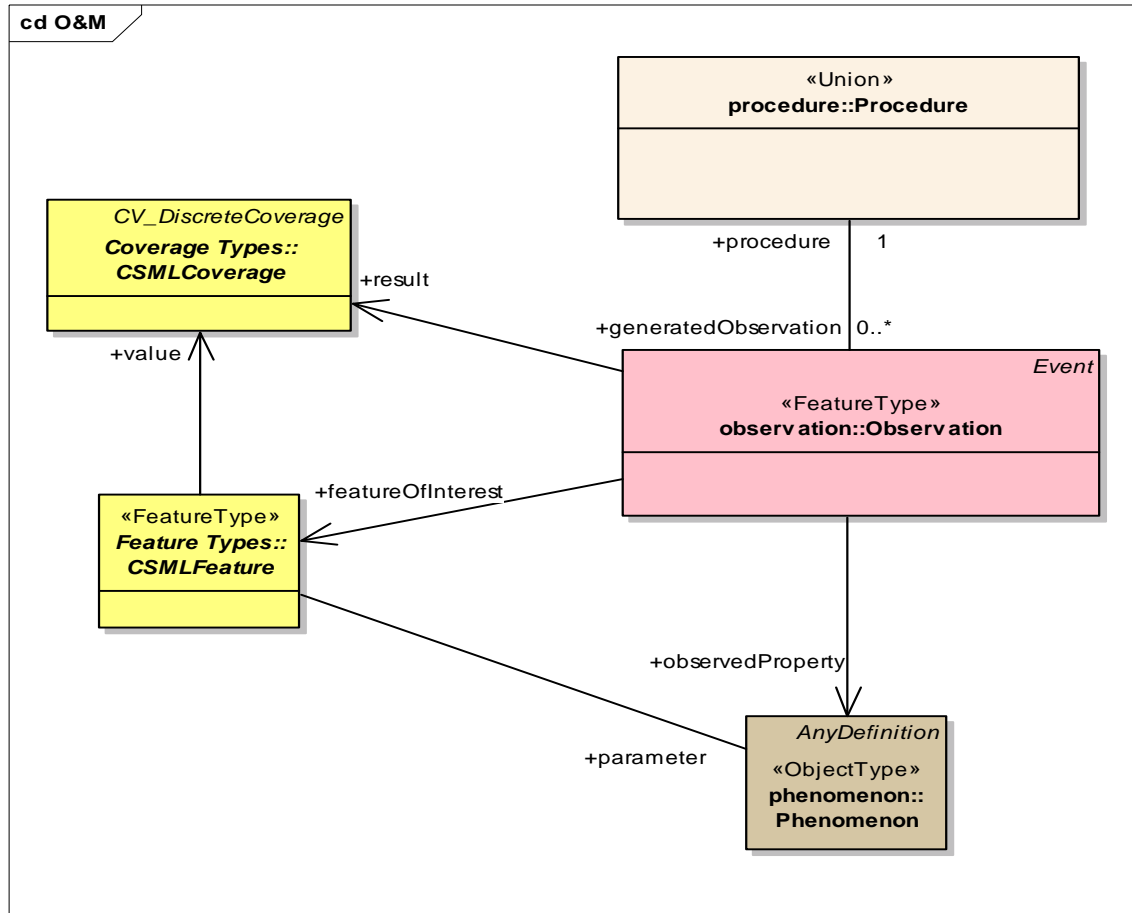

Fig. 2 – CSML & O&M Relationships (Woolf, 2007)

Through devising the competency questions and by examining the sample data schemas, several missing features (concepts) emerged that were considered important from an IMOS application ontology perspective and which were also considered generalisable and possibly worthy of elevation to a domain ontology. For example, all IMOS related data will be generated (or owned) by a "Project" and conceptually can be thought of as being associated with some type of specialised O&M "Event" called a "Deployment". A "Deployment" feature can be associated with (or hosted by) a "Deployment Platform" (e.g. vessel, buoy, satellite, float). IMOS observations may also be associated with an "InstrumentProcedure", a "ProcessingMethod" or various packaged combinations of both. These latter objects are specialised from the O&M "ProcedureType". The resulting merged model of specialised CSML and O&M concepts is depicted in Figure 3 and a GML Schema document was developed to accommodate the new and specialised features and objects. Re-used O&M and CSML components were accessed by importing the appropriate schema documents.
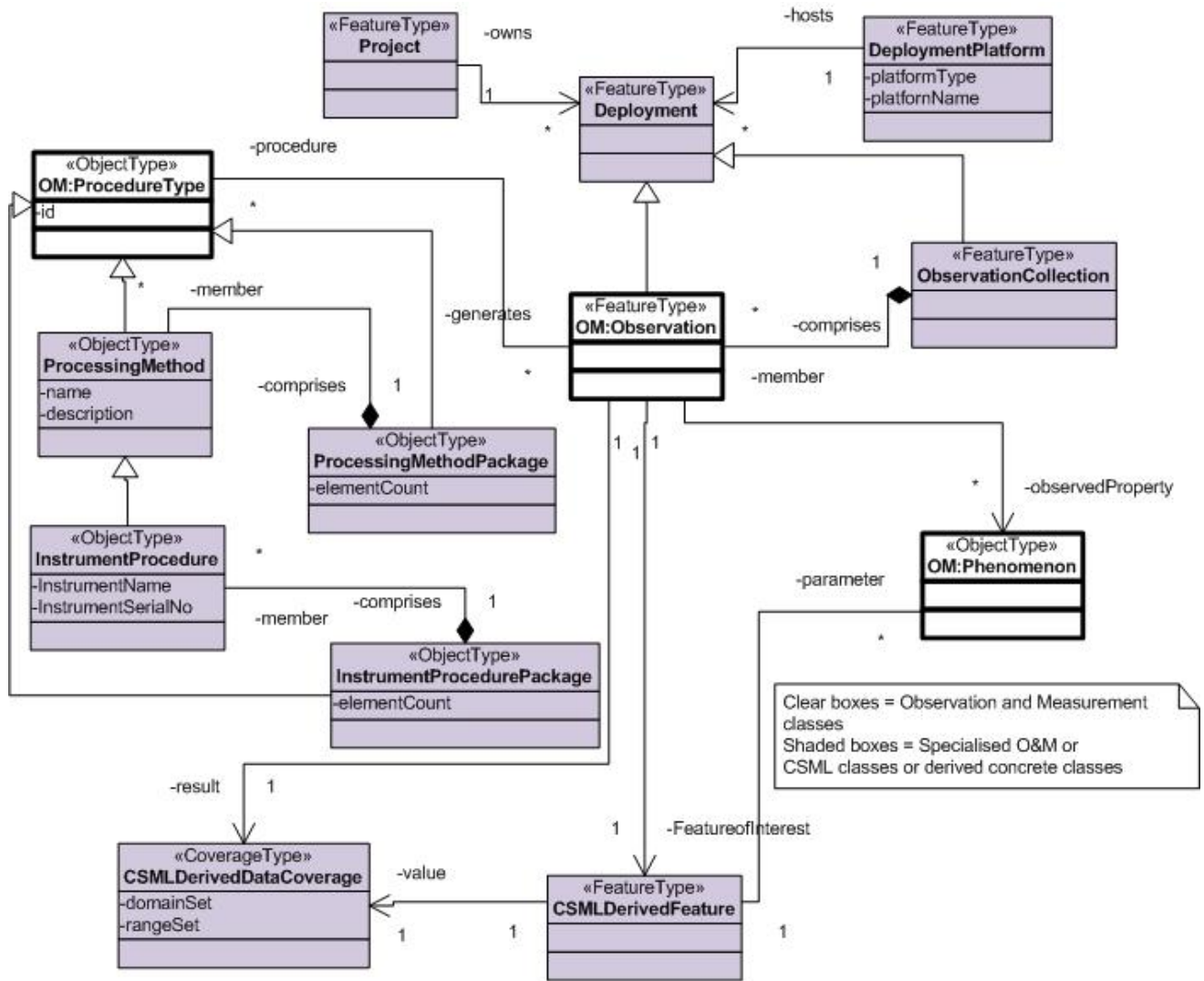
Fig. 3 IMOS Merged CSML/O&M Merged Model

One problem that is raised by merging the existing O&M and CSML schemas in this manner is that CSML includes the property "parameter" as a mandatory component in its model and O&M includes the property "ObservedProperty" also as a mandatory component. This means that within the merged model, "phenomenon" information has to be expressed twice as the value of two different, but otherwise identical properties, causing unnecessary redundancy in the schema.

The merged model also has to be wrapped in a modified CSML "Dataset" schema because features, units of measure, coordinate reference systems etc must now be embedded in a modified O&M "ObservationCollection", rather than in a CSML "FeatureCollection".

3.4   Sample Data Mapping

After creating the IMOS merged model, it was trialled and fine-tuned using CTD and CPR sample data. A brief description of these data streams and their encoding follows. Only in-line encodings were developed to explore the model.

3.4.1   CTD Data

A CTD instrument is a recording device which is lowered over the side of a vessel. It consists of a variety of sensors for recording data on physical phenomena as the instrument is lowered and raised through the water column. The instrument also has a number of bottles attached which collect water samples at specific depths. The bottle water is generally analysed for certain water chemistry phenomena and is also used to calibrate the temperature and salinity sensors. Each CTD deployment is conducted at a site (or station) along the vessel's track. The vessel is kept on station during the deployment but is still subject to movement because of currents and waves. Conceptually, the observations made can be

considered to be taken at a point location but in reality the start and end sampling positions may be slightly different from each other. The instrument travels down through the water column with depths derived from measurements taken by a pressure sensor. Each deployment of the CTD instrument generates a CTD observation i.e. a profile of phenomena vs depth pairs at a nominal point in the ocean. In the IMOS merged model the actual data values are represented through a specialised CSML ProfileCoverage, where the depths are encoded as part of the domain and each phenomenon value is encoded as a block in an array. See Figure 4 for the IMOS CTD data sample encoding pattern.

When fitting the sample data to the CSML (V2) specifications it became apparent that the properties described for representing the CTD data would lead to a very verbose encoding (see sample CSML CTD encoding prepared by Woolf using CSML(V2) specifications at http://aadc-maps.aad.gov.au/imos/CSMLExample_CTDProfile.xml). Alerted to this problem, Woolf then provided alternate encodings that significantly improved the encoding efficiency. One of these alternate encodings for in-line data representation was used in generating the IMOS sample CTD GML instance document viewable at http://aadc-maps.aad.gov.au/imos/CTDInstance.xml.

A further complication was that the IMOS data will need to carry quality control (QC) flags for each depth/phenomena value pair in every observation. While GML has provided for quality statements to be associated with entire features (e.g. through the "metadataProperty" which is an optional property) there is no facility in the present Abstract GML Coverage model to attach quality flags to values in the Coverage "rangeSet" property. This is a fairly significant limitation of GML in that GML properties cannot currently carry qualification metadata. This is a deficiency of GML and not the conceptual models of either O&M or CSML. A work-around to this problem was formulated by including quality flags in the phenomenon dictionary and correspondingly recording the QC flags as values in their own right in the "rangeSet" property. Ideally, however, these QC flags need to be addressed in a different manner and should be able to be directly associated with the parameter values that they pertain to in the "rangeSet" property. At present, this association is implied only through the naming convention used for the flag (e.g. Temperature_QC_Flag). This is a misuse of the phenomenon dictionary because its purpose is to list and define measured or observed phenomena, not the qualification attributes pertaining to these phenomena. Quality flags also needed to be associated with time and position data. Because GML currently makes no allowance for this, specific QC properties were defined that draw upon code-lists to carry this information and were included in the specialised IMOS observation feature.

```
<IMOS:Dataset>
{include and import statements for re-using other GML-based ontology elements}
        <IMOS:ObservationCollection>  {encompasses all deployments and observation members}
        {properties describing bounding box, time, etc for all deployments}
                <IMOS:Observation> {first observation member}
                {project, platform, location, time, procedure, composite phenomena properties}
                        <IMOS:FeatureofInterest> {CTD feature}
                        {CTD specific properties encoded in a specialised CSML Profile Feature}
                        </IMOS:FeatureofInterest>
                        <IMOS:Result> {CTD data values}
                                <IMOS:domainSet>
                                {CTD domain properties and values encoded in a specialised CSML Profile Coverage}
                                </IMOS:domainSet>
                                <IMOS:rangeSet>
                                {CTD range properties and values encoded in a specialised CSML Profile Coverage}
                                </IMOS:rangeSet>
                        </IMOS:Result>
                </IMOS:Observation>
                <IMOS:Observation> {second observation member – pattern repeats}
                        ……………
                        ……………
                </IMOS:Observation>
        </IMOS:ObservationCollection>
</IMOS:Dataset>
```

Fig 4. IMOS CTD Encoding Pattern

Some marine datastreams can be composed of observations generated from real, or conceptual compound instruments which can produce measures for multiple phenomena. These individual phenomenon measurements are often processed post collection using different processing techniques. A CTD is an example of a compound instrument which generates several types of measures, which can either be processed together or separately. In an encoded dataset it is useful and sometimes important to be able to distinguish which types of instruments have been used and what associated processing methods have been applied in generating particular phenomena. This can facilitate the discovery and data manipulation process in machine-to-machine transactions and assist with user evaluation of a dataset's fitness for use. The illustrative specialisations provided in the O&M specification, in general terms, met IMOS purposive requirements. The

"ProcedurePackage" concept was, however, adapted so that it was possible to structurally associate one or more processing methods directly with a specific instrument, if so desired. The rationale for this modification was based on an assumption that the encoded dataset would be easier to parse if there was a nested relationship between instruments and their data processing methods. In the example provided in the O&M specification, the "ProcedurePackage" can contain any number of instruments and any number of separate processing methods in a "om:CalculationProcedures" object, all of which can be serialised in any order and so it would be difficult to identify which processing methods should be associated with any given instrument(s).

The specialised O&M "Procedure Type" objects and their relationships, as shown in the UML model in Figure 3, have been devised to permit better levels of aggregation, where desirable, for expressing procedural information in IMOS data streams.

3.4.2   CPR Data

Plankton are sampled by towing a CPR instrument behind a vessel. This instrument contains a silk cloth wound around a small coil which unwinds at a known rate and is exposed to seawater before being wound back around another coil, which is then immersed in preservative. Planktonic biota adhere to the cloth as it is exposed to the seawater. The cloth is returned to the laboratory, cut into equivalent length strips representing a known number of nautical miles travelled and the adhered biota is identified and counted. While the vessel is underway other instruments record physical phenomena such as temperature, salinity and dissolved oxygen at the sea surface.

CPR data is also encoded using the pattern shown in Figure 4, except that an "ObservationCollection" can now be considered equivalent to a "tow" for CPR data (i.e. one continuous recording of a CPR instrument during a voyage tracing the vessel's track) and each "Observation" equates to a "tow segment" (i.e. a numbered segment of cut cloth with entrapped biota).

A problem was also encountered in fitting the sample data according to the CSML V2 specification. The most suitable CSML Feature for a CPR deployment appeared to be a "Trajectory Feature", represented by a distribution of irregularly timed observations embedded within a 2-, 3-, or 4-d compound spatiotemporal coordinate reference system (see CSML (V2), pp 26 Table 2). However, this CSML Feature type did not permit a good modelling of the "tow segment" domains. As a consequence, a standard GML MultiCurveCoverage was used instead. This was a far better match for modelling the data as it was then possible to encode each segment's domain as a line with as many points as needed to appropriately depict the coordinates of the ships track during collection of a particular CPR segment. The domain of the feature-of-interest in each "Observation" in the CPR case, as opposed to the CTD example, is not a vertical coordinate reference system (i.e. pressures equating to depths) but a line segment (tracing part of the ship's track) as defined by spatial coordinates. In a MultiCurveCoverage, the domain is partitioned into a collection of curves comprising of gml:MultiCurve data types. The coverage function then maps each curve in the collection to a value in the rangeSet. Woolf (personal communication, 7 June, 2007) confirmed that this choice might be more applicable for CPR-type data. A sample IMOS CPR GML instance document is viewable at http://aadc-maps.aad.gov.au/imos/CPRInstance.xml.

In addition to the problem of using the CSML Trajectory Feature to model an appropriate domain for CPR segments, it was necessary to develop specialised GML value objects as part of the MultiCurveCoverage rangeSet property to represent CPR biological data elements. Value objects in GML are simply pre-defined data structures that can be used to record values or measured quantities. The biological data elements used were derived from concepts defined in various biological ontologies championed by the international Biological Standards Group (TDWG, see http://www.tdwg.org/). These elements have been aggregated into a specialised IMOS object called "MarineBiotaStatistics" (see http://aadc-maps.aad.gov.au/imos/MarineBiotaStatisticV2.xsd for the GML schema). The types of biological information encoded in this object include:

- Taxon Concept Name: A Taxon Concept is a named classification unit (or taxon) as explicitly defined in a taxonomic treatment to which individuals, or sets of other taxon concepts are assigned (TDWG, 2007a). A Taxon Concept Name can be a scientific name or a vernacular name. If it is a scientific name it is governed by a biological code of nomenclature. In the case of IMOS data this may be the Codes For Australian Aquatic Biota (CAAB, see http://www.cmar.csiro.au/caab/) or the Register of Antarctic Marine Species (RAMS, see http://www.scarmarbin.be/rams.php?p=browser).
- Taxon Concept Life Science Identifier (LSID): a globally unique identifier (GUID) provided by some type of authority that represents a stable reference to a Taxon Concept (Kennedy *et al*, 2006). This property has been factored into the IMOS ontology in anticipation of GUID's becoming common place in the near future.

- TaxonomicPlacementFormal: A comma separated list of scientific taxon names that indicate (for administrative and data exploitation purposes only) the taxonomic placement of this object. The words should represent taxa of decreasing rank (TDWG, 2007b)
- TaxonRank: An enumerated rank of a taxon e.g. "species", "sub-phylum", "family", "sub-variety". (TDWG, 2007c).
- Taxon Count: Total number of individuals of a particular taxa observed in the sample,
- Taxon Maturation Stage: A textual description of the dominant life stages evident in the sample.

While the "MarineBiotaStatistics" object was devised with the CPR data in mind it could be made more general and re-used in other applications, for example in situations where statistics such as mean weight and mean length are calculated for specific taxa. These types of statistics are often recorded from analyses of fishing trawl and trap data. This "MarineBiotaStatistics" object is not appropriate, however, for observations made on individual biological specimens. Encoding the former type of statistical data also requires development of a new "composite phenomenon" which can be externally referenced in the sample schema (as in the example provided), or be declared locally.

3.5 Results of Evaluation

The results presented below are the salient issues arising from evaluating CSML and O&M and are not an exhaustive comparative summary of the two ontological models.

3.5.1 Structural Dimension

CSML conforms to GML encoding principles but has introduced a "Dataset" wrapper document to bind together in-line feature-based information and data storage artefacts encoded in a separate instance document. The storage descriptor instance document has a non-GML element at its root because GML has no valid abstract class available to represent such artefacts. However, despite these deviations from how GML has traditionally been applied, the syntax is valid XML and the novel approach taken allows a marriage of convenience between a feature-based model, as mandated by ISO and the OGC and other formatting approaches used within science that already have a significantly large user-base.

It was possible to extend the O&M and CSML ontological concepts to accommodate the IMOS sample data but some difficulties were encountered (e.g. the inclusion of quality flags for property values because of GML limitations). Both CSML and O&M rely on the concept of a phenomenon/parameter dictionary (or ontology) to express the observed properties of the features in question. These observed properties are declared primarily to help support discovery and exploitation of the encoded data values. In the case of biological data, the taxon concept names, which are the basic unit of currency and key to biological discovery paradigms, are unlikely to appear in phenomenon dictionaries. Using current GML constructs and patterns of expression the taxon concept names are more likely to be embedded at a lower level within the instance documents (as in the "MarineBiotaStatics" object example), effectively reducing their visibility for discovery purposes. The value entry in the "composite phenomenon" property is simply "taxon". For discovery purposes, when transferring biological data, reliance instead needs to be placed on linking taxon data in the GML instance document to an authoritative taxonomic names server (possibly via a LSID). These taxon concepts can then be mapped into a suitable repository in a services registry at service registration time and associated with specific services to facilitate querying against individual taxon concept names.

Both CSML and O&M were relatively modular in their construction. Both models import and include other schema and already re-use ontological components (e.g Sensor Web Enablement - SWE elements). Therefore, some difficulty was encountered when developing the derivative schemas to appropriately encapsulate all the nested sub-schemas and to validate them appropriately across the web using XMLSPY.

3.5.2 Functional Relevance Dimension

Sample CTD data was extracted from an AADC database and then served as an OGC Web Feature Service using some of the encoding patterns described in this paper. The tool selected was DEEGREE because it permits both querying and translation of underlying data, using XSLT, so that custom designed schema can be accommodated. By using DEEGREE and some existing public Java libraries from Unidata at the US University Corporation for Atmospheric Research (see http://www.unidata.ucar.edu/) it was also possible to query the underlying data using standard OGC Filter Language (see http://www.opengeospatial.org/standards/filter) but then send the data to a requesting client in binary (e.g netCDF) format instead of in GML. Using this approach, when data are housed in databases, CSML's non-GML compliant wrapper approach is possibly not even warranted and could be eliminated from the merged model. To serve data that is actually stored in legacy formats, however, the wrapper would still be required.

GML features all inherit a "metadataProperty" from the abstract feature class and this property provides a utility to develop a user-defined metadata package that relates to any feature in the instance document. This property was exercised in the "ObservationCollection" feature and used to reference an external ISO 19115 compliant dataset metadata record pertaining to all of the sampled features in the collection. Some metadata elements, drawn from the marine community metadata profile schema (AODCJF, 2006), were also instantiated in-line to identify the dataset owner, the dataset metadata custodian and dataset access conditions. The use of the "metadataProperty" in this manner assumed that all observations in the collection were associated with the one metadata record. It is highly possible, however, that data extracted from a database could be associated with multiple metadata records, depending on the granularity with which the data provider constructed his/her metadata records and their data service. If each observation in the collection was associated with a different metadata record it is possible to exercise the "metadataProperty" at the observation level. In this study, this approach was avoided in order to reduce verbosity and potential redundancy. This is based on the assumption that the community would agree a priori to the granularity required in constructing dataset level metadata and would then launch services congruent with this level of granularity.

Both CSML and O&M reference existing dictionaries and codelists. Many of these dictionaries are appropriate for reuse within IMOS, but IMOS will also need to develop some of its own resources in the short-term (e.g. for processing methods & project codelists and QCFlag dictionaries). There are a number of collaborative projects currently underway focussing on harmonising and serving vocabulary lists and dictionaries (e.g. Marine Metadata Interoperability Project) which can be leveraged. The NERC Data Grid (NDG), via the British Ocean Data Centre (BODC), has also instantiated a web service for delivering controlled vocabularies on-line (see http://www.bodc.ac.uk/products/web_services/vocab/) and is working on a dictionary of vertical coordinate reference systems. Where these lists meet IMOS needs they should be used. Controlled vocabularies already used by IMOS participants, such as the Global Change Master Directory (GCMD) codelists are already now accessible via this NDG vocab server.

3.5.3   Usability Dimension

A level of expertise has to be acquired to re-use and specialise CSML and O&M concepts and their properties. GML has quite an extensive specification (600 pages) and the language is not simple. Selecting the right CSML feature for a particular observation, or measurement type is not always immediately intuitive, despite examples provided in the CSML manual. During a work-shop, IMOS data managers conducted a collaborative desk-top exercise to assign each planned IMOS data stream to a CSML Feature Type, based on the CSML (V.2) Feature Type specifications. There was considerable debate and confusion about which would be the most appropriate Feature Type surrounding about 30% of the data streams (some of which were biological). Most data managers, however, agreed that if the CSML Feature Types could be used to represent IMOS data, the relatively small number of features that would be used to encode the data would provide efficiencies for IMOS developers in terms of the software that could be written to manipulate and visualise the data. Most participants also hoped to be able to leverage the work of overseas colleagues who would be working on software to address similar feature types.

At present there are few available CSML sample instance documents, which implies that CSML is an immature, but emerging language with relatively few real-life implementations outside of the UK NERC project. Woolf (personal communication, 21 March, 2007) indicated, however, that CSML will be the "starting point" for modelling feature types in the meteorology/ocean/atmosphere components of the European INSPIRE project (http://www.ec-gis.org/inspire/) and the GMES 'MyOcean' project which is developing an operational ocean forecasting capacity for Europe. CSML (V2) was released during 2007 and prior to that it had been relatively stable, with only minor updates in the previous 2 years. A lack of implemented O&M sample material on the internet, outside of the geoscience community, indicates that O&M is not yet in wide use in Australia or elsewhere. O&M was only released as a draft for comment in 2006 and a revision document will soon be available for the next version of O&M. O&M has also just been confirmed as an OGC standard. Both ontologies are being actively worked on and despite the current lack of implementation sites, the points made above give reassurance that both have the capacity to be long-lived and eventually pervasive in their adoption.

The consensus view of the IMOS data management experts, during the IMOS Standards workshop, was that both CSML and O&M provided considerable promise as a basis for developing an IMOS purposive ontology. However, development of IMOS centric encoding patterns and tools that ease the burden of transforming data into a CSML/O&M-based model would significantly lower the adoption barrier for IMOS data custodians and encourage uptake of these ontologies. A Feature-based Catalogue, capable of managing agreed community feature types, dictionaries, and encoding rules, which is also able to generate schema for "well-known" IMOS dataset types would considerably simplify the task of deployment by data service providers.

3.5.4   Maintenance Dimension

CSML is maintained at http://ndg.nerc.ac.uk/csml/. The site has a version controlled repository. Access to documentation and code is excellent and deprecated versions of the language are also accessible. Provision of more instance documents so that patterns described in theory can be examined in practice would increase the utility of the site. The O&M schema is maintained at https://www.seegrid.csiro.au/twiki/bin/view/ Xmml/ObservationsAndMeasurements. Like CSML it is maintained in a version controlled repository. The site is rich in examples and explanations about O&M schema constructs. Importantly the O&M twiki site also includes an "issues and change request facility" to enable user feedback and discussion on aspects of the schema.

3.5.5   Governance Dimension

Governance of CSML currently rests with its designer, but it is anticipated that as its user-base grows the governance mechanisms will broaden to encompass participation by peak bodies such as the World Meteorological and International Hydrographic Organisations, particularly with regards to feature specification. As a new OGC standard, O&M will now be subject to the OGC standards governance framework (see http://www.opengeospatial.org/ogc/programs/spec).

4   CONCLUSIONS

In evaluating CSML and O&M using the modified Annamalai & Sterling technique for constructing re-usable ontologies, it became apparent quite early in the process that a combination of both ontologies better satisfied the IMOS community selection criteria for construction of its purposive ontology, than either did individually. Given that the CSML and O&M designers had already recognised the utility of a harmonisation of their respective ontologies and had developed "hooks" between the two, it was not difficult to develop a specialised pattern of encoding for IMOS purposes, by harnessing both models.

The evaluation criteria used to assess ontological suitability were predominantly qualitative and although the approach was structured, the assessments lacked formal rigour and were subjective. The exercise was, however, practically beneficial in that the objective was to provide "expert" judgement about the applicability of existing ontologies in providing a basis for semantic data exchange within the IMOS community. This evaluation exercise has been able to demonstrate that a merged CSML and O&M ontological model is capable of carrying the semantic information required for at least 2 of the estimated 11 IMOS data streams, albeit with some supplementation. Furthermore, the evaluation exercise produced possible encoding patterns that can now be used to exploit the remainder of the data streams. In summary, the evaluation exercise itself has provided the foundation for construction of the IMOS data exchange ontology. An advantage of the approach used was that it was relatively rapid and was able to be performed by a domain, rather than a GML or ontology expert.

The development of encoding patterns and their applicability to the process used for evaluating ontologies should be tested in other domains. It would also be interesting to research which dimensions or criteria from this evaluation exercise, as well as, those used by others exploring this field of endeavour, are the most significant factors for decision-making in ontology selection exercises. A better understanding of these issues could lead to more practically applied robust evaluation techniques.

The IMOS community has two main challenges ahead. The first is to build on the encodings outlined in this paper and develop a features-based ontological model capable of representing all IMOS concepts and the second is to explore how to manage, govern and exploit this ontological model, through the development, or enhancement of appropriate software.

5.   REFERENCES

Annamalai M., Sterling L. (2003). Guidelines for Constructing Reusable Domain Ontologies, in Stephen Cranefield, Tim Finin, Valentina Tamma, Steven Willmott (eds), AAMAS03 Workshop on Ontologies in Agent Systems, CEUR Workshop Series, Vol. 73..

Australian Ocean Data Centre Joint Facility [AODCJF] (2006). Marine Community Profile of ISO 19115. Version 1.2. 2006-10-13. Retrieved in April 2007 from the WWW: http://www.aodc.gov.au/index.php?id=37.

Bainbridge S. (2007) An Initial Assessment of the Data Flows for IMOS Investment Areas and Implications for the eMII Project. March 2007, Unpublished.

Bennet N., Scott R., Brown M., O'Neil K., Lane M., Woolf A., Kleese van Dam K., Watkins J. (2006). Application of the NERC Data GRID Metadata and Data Models in the NERC Ecological Data Grid. Proceedings of the UK e-Science All Hands Meeting 2006. 18 - 21st September, Nottingham UK. Retrieved in May from the WWW: http://www.allhands.org.uk/2006/proceedings/papers/605.pdf

Corcho O., Gómez-Pérez A., González-Cabero R., Suárez-Figueroa M.C,. (2004). *Odeval: A Tool For Evaluating RDF(S), DAML+OIL, and OWL Concept Taxonomies*. AIAI 2004,IFIP WG12.6 -- First IFIP Conference on Artificial Intelligence Applications and Innovations, Toulouse France August, 22-27, 2004, A conference as part of IFIP World Computer Congress (WCC2004).

Cox S. (2006). Observations and Measurements. Discussion Paper. OGC 05-087r3 Version 0.13.0. Unpublished. Retrieved in March 2007 from the WWW: http://www.opengeospatial.org/standards/requests/37

Gangemi A., Catenacci C., Ciaramita M., Lehmann J. (2005) *A theoretical framework for ontology evaluation and validation*. In Semantic Web Applications and Perspectives (SWAP) -- 2nd Italian Semantic Web Workshop, Trento, Italy, 2005.

Gruber T. R. (1993). "What is ontology?" Retrieved March 2007 from the http://www-ksl.stanford.edu/kst/what-is-an-ontology.html

Gómez-Pérez, A.; Fernández, M.; de Vicente, A. (1996). *Towards a Method to Conceptualize Domain Ontologies.* Workshop on Ontological Engineering**.** ECAI'96. Budapest. Hungary, pp 41-52.

Guarino N., Welty C.A. (2004). An Overview of OntoClean. Ontolog. Conference Call. Ontologwiki. Retrieved in March 2007 from the WW: http://ontolog.cim3.net/file/resource/presentation/OntoClean-ChrisWelty_20041118/guarinowelty_final_v4.pdf.

Hartmann J., Spyns P., Giboin A., Maynard D., Cuel R., Suarez-Figueroa M.C., Sure Y. (2005). D1.2.3 Methods for ontology evaluation. KWEB/2004/D1.2.3./v1.3. A knowledgeWeb Deliverable. Work Package 1.2. Retrieved in March from the WWW: http://www.bioontology.org/evaluation_literature.html.

Kalfoglou Y., Schorlemmer M., Uschold M., Sheth A., Staab S. (2004). Semantic Interoperability and Integration. Seminar 04391 – executive summary, Schloss Dagstuhl – International Conference and Research Centre, September, 2004.

Kennedy J., Hyam R., Kukla R., Paterson T. (2006). Standard Data Model Representation for Taxonomic Information. OMICS – A Journal of Integrative Biology. Volume 10 (2), pp220-230.

Lake, R., Burggraf, D.S., Trinic, M., & Rae, L. (2004) *GML - Geography-MarkUp Language - Foundation for the GeoWeb.* Chichester, West Sussex, England: John Wiley and Sons, Ltd.

Lawrence B.N., R. Cramer, M. Gutierrez, K. Kleese van Dam, S. Kondapalli, S. Latham, R. Lowry, K. O'Neill and A.Woolf. The NERC DataGrid Prototype Proceedings of the U.K. e-science All Hands Meeting, (2003). S.J.Cox(Ed) ISBN 1-904425-11-9.

Lozano-Tello A., Gomez-Perez A. (2004). ONTOMETRIC: A method to choose the Appropriate Ontology. Journal of Database Management. Vol 15(2), pp1-18.

Sabou M., Lopez V., Motta E., Uren V. (2006). Ontology Selection : Ontology Evaluation on the Real Semantic Web. WWW2006, May 22-26, 2006, Edinburgh, UK.

Spyns, P. (2005) EvaLexon: Assessing triples mined from texts. *Technical Report 09, STAR Lab, Brussels, Belgium, 2005.*

TDWG. (2007a). Taxon Concept (LSID) Ontology. Retrieved in July 2007 from the WWW: http://rs.tdwg.org/ontology/voc/TaxonConcept.rdf

TDWG. (2007b). Ontology (LSID) Vocabularies - Globals. Retrieved in July 2007 from the WWW: http://rs.tdwg.org/ontology/voc/Common.rdf).

TDWG. (2007c). Taxon Rank (LSID) Ontology. Retrieved in July 2007 from the WWW: http://rs.tdwg.org/ontology/voc/TaxonRank

Uschold M. (2005). An ontology research pipeline. Applied Ontology (1). Pp13-16.

Woolf A., Lawrence B., Lowry R., Kleese Van Dam K., Cramer R., Gutierrez M., Kondapalli S., Latham S., O'Neill K., Stephens A. (2005). "Climate Science Modelling Language: Standards-based markup for metocean data", 85th meeting of American Meteorological Society, San Diego, Jan 2005.

Woolf A. (2007). Climate Science Modelling Language V2. Users Manual. Unpublished. Retrieved in March 2007 from the WWW: http://ndg.nerc.ac.uk/csml/

Appendix

Table 1 – Evaluation Measures

Dimension	Qualitative Measure
Structure	1. Conformance with language encoding principles and rules. 2. Encoding efficiency (ability to deliver complex and potentially voluminous data in compact structures). 3. Extensibility of ontology in terms of ability to easily add new concepts or specialise existing ones. 4. Modularity of ontology for re-usability.
Functional Relevance	1. Can the ontology readily meet the use-case goals? 2. Actual domain concept coverage (what fraction exists, even though assumption is that the coverage will be incomplete ?) 3. Harmonisation with ISO 19115 Metadata standard (i.e. what overlaps exist and is there scope to include ISO 19115 elements). 4. Applicability of included dictionaries and lists.
Usability	1. Complexity (in terms of user ability to model instance data using the ontology, i.e. level of expertise required). 2. Processing affordance (i.e. do the patterns chosen for constructing and encoding the concept have any potential operating synergy with service software that can recognise these patterns therefore improving scope to develop and associate re-usable data manipulation software ?). 3. Will the ontology be a "survivor" or a "one-minute-wonder"? 4. What is its current user implementation base? 5. Is it actively being worked on? 6. What is its state of flux (i.e. how often are there major revisions)? 7. Is it likely to become a standard, or parts of it be elevated to a domain ontology? 8. Could we readily convince our community-user base to use this ontology? 9. Assessment of any included ontology manipulation tools.
Maintenance	1. What is the maintenance base (i.e. in terms of people maintaining the ontology)? 2. Quality of access to XSDs. 3. Quality of access to instance samples. 4. Quality of access to conceptual models. 5. Quality of, and, access to documentation. 6. How often does the custodian release new versions? 7. Are old versions maintained and accessible?
Governance	1. Do the custodians encourage adopters to contribute to the base? 2. Is Governance transparent and participatory for the base concepts? 3. Is Governance participatory for maintained dictionaries and lists?

Table 2 – General Use-Cases

Use Case 1: Actors: Description:	*General Data Discovery & Selection* Data Users A system user interacts with a web-based client connected to a marine-themed, community-based services registry to search for and locate data of interest. The search paradigms offered encompass complex, multi-dimensional queries. Once data of interest has been found and a decision has been made to acquire the data, the user requests a copy of the data. The data is then sent to the user's browser client.
Use Case 2: Actors: Description:	*Data Integration* Data Users A system user interacts with a web-based client connected to a community-based services registry to acquire several similar datasets from a variety of sources. Having located the datasets of interest the user is able to visualise (e.g. plot) common attributes within these datasets, as if these attributes were drawn from a single data source. The user can then elect to have the variably sourced datasets combined into a single dataset and a copy of this integrated dataset sent to his/her browser client.

Table 3 Typical Competency Questions

1. Can I combine feature A (source A) with feature A (source B) for plotting purposes?
2. Is this named feature from source A the same type of feature as this named feature from source B?
3. Is feature A, a type of feature X?
4. What instruments record sea temperature?
5. What methods are used to sample pelagic biota?
6. Is specimen B taken from the same sampling station as specimen A?
7. What is the spatial accuracy of the location of this site?
8. What processing has been performed on dataset X and what dataset observation points were deemed to have passed these tests?
9. What other datasets and hence parameters were sampled during a particular deployment?
10. What project initiated the capture of specific data and what other datasets were captured by this project?