Interoperability in Costal Zone Monitoring Systems: Resolving Semantic Heterogeneities Through Ontology Driven Middleware

R.L. King, S. Durbha, and N. H. Younan

Department of Electrical and Computer Engineering , Mississippi State University, Mississippi State, MS 39762-9571, USA – rking@ece.msstate.edu

Abstract – Ontologies are widely recommended as a means of rectifying semantic heterogeneity. The advantage of using ontologies is that they can provide a conceptual schema regardless of a data set's format, structure, or size MSU is developing an ontological framework for resolving semantic heterogeneity problems in coastal zone data. This type of framework will provide the capability to (a) link the users to the knowledge, making integrated visualizations available; (b) provide search and query answering facilities; and (c) gather information at different levels of granularity, from the subcategory to the specific data level. Issues related to coupling such a system to models will also be discussed.

Keywords: Coastal zone, middleware, ontology, interoperability.

1. INTRODUCTION

The nation's coastal and ocean resources are under increasing pressure from population growth and development. Many organizations are involved in collecting data to measure the primary properties of coastal zones using a variety of methods ranging from remote sensing to in situ sensors and sampling. The understanding of the complex interrelationships within a coastal zone necessitates the exploration of strategies for innovative acquisition, integration, and data exploitation technologies for fully interchangeable, timely, and accurate geospatial data analysis and mapping.

Sharing of the generated datasets, information, and results, between geographically distributed organizations often proves to be challenging. This is due to the complicated steps involved in data discovery and conversion that result from the problems of syntactic, structural, and semantic heterogeneity in the datasets. The syntactic heterogeneity problems have been addressed to some extent by the standardization of metadata as advocated by multiple organizations. However, the lack of sufficient description of the meaning of the data along with a context may lead to the misinterpretation of data by users who are not involved in the original data acquisition process. Thus, semantic reconciliation is necessary to guarantee meaningful data sharing (i.e., the exchanged data is correctly interpreted and used).

2. INTEGRATED OCEAN OBSERVING SYSTEM

The Integrated Ocean Observing System (IOOS) is the U.S. contribution to the Global Ocean Observing System (GOOS) and to the oceans and coasts components of the Global Earth Observation System of Systems (GEOSS). The IOOS will provide data and information needed to significantly improve the nation's ability to achieve seven societal goals [Hankin, 2004]:

- Improve weather forecasts and predictions of climate change;
- Improve safety and efficiency of marine operations;
- Provide more timely predictions of natural hazards and their impacts;
- Improve national security;
- Reduce public health risks;
- Sustain, protect, and restore healthy marine and estuarine ecosystems; and
- Sustain, protect, and restore marine resources.

The IOOS is envisioned as a user-driven, coordinated, national and international network of observations, data management and communications, and data analyses that systematically acquires and disseminates data and information on past, present, and future states of oceans and coasts (including the U.S. Exclusive Economic Zone (EEZ), Great Lakes, and estuaries) [ROW, 2004]. More then 80 percent of the American population lives within 50 miles of a coast, a population component that has doubled in the past decade [NRC, 2004]. The coastal zone comprises many important onshore habitats (e.g., forests, rivers and streams, wetlands, estuaries, beaches, barrier islands), coastal infrastructure, as well as the coastal ocean. Therefore, it is imperative that strategies for innovative acquisition, integration, and data exploitation technologies be explored for fully interchangeable, timely, and accurate geospatial data and mapping that extends across the coastal zone.

The IOOS will consist of three subsystems [Hankin, 2004]:

- Observing Subsystem (remotely sensed and in situ measurements and their transmission from platforms);
- Modeling and Analysis Subsystem (evaluation and forecast of the state of the marine environment based upon measurements); and
- Data Management and Communications Subsystem (DMAC) (the integrating component)

Very simply, IOOS is a measurement, prediction, and integration system for the ocean component of the earth system.

2.1 Data integration through ontologies

Data integration tasks can fall into several categories of varying complexities [Walter and Fritsch, 1999]:

- Integration of data sets stemming from the same data-source with unequal updating periods.
- Integration of data sets represented in the same data-model, but acquired by different operators.
- Integration of data sets which are stored in similar, but not identical data-models.

- Integration of data sets from heterogeneous sources, which differ in data-modeling, scale, thematic content, etc.

The IOOS interoperability/integration task is that described by the more challenging last task.

We are demonstrating the applicability of the emerging semantic web technologies for semantics enabled resource and knowledge discovery for IOOS. The current implementation of IOOS addresses a subset of the problems of interoperability by solving the syntactic issues via OPeNDAP [OPeNDAP, 2004]. Nevertheless, complete and seamless interoperability is not achievable until the semantic issues of data and services are resolved. Our approach for achieving seamless interoperability between heterogeneous resources is based on their semantic content. In order to achieve semantic interoperability in a heterogeneous information system, the meaning of the information that is interchanged has to be understood across the systems. However, semantic conflicts can occur when two contexts do not use the same interpretation of the information [Visser, 2002]. This semantic heterogeneity exists in two forms, namely cognitive and naming [Bishr, 1998]. Cognitive semantic heterogeneity results from not having a common base of definitions between two (or more) groups (i.e. terms of reference). Defining such terms of reference amounts to constructing a shared ontology, or at the very least, points of overlap [Pundt and Bishr, 2002]. Naming semantic heterogeneity occurs when the same name is used for different concepts or different names are used for the same concept. It is not possible to undertake any semantic analysis until problems of semantic heterogeneity are resolved.

Ontologies are widely recommended as a means of rectifying semantic heterogeneity. An ontology is "a shared, formal conceptualization of a domain" [Gruber, 1993]. As this definition suggests, ontologies differ from data models in two significant aspects.

- Ontologies build upon a shared understanding within a community. This understanding represents an agreement over the concepts and their relationships that are present in a domain.
- Ontologies use machine processable, logic-based representations that allow computer manipulation. This includes transferring ontologies among computers, storing ontologies, checking the consistency of ontologies, reasoning about ontologies, etc.

3. METADATA

3.1 Metadata standards

Several metadata standards have been developed to address syntactic standardization [Hatton, 1997]. However, in systems with a large number of available data sources (e.g., IOOS); it is often not trivial to find the right set of data for a given task. Metadata, i.e. data describing a data set and its source, are often used to organize and manage such heterogeneous networks of collections. Metadata standards that originated in the environmental community and were specifically designed for environmental and geospatial data are the Content Standards for Digital Geospatial Metadata (CSDGM), developed by the Federal Geographic Data Committee (FGDC) [FGDC, 2004]. It is important to realize that metadata standards allow us to structure the file contents, but they do not help us define their semantics.

Figure 1 illustrates the need for rich metadata to provide a variety of services for domain application software. Where the structural metadata typically provides a full description of the physical parameters of the data file, the semantic metadata provides a meaning of the data along with a context, to allow the application to understand what it has read and how to use it. The advantage of using ontologies is they can provide a conceptual schema regardless of the format, structure, or size of a data set. Ontologies can be designed to semantically understand the content and structure of data present in the data set. Once we have semantic metadata, it is possible to enable interoperability between heterogeneous data sets through shared ontologies. Thus, content based discovery and retrieval of datasets is not only possible, but also achievable.

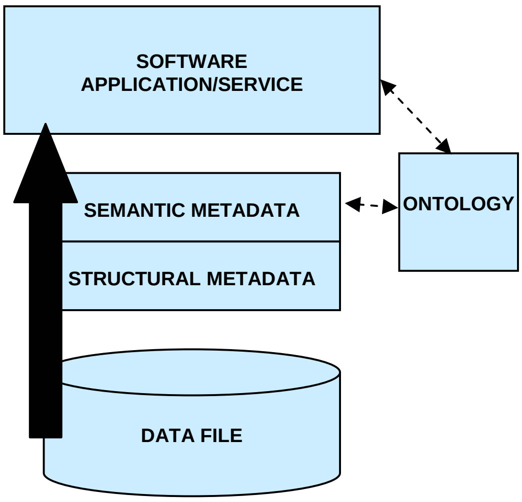

Figure 1. Application services require metadata that is interpretable through ontologies for resource discovery.

Another aspect of metadata standardization is structural and semantic integration. This is achieved through the development of a standardized terminology, which is used to fill the metadata models with information. In this integration the development of thesauri (for attribute data) and gazetteers (for geographic data) plays an important role. Although progress has been made with respect to syntactical, structural, and semantic standardization, the large number of different metadata standards and environmental thesauri illustrate a key problem of all standardization efforts. The user communities in the environmental field are too heterogeneous to allow creation of a single data model or a single terminology to satisfy all users [Visser, 2001].

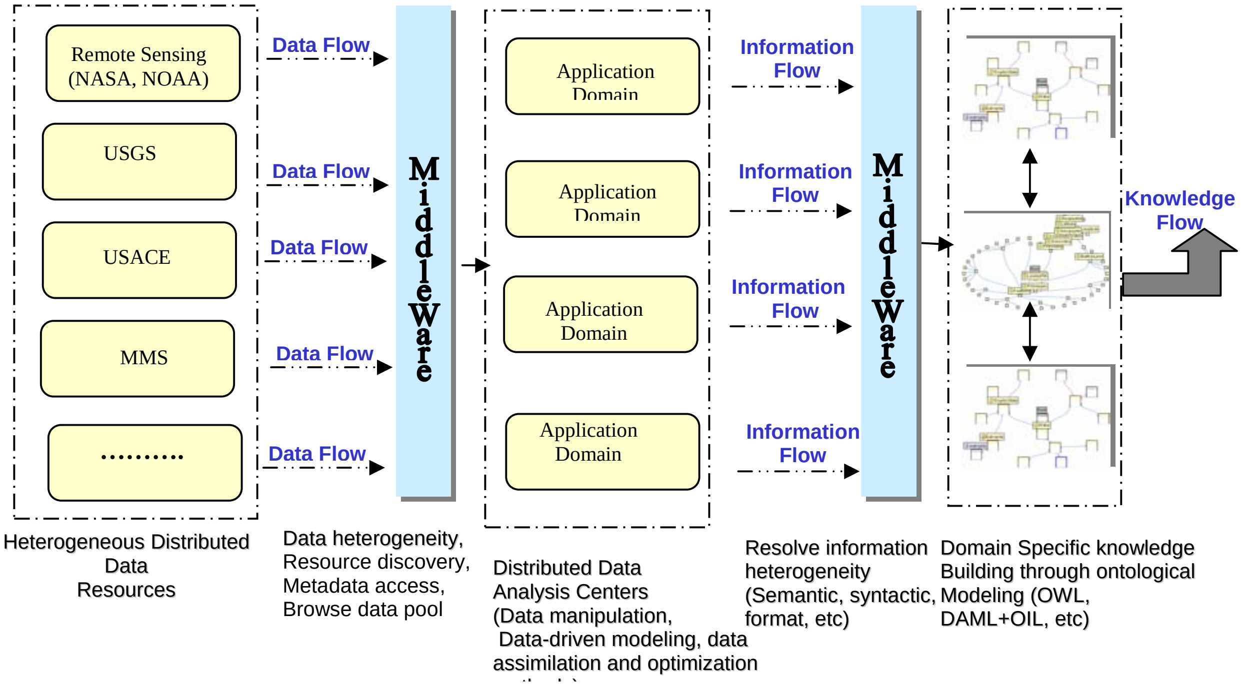

Figure 2. A process diagram for transforming distributed data resources into knowledge.

3.2 Metadata in IOOS scenario

The FGDC standard for geospatial metadata is widely favored in the ocean observation community. While the FGDC CSDGM is quite effective in classifying information relating to physical phenomenon, it is not as well suited to handle the biological aspects of many ocean processes. Also, the standard provides text based syntactic metadata with virtually no semantics and machine understandability. The metadata standard is also very complex with 334 different elements, 119 of which exist only to contain other elements [Parekh, 2004]. Integrated ocean observations networks are highly interdisciplinary in character and thus face enormous semantic problems. Since this type of semantic heterogeneity is already a problem for human experts in communicating with each other, it becomes even more challenging when attempting to integrate these terminologies automatically.

3.3 Semantic metadata model

Early approaches for semantic integration were mainly based on the use of thesauri to translate between specific vocabularies [Visser, 2001]. However, a more recent approach is the use of ontologies to handle the problem of implicit and hidden knowledge by making the conceptualization of a domain explicit. The advantage of this approach is that it represents a standard that is widely accepted by the oceans research community.

Since the initial observing subsystem for IOOS is to be built using existing assets as described in the final plan of IOOS [IOOS, 2004], we are developing semantic metadata through ontologies developed on top of the existing metadata in FGDC format. This task will be accomplished by the development of an Ocean Semantic Metadata (OSM) ontology that incorporates the elements from the FGDC-CSGDM. Some of the elements of the CSGDM are not sophisticated enough with regard to expressiveness or they lack formal semantics that are required to describe the concepts for ocean observations (e.g. Phytoplankton abundance). Thus, there is a need for additional qualifiers for those elements, which can be described in a language that provides formal semantics (e.g. Ontology Web Language (OWL)). Eventually, as depicted in Figure 1, OSM will contain process specific concepts and their relations through linking application level ontologies (e.g. coastal hazards, ocean biology, marine meteorology).

3.4 Knowledge discovery from multiple sources and methods through ontologies

Knowledge discovery from large heterogeneous networks of ocean observations and information products generated from modeling efforts is of paramount importance for IOOS. However, before knowledge can be discovered and shared it has to be formalized in such a manner that it is machine accessible and understandable.

Task or context-specific analysis of oceanographic data requires exploiting the relations between terms used to specify the data, to extract the relevant information, and integrate the results in a coherent form. As depicted in Figure 2 the *data* from various sources (MMS, NOAA, EPA, etc.) are transformed into *information* at different application domain data analysis centers. However, to achieve this, middleware is required that provides tools to browse and access the data resources for resolving the heterogeneity problems. *Knowledge,* defined in this context as usable information for a specific domain application, aids in effective decision-making. Domain specific knowledge building can be achieved through ontological modeling that provides functionalities for capturing knowledge structures. Ontologies can be used for data integration tasks (because of their potential to describe the semantics of information sources), to solve heterogeneity problems, and to provide varied levels of querying services that help exploring the knowledge at different levels of granularity. Therefore, we are building ontology driven middleware that enables the transformation of heterogeneous data and information into knowledge by intelligent utilization of relevant resources.

4. MIDDLEWARE FOR ONTOLOGY-DRIVEN BROKERING (MOB)

Middleware by its most general definition is any programming that serves to "glue together" or mediate between separate and often already existing programs. Middleware in distributed systems received its name from its relative position between the platform (OS and network services) and software applications. This layer is introduced in systems to hide the heterogeneity of the underlying components or applications and provide uniform access to their functions. Moreover most middleware provide higher-level services, such as messaging, transaction handling, and security. In short, middleware facilitates interoperability by hiding low-level access and by providing standard services.

We are developing a Middleware for Ontology-driven Brokering (MOB) that provides functionalities for ontology management, storage, query, and inference services. It also enables resource discovery and mediation through ontologies built on top of the metadata of the resources. The translation of metadata to semantic metadata through ontology-based approaches is also handled by MOB.

MOB will have the following features:

- An ontology server providing the basic storage services with functionalities for tracking, configuration management, access control, and security
- Mechanisms for knowledge management.
- Multi-protocol client access to allow different users and applications to use the system via the most efficient "transportation" media.
- Support for integration of variety of reasoning modules suitable for various domains

5. CONCLUSIONS

Semantic interoperability requires resolving various context-dependant incompatibilities. The context refers to the knowledge that is required to reason about the system for answering a specific query [Sudha, 2004]. Therefore, it is important to provide contextual knowledge of domain applications in order to ensure semantic interoperability. Each information source serves as a context for the interpretation of the information contained therein. This view implies that an information entity can only be completely understood within its context and we need to find ways to preserve the contextual information in the translation process. This line of reasoning leads us to conclude that:

- It is necessary to represent contextual knowledge of an information entity.
- Translation of an entity from one context to another requires the use of the contextual knowledge captured in the previous step.

Assuming that ontologies are used to capture the context of the information entities, then as we move from one context to another, there is a requirement to integrate ontologies. There are different approaches to ontology integration. A hybrid ontology approach consisting of a global shared ontology that encompasses all the local application level ontologies for a domain of interest (e.g. Coastal zone hazards) is being adopted for this work.

6. REFERENCES

Bishr, Y. (1998). Overcoming the semantic and other barriers to GIS interoperability. Int. J. Geographical Information Science, 12 (4), pp. 299-314.

Federal Geographic Data Committee (2004). Content Standard for Digital Geospatial Metadata, Version 2, FGDC-STD-001-1998 http://www.fgdc.gov/metadata/contstan.html

Gruber.T.R. (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5:199–220.

Hankin, S. and the DMAC Steering Committee (2004). Data Management and Communications Plan for Research and Operational Integrated Ocean Observing Systems: I. Interoperable Data Discovery, Access, and Archive, Ocean.US, Arlington, VA, 292 pp. http://www.dmac.ocean.us/dacsc/imp_plan.jsp

Hatton, B. (1997). Distributed Data Sharing. Paper presented at the AURISA 97, Christchurch, New Zealand.

IOOS (2004). IOOS First Annual Implementation Conference, http://www.sccoos.org/docs/Annual-IOOS-Goals.pdf

National Research Council Committee on National Needs for Coastal Mapping and Charting, (2004). A Geospatial Framework for the Coastal Zone: National Needs for Coastal Mapping and Charting, National Research Council, Washington, D.C., 149 pp., http://www.nap.edu

OPeNDAP (2004). http://opendap.org/

Parekh, V., J. Gwo, and T. Finin (2004). Ontology based Semantic Metadata for GeoScience Data, International Conference of Information and Knowledge Engineering, June 2004.

Pundt, H. and Bishr, Y. (2002). Domain ontologies for data sharing-an example from environmental monitoring using field GIS. Computers and Geosciences, 28, 95-102.

ROW (2004). http://www.ocean.us/documents/docs/ROWProceedings-composite-v3.doc

Sudha Ram, Jinsoo Park (2004). Semantic Conflict Resolution Ontology (SCROL): An ontology for detecting and resolving data and schema-level semantic conflicts, IEEE Transactions on Knowledge and Data engineering, Vol.16 no.2 February, 2004.

Visser U, Stuckenschmidt H, Schuster G, et al.(2002). Ontologies for geographic information processing. Computers and Geosciences, 28 (1): 103-117.

Visser, U., Stuckenschmidt, H., Wache, H., & Vögele, T. (2001): Using Environmental Information Efficiently: Sharing Data and Knowledge from Heterogeneous Sources, In C. Rautenstrauch & S. Patig (Eds.), Environmental Information Systems in Industry and Public Administration (pp. 41-73). Hershey, USA & London, and UK.

Walter, V. & Fritsch, D (1999). Matching Spatial Data Sets: Statistical Approach, International Journal of Geographical Information Science 13(5), 445–473.