Maritime Information Markup and Use in Passage Planning

Raphael Malyankar
Dept. of Computer Science and Engineering
Arizona State University, Tempe, AZ 85287-5406.
E-mail: `rmm@acm.org`
URL: http://www.eas.asu.edu/∼gcss/research/navigation/

Abstract

This paper describes the integration of maritime information from multiple sources in the context of passage planning for coastal voyages. It begins with describing the construction of a computational ontology for maritime information and nautical chart symbology. The use of this ontology in defining a markup language is then described, followed by a discussion of the use of this ontology and markup language in a demonstration information retrieval application.

1 Introduction

This paper describes the creation of an information model for maritime information, and the use of this information model in markup and integrated retrieval from multiple sources in the context of passage planning for coastal vessels.

Government interest arises from the fact that collecting, preparing, and distributing this information is part of the missions of a number of Federal agencies, particularly NOAA, NIMA, and the USCG. Some of it is already being distributed in computerized form, often via the Web. The purpose of the project is to facilitate smarter use of the available information, that is, its use in intelligent software; and, further, to enhance the interoperability of existing and future software and databases by providing a mutually intelligible means of expressing and sharing this information.

The creation of a model of available knowledge and information constitutes 'ontological engineering'. For our project, ontological information is acquired from multiple sources, including standards documents, database schemas, lexicons, collections of symbology definitions, and also from semi-structured documents. The computational ontology[1] thus created is being used to create an XML-based markup language (Maritime Information Markup Language — MIML) for tagging documents within this domain. This markup language is used (as an initial demonstration of the kind of application that will be enabled) in a Web-based *passage planning* system that retrieves and presents necessary and relevant information for mariners planning a voyage, from Web sites, digital nautical charts, and marked-up text documents.

The rest of this paper describes the sources of ontological knowledge and the contribution of each source to the overall ontology. An overview of a markup language for maritime information derived from this ontology follows. The use of this markup language and ontology in a demonstration application is then described.

[1] In artificial intelligence, an *ontology* is a set of definitions of the concepts that exist in a particular domain and the relationships between them. A *computational ontology* is a a collection of terms, formal definitions, and constraints, which can be processed by software, and which increases the scope of computational methods applied to the relevant domain.

2 Ontology Construction

Ontological knowledge was derived from the following sources:

Standards Documents: A normative standard for digital nautical chart content is the IHO (International Hydrographic Organization) S-57 Transfer Standard for Digital Hydrographic Data [6]. The 'object catalog' section of this document consists of a list of chart entities, definitions, and entity attributes, which gives us a collection of domain entities that can be considered canonical as far as the scope of the standard goes. Extraction from this 'object catalog' was automated using graph traversal algorithms that exploit links between entities and attributes. The automated extraction resulted in 173 classes. A comparison of 10% (selected at random) of the extracted information with the original source indicated error rates of 8% to 20% (for different categories of ontological knowledge — classes/types/attributes). The additional effort needed to reduce this rate in the automated extraction was not undertaken, as it proved no very laborious task to make the corrections by hand (about 10 hours for a non-expert who compared the extracted ontology with the original source).

The Spatial Data Transfer Standard [3] was another source. The parts we used were the list of 'included terms' (analogous to a hyponym list) and attribute definitions. Extraction from this was less satisfactory in some ways, since these sections are less rigorous than the object catalog of the S-57 standard, but, on the other hand, the lists cover more of the terms used in practice.

Databases: The primary digital chart database we have used so far is the set of sample Digital Nautical Chart (DNC) data files available from the National Imagery and Mapping Agency (NIMA), covering the San Diego Harbor and approaches. The DNC database have somewhat more *semantic* structure than the aforementioned standards, consisting as it does of feature classifications organized by 'layers', for example, environmental features, cultural features, land cover features, etc. Induction of ontological knowledge from this consisted of mapping the structure to a class hierarchy. Taxonomical information that could be directly extracted from the table names in this database therefore consists of relationships between the abovementioned features/classes. Approximately 134 classes were mined from this database.

As with the S-57 standard, this database and schema covers only chart entities, and the terminology is even more restricted (and to some extent, more linguistically opaque) than the S-57 standard, due to the use of abbreviated names for entities and attributes, and the lack of textual definitions.

Lexicons and Symbology Definitions: We also used the Stanford Medical Informatics group *Protégé* tool [4] and a standard collection of symbology definitions from the National Oceanic and Atmospheric Administration (NOAA *Chart No. 1*) [11] to create an ontology of navigation aids, hazards, and other entities. *Chart No. 1* is a list of the symbology used in nautical charts accompanied by brief definitions of what each symbol stands for. It is organized semantically (in that related symbols are in the same section or subsection). Approximately 500 classes were created from this source. Definitions available within this document were supplemented by using a widely popular publication on navigation and seamanship, (*Chapman Piloting* [8]) and an online dictionary of chart terms (discovered and used by the creator, a student unfamiliar with nautical terms). Ontology creation based on these documents consisted of manual entry of information using Protégé, due to the lack of electronic versions of the symbology definitions.

Semi-structured material: The *United States Coast Pilot* is a 9-volume series containing information that is important to navigators of US coastal waters (including the Great Lakes). Included are photographs, diagrams, and small maps. The flow of text follows the coastline geographically, e.g., from north to south. This

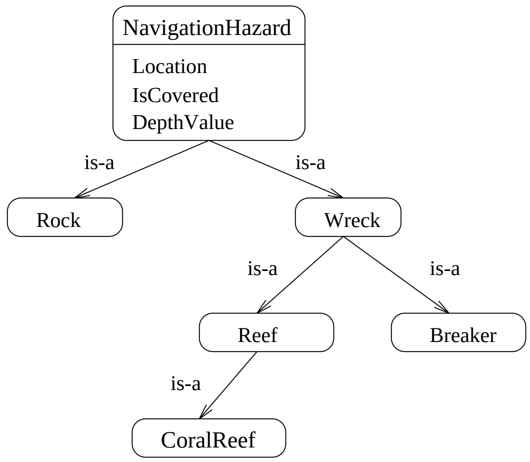

Figure 1: Part of the taxonomy derived from NOAA Chart No. 1

is a ‘lightly structured’ document, with each volume containing a preliminary chapter containing navigation regulations (which includes a compendium of rules and regulations, specifications of environmentally protected zones, restricted areas, etc.), followed by chapters dealing with successive sectors of the coast. Each chapter is further divided into sections (still in geographical order); each section is further divided into sub-sections and paragraphs describing special hazards, recognizable landmarks, facilities, etc.

The internal structure of subsections and paragraphs provides taxonomic hints, indicating, for example, which leaf entities are categorizable as sub-classes of weather conditions, as well as providing a small amount of additional taxonomical information that extends taxonomies derived from other classes (e.g., tide races as a form of navigational hazard). The *Coast Pilot* is normative (in the sense of using well-understood terms) and comprehensive. A version marked up with XML would have proved invaluable for ontology learning, but there is no such version available at this time (indeed, it is the aim of this project to create such a marked up version).

There is a certain amount of overlap in the ontological information derived from the different sources above, in addition to structure mismatches (for example, information that is denoted as an attribute in one ontology may be used for sub-classing in another). The reuse of other computational ontologies was explored early on, but they (e.g., the SENSUS ontology in Wordnet) turned out to define terms in usages that are either not relevant, or ‘wrong’ in the maritime information domain — for example, a ‘bridge’ is a trafficable passageway for land travel, but an obstruction or landmark for waterborne vessels.

We have discovered that though there is a certain amount of duplication between the above sources, they are largely independent and produce different parts of the taxonomy for the maritime information domain as a whole. It is therefore necessary to merge the ‘sub-ontologies’, i.e., construct a larger ontology that contains all the terms and that reconciles discrepancies between them. Our approach to the merging problem is described elsewhere [9]. Figure 1 shows part of the ontology created.

3 Markup Language

The Maritime Information Markup Language (MIML) follows the rules established for XML by the W3C consortium. Tags in this language are derived from object names in the object catalog, symbology terms in NOAA Chart No. 1, feature names in the sample Digital Nautical Chart database, and other entities defined from weather reports and other sources of ontological knowledge. For example, the <Anchorage> and <Pier> tags in Figure 2 are directly derived from the corresponding class names in the ontology. A second

```
<Chart>
 <ChartNumber>18773</ChartNumber> <ChartNumber>18772</ChartNumber>
 <Location> San Diego Bay is 10 miles NW of the Mexican boundary</Location>
 <Description>San Diego Bay is where California’s maritime history...</Description>
 ...
 <Anchorages>General anchorages, special anchorages, and ... </Anchorages>
 <Tides> The mean range of tide is 4.0 feet at San Diego ... </Tides>
 <Currents> The currents set generally in the direction of...</Currents>
 ...
<Wharves>
 The San Diego Unified Port District owns the deepwater commercial facilities in...
 <PierArea>
  <Pier  name ="B Street Pier, Cruise Ship Terminal">
  (32 deg. 43’02"N., 117 deg. 10’28"W.): 400-foot face, 37 to 35 feet alongside;
  1,000-foot N and S sides, 37 to 35 feet alongside;...
  </Pier>
  <Pier  name ="Broadway Pier, S of B Street Pier">
  135-foot face, 35 feet alongside; 1,000-foot N and S sides, 35 feet...
  </Pier> ...
  <Pier  name ="Tenth Avenue Marine Terminal">
   <Berth  name="Berths 1 and 2">
    Concrete bulkhead, 1,170 feet of berthing space; 27 feet alongside...
   </Berth>
   <Berth  name="Berths 3 and 6"> ... </Berth>
   <Berth  name="Berths 7 and 8"> ... </Berth>
  </Pier>
 <PierArea>
</Wharves>
...
</Chart>
```

Figure 2: Marked-up fragment of Chapter 4, Volume 7 of the *Coast Pilot*

class of tags was also needed to denote information elements within the document, for example the <Chart> tags, which denote sections that pertain to a specific (identified) nautical chart in the NOAA chart numbering system, and the <Description> tag, which is used to denote general text information that cannot be placed into a more specific category. While the first class of tags for MIML were derived from the ontologies mentioned earlier, the DTDs (Document Type Definitions) for the Coast Pilot were prepared ‘by hand’, with tags in the second class being invented as necessary.

Figure 2 shows a marked-up fragment of Volume 7 of the *Coast Pilot*. This fragment is part of Chapter 4, which covers the California coast from San Diego to Point Arguello, and describes, amongst other things, weather, navigation hazards, aids for navigation, local regulations, contact information, harbor facilities, etc. The section in the figure comes from the portion describing harbor facilities in San Diego Bay. (Ellipses denote material left out of this figure for brevity’s sake.)

4 Prototype Passage Planning Web Site

A ‘passage plan’ is, for the purposes of this project, an answer to the questions: ”How do I get from X to Y? What will I encounter on the way, and what will I find when I get there? What do I need to know for this particular journey”? Passage planning involves not just plotting a safe route, but also includes generating a report about hazards that may be encountered, facilities available along the route and at the destination, weather and tide conditions that may encountered during the voyage, etc. The passage plan depends on the type of vessel and the purpose of the journey, since information that may be of interest to a freighter may be irrelevant to a small pleasure craft. The use of these concepts, and of MIML is demonstrated in the prototype site described next. The demonstration prototype in its current stage of development does not tackle the route-planning problem ("how do I get from X to Y"), because similar issues have long been addressed in path planning research within artificial intelligence, and the computational magnitude of this particular problem prevented anything more than a superficial solution with available resource. However, the demonstration does attempt to deal with the other components of what we call the the "passage planning question", though the sources used are by no means the only sources for this information.

4.1 Content sources for prototype site

The 'content' sources for passage planning are Web sites with real time information, the Coast Pilot, and programs that generate information as and when required. They fall into the following categories:

Static textual documents: The primary text document currently being used is the Coast Pilot, described earlier. It includes descriptions of particular items of interest, some of which are also shown in nautical charts (such as lighthouses and beacons), and other descriptions which are either not available in the nautical charts and other places, or not apparent from them, such as special local tidal dangers. (Where information in the CP duplicates that in other sources, we interpret it as emphasizing important features and dangers.) It also contains a few diagrams and photographs taken from a mariner's point of view, information on anchorages, etc., and pointers to other sources, for example, to the U.S. 'Port Series' for more information on facilities at a specific port. Markup of the Coast Pilot with MIML tags is currently being done manually; we hope to automate part of this markup process in the future.

Web sites with real-time information: Certain information is being made available in near-real-time by both official and unofficial sources, especially weather conditions, forecasts, and warnings. The National Databuoy Center (NDBC) Web site provides recent weather data from databuoys all along the U.S. coastline. Certain marinas have also begun putting local conditions on their Web sites. Tide predictions are available from another NOAA site. The prototype incorporates information from databuoys (via the NDBC Web site) into the information it collects for a specific 'passage plan'.

Chart databases: Data extracted from DNCs is loaded into a local database (due to the difficulty of querying DNC files). The contents are essentially tables of features and their attributes. These tables are easily transformed into an object-relational database form. This source provides information about such items as coastlines, marker buoys, lighthouses, depth measurements, and other nautical chart features.

Dynamically generated content: Certain content (tide predictions) is generated by programs residing on the local web server.

4.2 Capabilities demonstrated by prototype site

The capabilities demonstrated in the prototype are:

Extraction from XML documents: Relevant elements from the marked-up Coast Pilot are extracted in response to a user query. Relevance is judged based on proximity to the location(s) specified (and the route

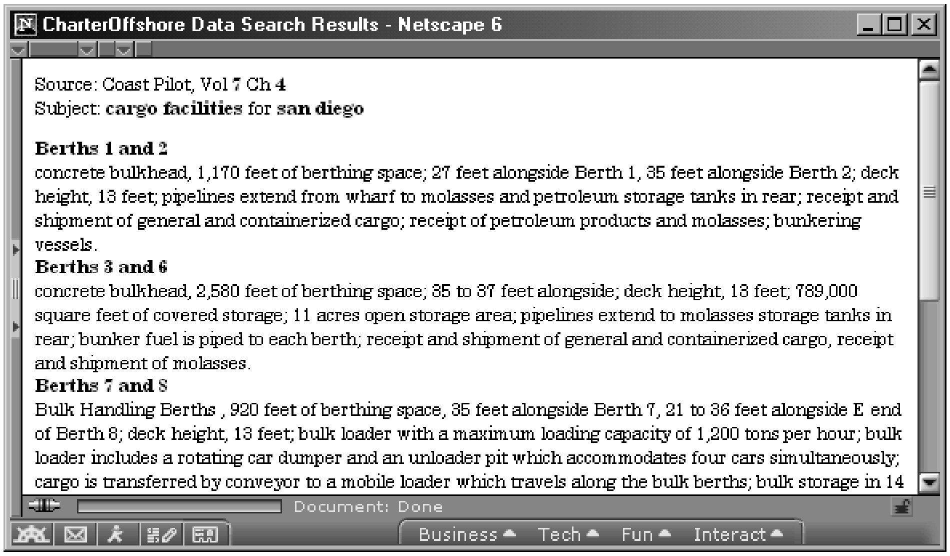

Figure 3: Elements describing cargo facilities, extracted from the *Coast Pilot*

between source and destination), the type of vessel and purpose of the voyage, and in response to the optional question mentioned earlier. Figure 3 shows the response to a question about cargo facilities at San Diego.

Real-time information presentation: Information about weather conditions as reported by NDBC data buoys is retrieved from the NDBC web site, and processed into a form suitable for presentation, this time with MIML markup added automatically.

Data retrieval: The databases created from DNCs are queried with SQL queries and the results transformed into forms suitable for presentation. This transformation currently involves statistical post-processing of the information, the nature of this post-processing depends on the form of the query, especially when the retrieval is raw material for an response to the optional user questions mentioned earlier. Transformation is discussed further in the next paragraph.

Simple question answering: The prototype is able to answer questions posed using a limited vocabulary and syntax. Questions can be asked in ways that are close to natural language (e.g., “can I anchor off San Diego”). Pattern-matching is used to transform this natural-language question into a query that can be executed by the database back-end. The techniques used in information retrieval and processing of retrieved information may involve the capabilities described earlier in this section. In some cases, the raw data retrieved from the database undergoes post-processing depending on the form of the question; for example, the questions “show the sea floor off …” and “can I anchor near …” both retrieve the same raw data (sea floor characteristic data points), but the first form produces a table of values, while the second combines the the retrieved values into an assessment of the general sea floor description in the same location.

4.3 Functioning of Prototype

The primary interface with the user consists of a form to be filled out with information about the journey, including the location (either source-destination or a single point), type of vessel (cargo, sail, etc.), time of journey, and, optionally, specific questions about such items as anchorages, local facilities, depths, etc. The

Web server transforms the form into a collection of sub-queries, each formulated for the specific knowledge sources available (here, Web sites, DNC database, marked-up CP files, and a tide prediction program). The transformation process is based on matching keywords or key phrases from user questions, combined with table-based lookups of menu selections (e.g., type of vessel), and use of the ontology to decide specifically what information must be obtained from the information source. The answers obtained after executing individual sub-queries are combined into a 'passage plan' or mini 'portolan chart' that is customized for the specific voyage.

The current version of this page limits its search to the four sources mentioned earlier and answers a limited range of questions, being constrained by the limited richness of structure of the sources (e.g., the CP is marked up with MIML tags only at sentence- or paragraph-level detail, whereas tagging parts of sentences is required for more sophisticated retrieval). Efforts to integrate more sources and enhance the expressiveness of the markup language are underway.

5 Related Work

In a paper on agents for information gathering, Knoblock and Ambite [7] describe the use of a domain model in formulating queries for different knowledge sources represented by different agents. Noy and Musen [12, 13] describe an algorithm and tool for merging ontologies in Protégé. Chalupsky [1] describes OntoMorph, a tool for translating symbolic knowledge from one KR formalism to another, and describes ontology alignment in [2]. Hovy [5] describes a procedure for ontology alignment and heuristics for suggestions, including pattern matching on strings, hierarchy matching and data/form heuristics. Ontology analysis and merging in *Chimæra* is described in [10]; the techniques used include syntactic analysis of class and slot names, taxonomic resolution, and semantic evaluation (for example, slot/value type checking and domain-range mismatches).

6 Summary and Future Work

The primary purpose of this paper was describing the creation of a markup language for maritime information (MIML) from computational ontologies and the use of this language in information retrieval from documents and other sources used in the field. MIML is still in the early stages of development, and needs to go through a standards process before it can gain wide acceptance in the field. Research plans for the future include demonstrating capabilities beyond information retrieval, especially intelligent reasoning, using the retrieval and access capabilities provided by markup; this will involve 'drill-down' markup, to lower levels than in the sample fragment of Figure 2. Querying of large databases of XML documents, the use of markup and ontologies in delivery of information to users, and the use of markup in updating databases (and documents) and in translating between heterogeneous databases will be investigated. The implications for our research of the Resource Description Framework (RDF) and DARPA Agent Markup Language (DAML), which are both currently still under development, will also be examined.

Acknowledgements

The efforts of Koi-Sang Leong, Helen Wu, and Rashmi K Iyengar in creating the ontology and markup of the Coast Pilot are gratefully acknowledged, as are the efforts of Suresh Lalwani and Jayagowri Renuka in programming for the demonstration described. This work was partially supported by the National Science Foundation under grant EIA-9983267, NOAA, Maptech, and the U.S. Coast Guard. The opinions, findings, conclusions and recommendations expressed in this material are those of the author and do not necessarily reflect the views of these entities.

References

[1] H. Chalupsky. Ontomorph: a translation system for symbolic knowledge. In A.G. Cohn, F. Giunchiglia, and B. Selman, editors, *Principles of Knowledge Representation and Reasoning: Proceedings of the Seventh International Conference (KR2000), San Francisco, CA*. Morgan Kaufman, 2000.

[2] H. Chalupsky, E. Hovy, and T. Russ. Progress on an automatic ontology alignment methodology, 1997. ksl-web.stanford.edu/onto-std/hovy/index.htm.

[3] FGDC. Spatial data transfer standard. Federal Geographic Data Committee, U. S. Geological Survey. Proposed standard, 1998.

[4] W. E. Grosso, H. Eriksson, R. W. Fergerson, J. H. Gennari, S. W. Tu, and M. A. Musen. Knowledge modeling at the millennium (the design and evolution of Protege-2000). Technical report, Stanford University, Institute for Medical informatics, Stanford, CA, 1999. Technical Report SMI-1999-0801.

[5] E.H. Hovy. Combining and standardizing large-scale, practical ontologies for machine translation and other uses. In *Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC). Granada, Spain*, 1998.

[6] International Hydrographic Organization. IHO transfer standards for digital hydrographic data, edition 3.0, 1996.

[7] C. A. Knoblock and J. L. Ambite. Agents for information gathering. In J. M. Bradshaw, editor, *Software Agents*, chapter 16. MIT Press, Cambridge, MA, 1997.

[8] Elbert S. Maloney. *Chapman Piloting: Seamanship and Boat Handling*. Hearst Marine Books, New York, 63rd edition, 1999.

[9] R. M. Malyankar. Acquisition of ontological knowledge from canonical documents. In *IJCAI-2001 Workshop on Ontology Learning*. Seattle, WA, 2001.

[10] D. McGuiness, R. Fikes, J. Rice, and S. Wilder. An environment for merging and testing large ontologies. In *Proceedings of the Seventh International Conference on Principles of Knowledge Representation and Reasoning (KR2000), Breckenridge, Colorado*, April 2000. Tech. report KSL-00-16, Knowledge Systems Laboratory, Stanford University.

[11] National Oceanic and Atmospheric Administration. Chart no. 1: Nautical chart symbols, abbreviations, and terms, 1997.

[12] N. F. Noy and M. Musen. SMART: Automated support for ontology merging and alignment. In *Twelth Workshop on Knowledge Acquisition, Modeling, and Management, Banff, Canada*, 1999.

[13] N. F. Noy and M. A. Musen. PROMPT: Algorithm and tool for automated ontology merging and alignment. Technical report, Stanford University, Institute for Medical informatics, Stanford, CA, 2000. Technical Report SMI-2000-0831.