Using Core Ontology for Domain Lexicon Structuring

Rita Marinelli* **Adriana Roventini*** **Giovanni Spadoni****

*Istituto di Linguistica Computazionale, C.N.R.
Area della Ricerca Via Moruzzi 1, 56124 Pisa Italy
e-mail: Rita.Marinelli@ilc.cnr.it – Adriana.Roventini@ilc.cnr.it
**President & C.E.O. S. Spadoni s.r.l. Shipping Agency, Livorno, Italy
g.spadoni@saurospadoni.it

Abstract

The users' demand has determined the need to manage the growing new technical maritime terminology which includes very different domains such as the juridical or commercial ones. A terminological database was built by exploiting the computational tools of ItalWordNet (IWN) and its lexical-semantic model (EuroWordNet).

This paper concerns the development of database structure and data coding, relevance of the concepts of 'term' and 'domain', information potential of the terms, complexity of this domain and detailed ontology structuring recently undertaken and still in progress.

Our domain structure is described defining a core set of terms representing the two main sub-domains specified in 'technical-nautical' and 'maritime transport' terminology. These terms are sufficiently general to be the root nodes of the core ontology we are developing. They are mostly domain-dependent, but the link with the Top Ontology of IWN remains, endorsing either general and 'foundation' information, or detailed description directly connected with the specific domain. Through the semantic relations linking the synsets, every term 'inherits' the top ontology definitions and becomes itself an integral part of the structure. While codifying a term in the maritime database, the reference is at the same time allowed to the Base Concepts of the terminological ontology embedding the term in the semantic network, showing that upper and core ontologies make it possible for the framework to integrate different *views* on the same domain in a meaningful way.

1. Introduction

Our research was originated by a request on the part of specialized users for a terminological maritime dictionary written in Italian but referred to the English language.

The users' demand has determined the need to manage the growing new technical terminology which also includes very different domains such as the juridical or the economic one.

By exploiting the computational instruments of ItalWordNet (Roventini *et als.*, 2003) and its lexical-semantic model EuroWordNet (Vossen, 1999), a terminological database was built containing about 2,500 synsets (Marinelli *et als.*, 2003). This allowed us to overcome the concept of 'dictionary', and obtain data not only described (by definition), but also codified (by semantic relations), managed automatically in an easy manner and linked to the corresponding closest concepts in English through the Inter-Lingual Index (ILI).

2. Types of Relations

We started to design the top level terminological data base, identifying the most relevant and representative domain concepts, taking into account terms:

1) belonging either to the generic or to the specialized lexicon.
2) having a large number of hyponyms.
3) significant (only) in that knowledge field.

The next step was a top-down development process defining and coding more specific concepts to populate and enlarge the database.

The lexical semantic relations provided by the EWN/IWN model were employed. The **Internal relations** allow information encoding in the form of lexical-semantic relations between pairs of synsets[1]. Synonymy and hyponymy are the most important relations encoded; this linguistic model is very rich and contains many other lexical-semantic relations such as *part of*, *cause*, *purpose*, *sub-event*, *belong-to-class*, etc.

The **Equivalence relations** link the Italian synsets with the closest concepts (synonyms, near synonyms, etc.) in the Inter-Lingual Index (ILI), a separate language-independent module containing all WN1.5 synsets but not the relations among them. The possibility to use IWN and the terminological DB for multilingual applications is ensured by this link.

The **Plug-in relations** allow to connect the specialized wordnet to the generic one linking a terminological sub-hierarchy (represented by its root node) to a node of the generic wordnet.

[1] A *synset* is defined as set of synonymous words belonging to the same Part-of-Speech (PoS) that can be interchanged at least in a context.

Up until now our database has been connected, by means of the *plug_in* relations, to the general database and to the Top Ontology (TO) which IWN inherited from EuroWordNet (Marinelli *et als*., 2004).

We outline a new domain ontology design, to better define the boundary of this research, setting the grounds of the terminological concepts and gaining more functional information.

3. The Concept of Term

Before defining the ontology, a reflection is necessary about the concept of 'term', the 'relevance' of each term, the knowledge potential of the terminological lexicon, together with the possibility of manipulating this knowledge with considerable cognitive effects, specifying how to represent it as a concrete (suitable to be instantiated) data structure.

From the cognitive point of view the meaning potential of a term can be explained by the importance it has as input that satisfies our expectations of relevance.

The search for relevance is a basic feature of human cognition, which communicators may exploit, improving their knowledge on a certain topic.

According to relevance theory, an input is relevant to an individual when its processing in a context of available assumptions yields a positive cognitive effect. (Wilson and Sperber, 2002).

The terms are a means of knowledge information; actually, linguistics, philosophy and the technical-scientific disciplines consider terminology as a 'conjunction' of units with an essential aim, and, therefore, with a functional value (Cabré, 2000). In the different applications a twofold function of the terminological units is activated: the specialized knowledge representation and its conveyance. The terms are used in specialized communication, characterized by linguistic and pragmatic factors: pragmatics studies how the meaning potentials are completely specified and actually used by the speakers; terms deserve a new more dynamic approach, considering that meaning is not only 'content', but a way to change the state of information of the speakers (Chierchia, 1997).

Specialized communication admits different levels of specialization, various degrees of knowledge opacity, several indexes of cognitive and terminological density and distinct aims; and taking this into account means considering the terms with all the meaning and knowledge potential they can have (Cabré, 2003).

4. Domain Complexity

The second consideration about specialized lexicon structuring is concerned with the nature and structure of domain.

Domains may be more or less specific, more or less tangled (Poli, 2002). The maritime domain also includes many other knowledge fields ranging from meteorology to astronomy, from law and maritime contracts to transport technology. The detailed structuring of a context of analysis with respect to its sub-domains is a very complex task. As a matter of fact, in our lexicon we find different levels of specificity depending both on the hierarchical structure of taxonomies and on the many lexical items coming from various disciplines strictly connected with maritime navigation and maritime transport. Therefore, we thought it was necessary to describe our domain structure taking this complexity into account. To this purpose a comprehensive set of basic concepts is required, organized so as to admit the existence of different possible pathways among sub-domains under a common conceptual framework (Gangemi, 2005).

5. Structure Description

Our domain structure is described defining a 'core' set of concepts representing the main two sub-domains specified in maritime terminology: *technical/nautical* (nautics) and *maritime transport* (transport) domain. These are sufficiently general to be the root nodes of the core ontology we are developing: they have to be supported by specialized documentation and studied by ontological engineers and domain experts working in close collaboration (Marinelli and Roventini, 2005). The goal of core ontology is to provide a global and extensible model into which data originating from diverse sources can be mapped and integrated.

5.1 The Core Concepts Identification

Identification of the most relevant concepts has been carried out following two different criteria. As far as the technical/nautical terminology is concerned, we considered the Glossary edited by the Harbour Master of Leghorn (Tuscany) and the Italian Navigation Code, as a starting point for choosing the most recurring and significant concepts and laying down a first categorization: the most interesting and representative patterns e.g. equipment (*attrezzatura*), direction (*governo*), steering (*conduzione*), etc., were highlighted, each one incorporating a set of related concepts into which it is divided.

On the other hand, with regard to maritime transport, the various stages of the 'import/export' operation process were singled out, e.g. loading (*operazioni di carico*), stowage *(stivaggio)*, embarkment *(imbarco)*, freight rating *(tassazione)*, etc., that are the main phases of the path that it is necessary to follow so that a cargo (goods or passengers) can actually be transported to its destination. Different 'perspectives' lying across this domain, e.g. juridical or commercial, were also distinguished. For each of these phases and perspectives a representative concept was chosen to be developed and considered as a node to be fleshed out within its own framework.

When it was possible we used official reference criteria or standards for high level classification.

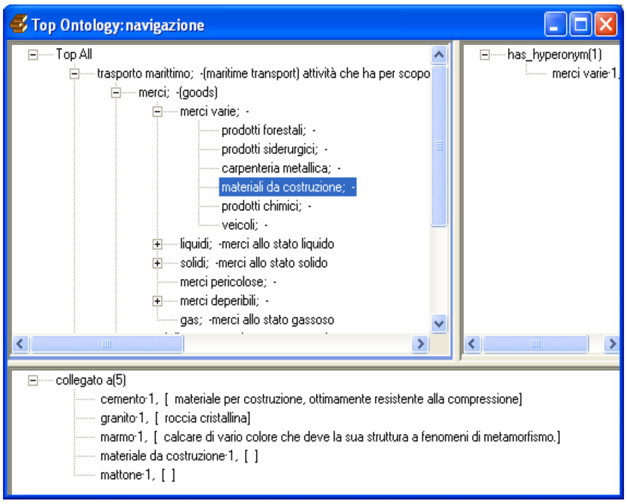
5.2 Knowledge Representation

As a first step, it was important to produce a comprehensive list of the most salient terms for the beginning of our work. Therefore, we started with the definition of the most general concepts in the domain and the subsequent specialization of the concepts (top-down development process), defining the most specific concepts, and then grouping them under more general concepts (bottom-up development). This 'combination' approach may be considered the easiest way of developing an ontology, since the concepts 'in the middle' tend to be the most descriptive concepts in the domain (Rosch, 1978).

Therefore we started with a few top-level concepts such as 'goods' (*merci*), and several specific concepts used for goods classifying, for example, 'materiali da costruzione' (construction materials) and 'prodotti chimici' (chemical products). Afterwards we related them to a middle-level concept, such as 'merci varie' (general cargo).

We then divided and distributed the many types (hyponyms) of construction materials or chemical products, by structuring a number of middle-level concepts and their hyponyms at various levels.

In the Fig. 1 the concept 'materiali da costruzione' (construction materials) is highlighted and there is evidenced the link either with the hyperonym 'merci varie' (general cargo) or with the hyponyms, e.g.: 'cemento' (cement), 'granito' (granite), etc.



```
⊟ Top All
   ⊟ trasporto marittimo; -(maritime transport) attività che ha per scopo
      ⊟ merci; -(goods)
         ⊟ merci varie; -
               prodotti forestali; -
               prodotti siderurgici; -
               carpenteria metallica; -
               materiali da costruzione; -
               prodotti chimici; -
               veicoli; -
         ⊞ liquidi; -merci allo stato liquido
         ⊞ solidi; -merci allo stato solido
           merci pericolose; -
         ⊞ merci deperibili; -
           gas; -merci allo stato gassoso

⊟ has_hyperonym(1)
     merci varie·1,

⊟ collegato a(5)
     cemento·1, [ materiale per costruzione, ottimamente resistente alla compressione]
     granito·1, [ roccia cristallina]
     marmo·1, [ calcare di vario colore che deve la sua struttura a fenomeni di metamorfismo.]
     materiale da costruzione·1, [ ]
     mattone·1, [ ]
```

Fig. 1 'Materiali da costruzione' *(construction materials)*

In the set of core concepts proposed to represent our domain ontology, a non rigid categorization methodology was used. Regarding the concept 'merci' (goods), for instance, a classification is shown, in the figure above, which takes into account the criteria used by the Leghorn Port Authority. However, it is possible to insert different types of classification, e.g. the reference codes used by ISTAT, Istituto Nazionale di Statistica (National Statistics Institute), so that both classification systems can be provided for helping and supporting, e.g.:

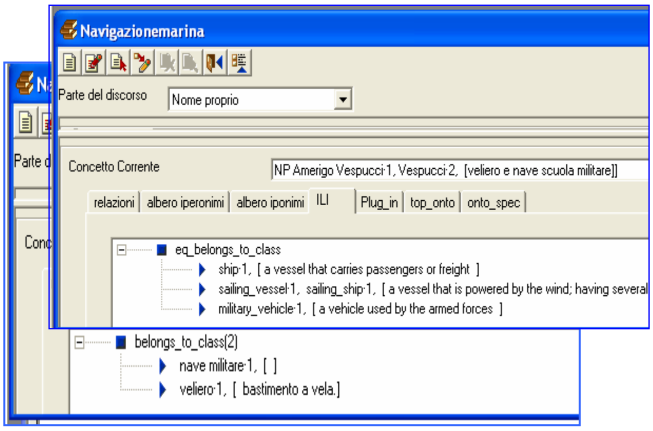
autoveicoli nuovi e usati (new and second-hand motor vehicles) *has hyperonym* materiali da trasporto (transport materials) (according to ISTAT);

autoveicoli nuovi e usati (new and second-hand motor vehicles) *has hyperonym* veicoli (vehicles) (according to Port Authority).

In defining a class hierarchy, we have to consider that the ontology should not contain all the possible information about the domain, but should try to guarantee consistency. Another example of the issue of establishing the criteria for classification is, the concept 'nave' (ship): it has a large number of hyponyms but, as we said above, they can be classified from different points of view: on the basis of the type of propulsion (oars, sails, propeller), of the use for which they were built (transport of goods or passengers, competitions, war operations, etc.), of the place where they move (river, lake, sea).

In this maritime domain, for example, we could classify the concept 'ship' into military, passengers, or cargo ships, considering the different kinds of uses. Alternatively, from a different point of view, we could divide the concept of ship into sailing or propeller ships.

As allowed by most knowledge-representation systems, multiple inheritance in the concepts hierarchy is represented: a concept can be a 'sub concept' of more than one concept. For example, if in the domain specific ontology we defined the two separate classes of sailing vessels and military ships, the *Vespucci* would inherit both concepts as it is both a sailing vessel and a military ship, as shown in Fig. 2 and in Fig. 3:



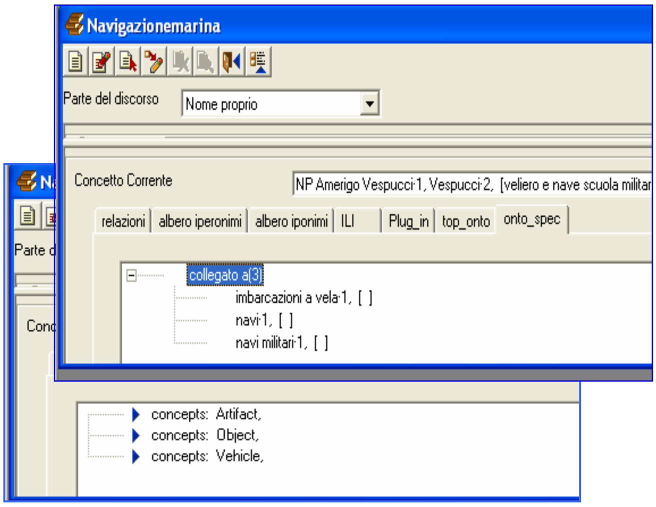
Parte del discorso: Nome proprio

Concetto Corrente: NP Amerigo Vespucci·1, Vespucci·2, [veliero e nave scuola militare]]

```
⊟ eq_belongs_to_class
     ship·1, [ a vessel that carries passengers or freight ]
     sailing_vessel·1, sailing_ship·1, [ a vessel that is powered by the wind; having several
     military_vehicle·1, [ a vehicle used by the armed forces ]

⊟ belongs_to_class(2)
     nave militare·1, [ ]
     veliero·1, [ bastimento a vela.]
```

Fig. 2 Internal and Equivalence Relations



Parte del discorso: Nome proprio

Concetto Corrente: NP Amerigo Vespucci·1, Vespucci·2, [veliero e nave scuola militar

```
⊟ collegato a(3)
     imbarcazioni a vela·1, [ ]
     navi·1, [ ]
     navi militari·1, [ ]

     concepts: Artifact,
     concepts: Object,
     concepts: Vehicle,
```

Fig. 3 Link with IWN TO and with Domain Ontology

Another matter of concern is the treatment of concepts such as 'acido' (*acid*), which is a hyponym of 'substance' in the generic database, but also a type of dangerous goods, depending on the ‘focal orientation’ in which the is-a relationship is viewed.

Among the several viable alternatives, it is a matter of deciding which one could work better for the planned task, or would be more intuitive, more extensible, and more maintainable: an ontology is a model of reality of the world and the concepts in the ontology must reflect this reality (Friedman Noy and McGuinness, 2001).

6. Modeling the Domain Structure

Our analysis and modelling processes are guided by domain-independent criteria and relations, i.e. by the upper ontology provided by the IWN model. The model of this structure, like the database, is WN-like: the most important relations are the (vertical/hierarchical) is-a relations and among the ‘horizontal’ relations, a subset is exploited (*is means, for purpose, causes*) and also integrated with new semantic relations more suitably tailored to the specific needs (*event location, event time*), e.g.:

Fiera Internazionale della Nautica (International Nautics Fair) *event location* Genova

Fiera Internazionale della Nautica (International Nautics Fair) *event time* 20-28 Novembre 2005.

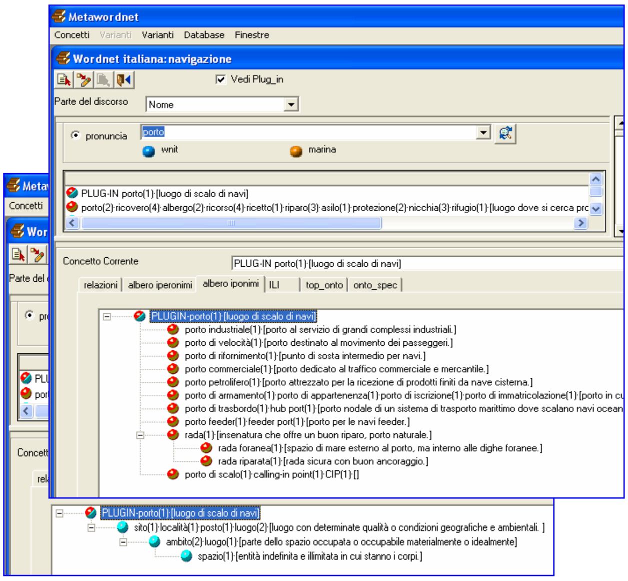
As a matter of fact, in the maritime domain, there are a large number of terms to be coded as ‘events’, therefore new semantic relations are necessary to represent the space-temporal dimension in a more exhaustive manner. A study is in progress to examine this subject in more detail.

The subset of the above-mentioned relations seems to be the most appropriate to characterize the main conceptual schemas that people of the technical-nautical or maritime transport ‘world’ actually use, namely activity plans, programs involving particular devices for cargo stowage, shipping goods, navigation management, etc. (Marinelli and Spadoni, 2006).

While the core concepts are mostly domain dependant, the link with the Top Ontology of IWN continues to exploit the *plug-in* relations: in this way it is possible to guarantee either general and ‘foundation’ information, or detailed information directly connected with the specific domain. The tool we are using to build the specific ontology also allows an ‘integrated’ consultation of the terminological database, highlighting that every term ‘inherits’ the IWN Top Ontology definitions and becomes an integral part of the structure; at the same time, while codifying a term in the maritime database, reference to the concepts of the domain ontology is allowed, embedding the term in the terminological network.

We can consider a concept in the generic database and the relative concepts linked by the semantic relations as an anchor point. We wish to give prominence and development to these concepts within the appropriate framework of the terminological database. In this way the generic database IWN is really not only a theoretical but also a pragmatic model, giving the drift about the (initial) paths to follow each time a concept is to be improved, enlarged and refined in the terminological database. In this development process it is likely to find concepts in the terminology that are not present in IWN but worth including in the database; or it may be necessary to codify concepts or missing relations, proceeding by approximation and considering the results step by step, exploiting the most effective and operative aspects of the two databases, in a dynamic process aimed at knowledge management and interoperability.

The example of ‘porto’ (harbour) is shown hereafter as it appears in the integrated consultation of the tool:



☑ Vedi Plug_in

Parte del discorso: Nome

● pronuncia: porto

wnit   marina

PLUG-IN porto(1)·[luogo di scalo di navi]

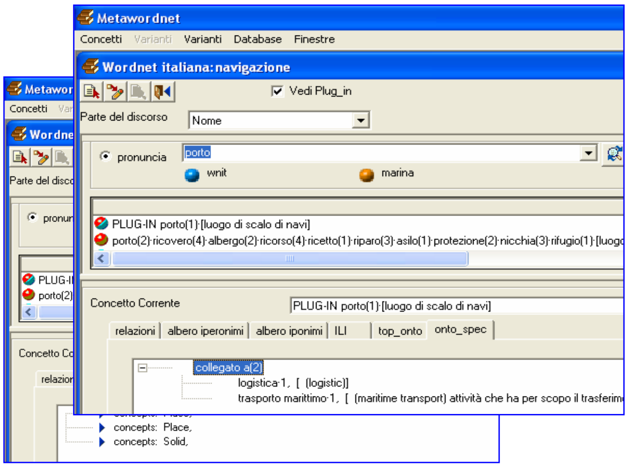
porto(2)·ricovero(4)·albergo(2)·ricorso(4)·ricetto(1)·riparo(3)·asilo(1)·protezione(2)·nicchia(3)·rifugio(1)·[luogo dove si cerca prc

Concetto Corrente: PLUG-IN porto(1)·[luogo di scalo di navi]

```
PLUGIN-porto(1)·[luogo di scalo di navi]
    porto industriale(1)·[porto al servizio di grandi complessi industriali.]
    porto di velocità(1)·[porto destinato al movimento dei passeggeri.]
    porto di rifornimento(1)·[punto di sosta intermedio per navi.]
    porto commerciale(1)·[porto dedicato al traffico commerciale e mercantile.]
    porto petrolifero(1)·[porto attrezzato per la ricezione di prodotti finiti da nave cisterna.]
    porto di armamento(1)·porto di appartenenza(1)·porto di iscrizione(1)·porto di immatricolazione(1)·[porto in cu
    porto di trasbordo(1)·hub port(1)·[porto nodale di un sistema di trasporto marittimo dove scalano navi ocean
    porto feeder(1)·feeder port(1)·[porto per le navi feeder.]
    rada(1)·[insenatura che offre un buon riparo, porto naturale.]
        rada foranea(1)·[spazio di mare esterno al porto, ma interno alle dighe foranee.]
        rada riparata(1)·[rada sicura con buon ancoraggio.]
    porto di scalo(1)·calling-in point(1)·CIP(1)·[]
```

```
PLUGIN-porto(1)·[luogo di scalo di navi]
    sito(1)·località(1)·posto(1)·luogo(2)·[luogo con determinate qualità o condizioni geografiche e ambientali. ]
        ambito(2)·luogo(1)·[parte dello spazio occupata o occupabile materialmente o idealmente]
            spazio(1)·[entità indefinita e illimitata in cui stanno i corpi.]
```

Fig. 4 Downward and upward relations



☑ Vedi Plug_in

Parte del discorso: Nome

● pronuncia: porto

wnit   marina

PLUG-IN porto(1)·[luogo di scalo di navi]

porto(2)·ricovero(4)·albergo(2)·ricorso(4)·ricetto(1)·riparo(3)·asilo(1)·protezione(2)·nicchia(3)·rifugio(1)·[luogo

Concetto Corrente: PLUG-IN porto(1)·[luogo di scalo di navi]

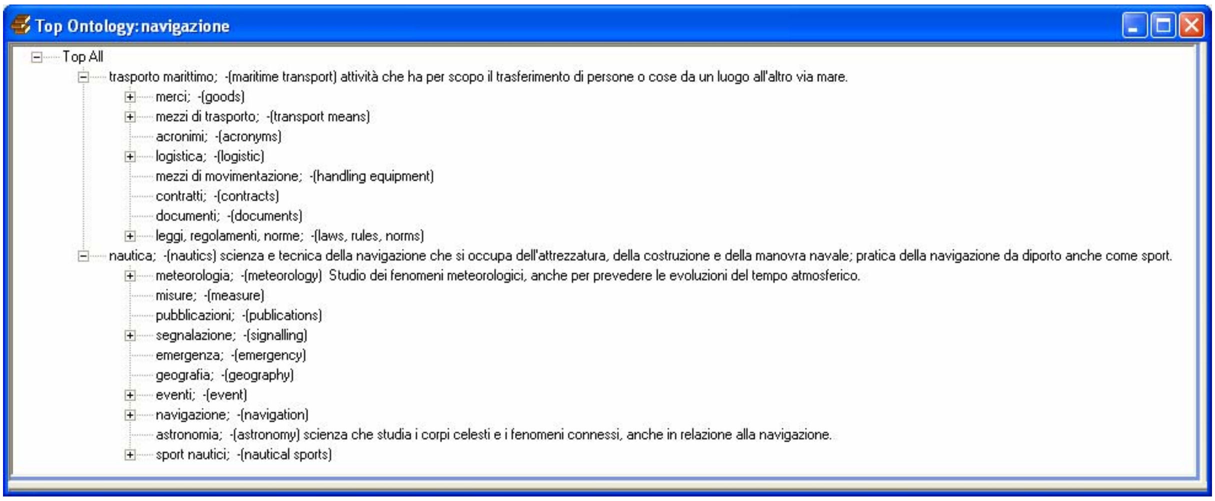
```
collegato a(2)
    logistica·1, [ (logistic)]
    trasporto marittimo·1, [ (maritime transport) attività che ha per scopo il trasferim
```

```
    concepts: Place,
    concepts: Solid,
```

Fig. 5 The links to Domain Ontology and IWN TO

Upper and core ontologies allow the framework to integrate different ‘views’ on the same domain in a meaningful way. We describe our domain structure taking into account the need to manage the ever increasing new technical terminology. We define a common body of represented knowledge for users who need to share information in this domain, for professionals and not professionals who wish to reuse domain knowledge, to clarify and separate domain and operational knowledge.

What follows is the 'core' set of concepts representing the main two sub-domains specified in maritime terminology:



```
⊟ Top All
    ⊟ trasporto marittimo;  -(maritime transport) attività che ha per scopo il trasferimento di persone o cose da un luogo all'altro via mare.
        ⊞ merci;  -(goods)
        ⊞ mezzi di trasporto;  -(transport means)
          acronimi;  -(acronyms)
        ⊞ logistica;  -(logistic)
          mezzi di movimentazione;  -(handling equipment)
          contratti;  -(contracts)
          documenti;  -(documents)
        ⊞ leggi, regolamenti, norme;  -(laws, rules, norms)
    ⊟ nautica;  -(nautics) scienza e tecnica della navigazione che si occupa dell'attrezzatura, della costruzione e della manovra navale; pratica della navigazione da diporto anche come sport.
        ⊞ meteorologia;  -(meteorology)  Studio dei fenomeni meteorologici, anche per prevedere le evoluzioni del tempo atmosferico.
          misure;  -(measure)
          pubblicazioni;  -(publications)
        ⊞ segnalazione;  -(signalling)
          emergenza;  -(emergency)
          geografia;  -(geography)
        ⊞ eventi;  -(event)
        ⊞ navigazione;  -(navigation)
          astronomia;  -(astronomy) scienza che studia i corpi celesti e i fenomeni connessi, anche in relazione alla navigazione.
        ⊞ sport nautici;  -(nautical sports)
```

Fig. 6 The core concepts representing the two sub-domains specified in maritime terminology

Most of the concepts that are shown in Fig. 6 are in the plural form: our objective is to indicate, in this way, the whole category, the whole set extensionally represented by its elements. The plural is also a heritage from the (Italian and English) glossaries lexicon and, most of all, from the 'Italian Navigation Code', where the titles of the main paragraphs and subjects dealt with generally contain words in the plural, e.g.: 'navi e galleggianti' (ships and floats) (op. cit.: 30), 'categorie della gente di mare' (sea people categories) (op. cit.: 60), 'zone portuali' (harbour zones) (op. cit.: 65), etc.

7. Final Remarks

Multiple inheritance, categories without rigid boundaries, a categorization performed from different points of view together with the possibility of being enlarged and provided with new details, make the system a flexible dynamic structure, where no account is suggested of absolute levels of categorization.

Our approach to ontology building is not to create a rigid system with reduced freedom of interpretation, but to admit and navigate alternative interpretations, conceiving different contexts of use which have to be promptly highlighted for effective usefulness granting information integration, interoperability without overlapping, clarified information tested and officially validated.

The aim of our project is to underline the dynamic cognitive processes leading to ontology organization that are strictly connected with dynamic applicative and pragmatic processes, integrated in an iterative (even recursive) enriching, refining, tuning cycle: we view ontology design as a creative process, trying to guarantee not completeness, but consistency (Gruber, 1993) and we can assess its quality by enlarging, testing and refining, and, actually using it (Friedman Noy and McGuinness, 2001).

References

Cabré Castelvì M. T., La terminologia: representacion y comunicacion, Institut Universitari de Linguistica Aplicada, Barcelona, 2000.

Cabré Castelvì M.T., Theories of terminology, Terminology 9:2 (2003), 163-199.

Chierchia G., Semantica, Bologna, Il Mulino, 1997.

Codice della Navigazione e relativi regolamenti, Milano, Giuffrè, 1971.

Croft W., Cruse D. A. , Cognitive Linguistics. Cambridge University press, 2004.

Friedman Noy N., McGuinness D.L., "Ontology Development 101: A Guide to Creating Your First Ontology". Stanford Knowledge Systems Laboratory Technical Report, 2001.

Gangemi, A., Development of an Integrated Formal Ontology and an Ontology Service for Semantic Interoperability in the Fishery Domain, Draft project plan v.7, CNR – Institute of Cognitive Sciences and Technologies, Ontology and Conceptual Modeling Group, 2005.

Gruber, T.R. , A Translation Approach to Portable Ontology Specification. Knowledge Acquisition 5: 199-220, 1993.

Marinelli R., Roventini A., Spadoni G. , Linking a subset of maritime terminology to the Italian WordNet. Proceedings of the Third International Conference on Maritime Terminology, Lisbon, 2003.

Marinelli R., Roventini A., Enea A. , Building a Maritime Domain Lexicon: a Few Considerations on the Database Structure and the Semantic Coding. LREC

2004: Fourth International Conference on Language Resources and Evaluation, held in Memory of Antonio Zampolli. Lisbon, Portugal, 26th, 27th & 28 May 2004, Proceedings, Volume II, Paris, The European Language Resources Association (ELRA), 465-468, 2004.

Marinelli R., Roventini A., Some Considerations about the Italian Maritime Lexicon Structuring. In Proceedings of the IX Simposio Internacional de Comunicación Social, Santiago de Cuba, 24-28 de Enero de 2005. 635-639, 2005.

Marinelli R., Spadoni G., Some Considerations in Structuring a Terminological Knowledge Base. In Proceedings of the Third International WordNet Conference, GWC 2006, (forthcoming).

Poli R., Ontological Metodology, in Human Computer Studies, (2002), 56, 639-664.

Rosch, E. , Principles of Categorization. Cognition and Categorization. R. E. and B. B. Lloyd, editors. Hillside, NJ, Lawrence Erlbaum Publishers: 27-48, 1978.

Roventini, A., Alonge, A., Bertagna, F., Calzolari, N., Cancila, J., Girardi, C., Magnini, B., Marinelli, R., Speranza, M., Zampolli, A.: ItalWordNet: Building a Large Semantic Database for the Automatic Treatment of Italian, in: «Linguistica Computazionale», vol. XVI-XVII (2003), pp.745-791, Giardini, Pisa.

Roventini A., Marinelli R., Extending the Italian WordNet with the Specialized language of the Maritime Domain. Proceedings of the Second International WordNet Conference, GWC 2004, 193-198.

Vossen, P. (ed.): EuroWordNet General Document, 1999.http://www.hum.uva.nl/~EWN

Wilson D., Sperber D., Relevance Theory. UCL Working Papers in Linguistics 2002. 14: 249-287. Also published in G. Ward and L. Horn (eds.) Handbook of Pragmatics. Oxford: Blackwell, 607-632. Available at http://www.phon.ucl.ac.uk/home/deirdre/