Adding information to a terminological database by means of image files

Rita Marinelli
Istituto di Linguistica Computazionale C.N.R.
Via Moruzzi 1 Pisa, Italy
rita.marinelli@ilc.cnr.it

Giovanni Spadoni
S. Spadoni s.r.l. Shipping Agency Via delle Cateratte
90 Livorno, Italy
g.spadoni@saurospadoni.it

Sebastiana Cucurullo
Istituto di Linguistica Computazionale C.N.R.
Via Moruzzi 1 Pisa, Italy
nella.cucurullo@ilc.cnr.it

Abstract

A lexical semantic database containing terms belonging to the specialized lexicon of the maritime navigation and maritime transport was built according to WordNet/EuroWordNet model. Our paper present a project planning the enrichment of the terminological database by means of a set of images. A short description is given about a) the structure of the terminological database and the domain conceptual modelling; b) the various features of the database management tool, and, among all, the possibility of visualizing, on demand, the image which is associated with the term being sought, contributing to clarify and refine the meaning of the term, increasing its information and communication effectiveness.

1 Introduction

MariTerm is a database structured according to the EuroWordNet/ItalWordNet model, in the frame of the WordNet philosophy: the relational structure of the database is of lexical semantic type. An approximate 4000 lemmas are codified, which belong to the specialized lexicon of the technical-nautical and maritime transport domain (Marinelli and Roventini, 2006).

The objective of this project was to create a terminological resource that could be a support for management of the terms belonging to this domain that are used with an increasing frequency in spoken and written texts and, in general, in everyday life. Our study was guided by the need for a useful instrument for work and didactic activities and, in general, for various types of communication contexts. This paper presents a research recently undertaken and still in progress, which focuses on the improvement of the maritime database by providing visual information by means of a set of images.

In the following sections we describe: a) the structure of the terminological database and the domain conceptual modelling; b) the database management tool which allows consultation of the terminological database, updating of the set of data and, among the various features, visualization on demand of the image which is associated with the term being sought; c) final remarks and conclusions.

2 The Database Structure

The relational structure of the database provided by the model is represented in terms of:

a) Internal relations: which link synsets (sets of synonyms [1]) in hierarchical relationship (vertical relations), by means of hyperonymy/hyponymy relations, or in meronymy, entailment, role, etc. relationship (orizontal relations): the use of vertical (hyperonymy/hyponymy) relations leads to the definition of the most basic level of categorization namely "the most inclusive (abstract) level at which the categories can mirror the structure of attributes perceived in the world" (Rosch, 1988), while the use of the horizontal dimension for categorization implies the improvement of the distinctiveness and flexibility of categories.

Each synset is ontologically classified, on the basis of its hyperonym, in terms of the IWN Top Ontology (TO), i.e. a hierarchy of language-

[1] In the latest version of WordNet, (WN 3.0), "synset" is defined as "a set of one or more 'synonyms' or 'variants'", e.g.: *imbarcazione, natante* (vessel), *naufragare, colare a picco, affondare* (to sink).

independent concepts reflecting essential semantic distinctions, e.g.: *navigazione* (navigation) → Agentive, Dynamic, Purpose.

b) Equivalence relations: connect the Italian synsets with the closest concepts (synonyms, near synonyms, hyperonyms, etc.) of the Inter Lingual Index (ILI [2]). When possible an eq_synonym or eq_near_synonym relation is used, otherwise an eq_has_hyperonym relation is coded, e.g.:

nolo eq_synonym *freight*
nolo prepagato eq_has_hyperonym *freight*

by these links to the ILI, the terms are also connected to the Top Ontology (TO).

c) Plug-in relations: allow the linking of a synset of the specialized wordnet to the generic (IWN), (e.g.: "*porto*" is present in both the databases); in such a way a terminological sub-hierarchy (represented by its root node) is connected to a node of the generic wordnet. By means of the plug-in relations the tool we are using to manage the terminological database and the specific ontology also allows an "integrated" consultation of the database; it shows that if a synset is found in both databases (and is plugged-in), the synset belonging to the specific domain partially "obscures" the generic one: downward (hyponymy) and horizontal relations (part_of relations, role relations, etc.) are taken from the terminological wordnet, while upward (hyperonymy) relations are taken from the generic one.

3 Domain Structuring

In the integrated consultation a term is plugged in its hyperonym or synonym in the generic lexicon and the link with the upper part of the taxonomic chain can be shown visualized by the tool. Since the top ontology of a concept in the database is fully defined through its hyperonym, it is possible to see the highest concepts (TO) of the generic network to which the term is connected.

We deemed necessary to provide the terms with a specific ontology to better complete and support the functional value of terms as means of knowledge information. Following Cabré (2000), the terminological units have a double function: the specialized knowledge representation and its conveyance. The importance of a term is assessed according to the place it has in the conceptual structure of a domain following precise criteria.

The domain structure was outlined designating a "core" set of concepts which represent the two main sub-domains specified in maritime terminology: technical/nautical (nautics) and maritime transport (transport) domain and the various disciplines embraced by maritime domain. They range from astronomy to geography, from transport logistics to meteorology.

A comprehensive set of basic concepts was worked out and organized by the suggestions of ontological engineers and domain experts (Marinelli et al., 2006) so as to constitute the hook up points of the domain modelling, admitting the existence of different possible pathways among sub-domains under a common conceptual framework (Gangemi, 2005).

Two different criteria were followed to distinguish the most relevant concepts: i) for the technical/nautical terminology, we used the Glossary edited by the Harbour Master of Livorno (Tuscany) and the Italian Navigation Code, as a starting point for choosing the most frequently recurring and significant concepts and laying down a first categorization: the most interesting and representative patterns e.g. *attrezzatura* (equipment), *governo* (direction), *conduzione* (steering), etc., each incorporating a set of related concepts into which it is divided, were highlighted; ii) for maritime transport, the various stages of the "import/export" operation process were singled out, e.g. *operazioni di carico* (loading), *stivaggio* (stowage), *tassazione* (freight rating), etc., which are the main phases of the path necessary to follow so that a cargo (goods or passengers) can actually be transported to its destination. A representative concept was designated for each of these phases and perspectives and it was considered as a node to be fleshed out and developed within its own framework. When it was possible, we exploited official reference criteria or standards for high level classification, namely the criteria used by the Leghorn Port Authority and the codes used by ISTAT, *Istituto Nazionale di Statistica* (National Statistics Institute) for the classification of goods.

The definition of the criteria for classification is a crucial issue: the concept "*porto*" (harbour/port), for example, has many hyponyms but they can be classified from different points of view: with reference to harbour location (lake, sea, river), or to the specific use (commercial, industrial,

[2] An unstructured version of WordNet 1.5, containing all its synsets but not the relations among them.

military), or to the logistic services offered (*rifornimento*/ bunkering, *immatricolazione*/ registry) (Marinelli and Spadoni, 2007).

Each term is connected to one or more domain dependent concepts belonging to this "core" set; at the same time, the plug in relations described above bridge the term to the TO of IWN. The knowledge of a term is assured from both a general, foundation perspective and a specialized point of view, directly connected with the specific knowledge field. In the "integrated" consultation of the terminological database, the tool shows that every term can "inherit" the IWN Top Ontology definitions thus becoming an integral part of the structure; 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. The example of "*porto*" (harbour) is shown hereafter as it appears in the integrated consultation of the tool:

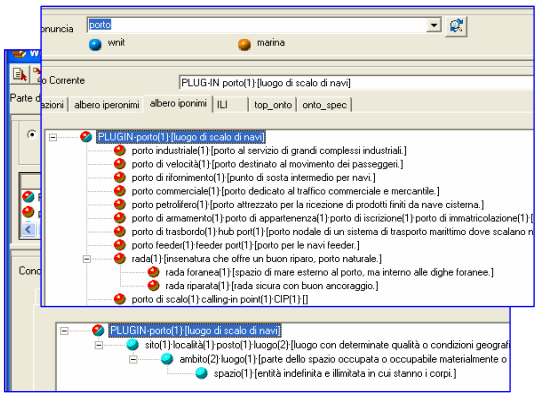

Figure 1. Downward and upward relations

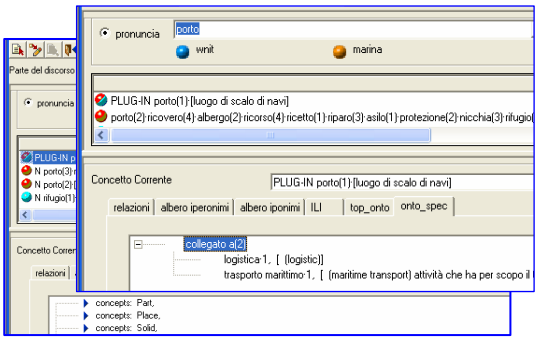

Figure 2. The links to the Domain Ontology and IWN TO

4 Increasing Lexical Coverage

The database has been enriched increasing its lexical coverage with a set of acronyms and abbreviations used in particular work environments (research, medicine, and, especially, transport) where the English language is very much used or even prevails. We mean acronyms like ASAP (As Soon As Possible), AGW (All Going Well), or WP (Weather Permitting), SHEX (Sunday Holidays Excluded). They hardly ever appear in literary texts, in newspapers, or in spoken language, but are included in the jargon of every day conversation belonging to the import/ export world and in maritime terminology in general. Universally recognized, they are fundamental and necessary for informal e-mail communications, for actual effective economy purposes. A set of proper names has also been added to the database, representing the most important ports and well known national and international Transport organizations. A group of terms belonging to maritime meteorology has also been codified: among the knowledge fields that are included in the maritime domain, Meteorology has a particular relevance. In fact, weather forecasts' accuracy makes it possible to plan the most "economical" and safest routes, in order to maintain the scheduled "transit time" between ports, to program cargo operations minimizing idle time and consequent costs, due to bad weather conditions. The weather component plays a significant role in maritime contracts as, e.g., the Expected Time of Arrival (ETA) for a ship into a port is always computed "Weather Permitting" (WP) and the calculation of the "lay time", the maximum time that the maritime contract assigns to perform the cargo operations, is always based on a fixed number of "Weather Working Days" (WWD) (Marinelli and Spadoni, 2007).

5 Enriching the Database with Images

In the last months, also the tool for the system management has been improved with new capabilities: it is possible to visualize the image illustrating the term being sought. An archive contains a set of images stored in such a way that a link can be created between each term of the database and the corresponding image. The archive can be updated by adding new files or replacing the old images with more recent ones. In this way, the information potential of the synset visualized by the database management tool is enriched and the imagery of the user is "guided" to the yielding of a more adequate knowledge of the term, abreast with the times:

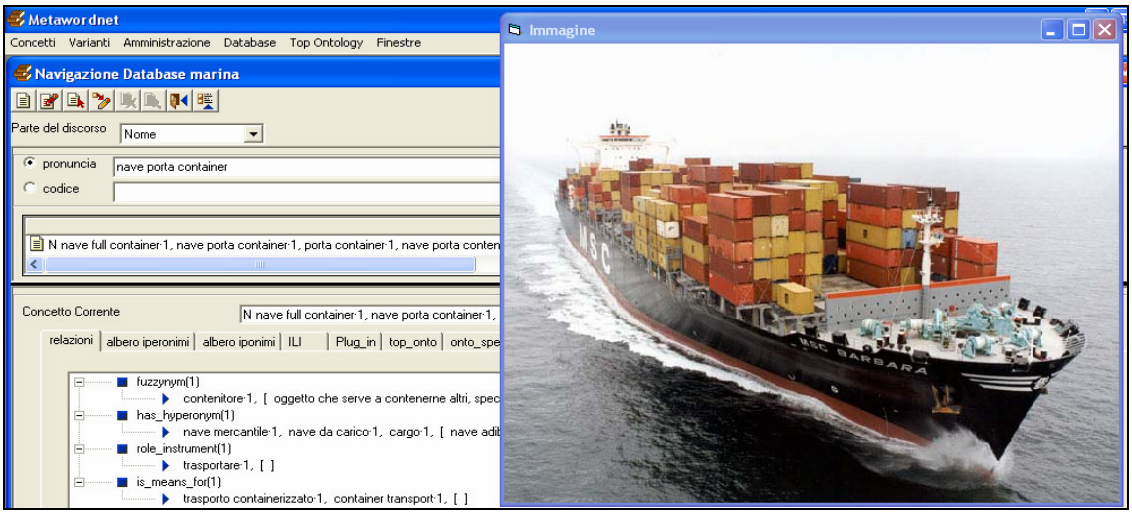

Figure 3. Example for "*nave porta-container*" (full container ship)

The database management tool allows connection on demand of the image selected to the term required: the image is activated by clicking on a button in the tool interface and can be substituted or even erased if not useful.

The new tool potential can be exploited for the sub-domain of "maritime transport" as well- which already includes a section "documentation"- inserting images of the standard documents that are used in the various phases of the logistic chain (standard charter-parties and bill of lading forms, documents of transport, international custom forms, etc.). The capability of the system can be boosted by making available to the users the immediate reference to the templates of the documents, whose definition is normally considered not sufficient for a full understanding of the processes involved.

A set of terms (about 150) that are the most frequent in the Maritime Corpus available in the Institute of Computational Linguistics [3] was considered the starting point of our research. They are representative of this domain and images could be easily obtained. These terms belong either to the generic or to the specialized lexicon and have a large number of hyponyms which are relevant to this knowledge field.

The images collected up to now were retrieved from different sources (on the basis of the domain expert's suggestions) [4] : web sites, personal photos, private archives, books and specialized publications, etc. They were also supplied by the Naval Academy of Livorno (Italy), by the CoMMA–Med Laboratory of the Institute of Biometeorology (C.N.R.) and by the Porto Livorno 2000, a passenger terminal managing company.

One or more images were found for each term of this set and stored in the image archive in such a way as to be linked to the identification number (id) of the synset with a one-to-one correspondence. It is possible to choose the most suitable image as example of the term and to compile a file for every image containing organized information: subject (the object represented), the source, the date, the type/kind (photo, video, drawing, etc.), a short description of the object represented, plus technical characteristics (resolution, dimensions, etc.), a field recording an inventory number; some fields will also be provided containing the reference to other related images and a field with one or more words to be used as keywords by the database management system.

The example of "*vela aurica*" (fore and aft sail) and the file/card with all information about the image are shown in the figure 4 and 5 as they appear in the consultation tool.

The content of the image archive can be visualized, saved and printed, as well as the file that contains catalographic information, allowing to determine the accuracy of data by checking them with various sources. In the near future a database including these descriptive catalographic files will be designed so as to be able to support our project. In such a way the set of images, chosen and structured on the basis of precise technical criteria, correspond to the set of terms and to a catalographic database of catalogued descriptive files.

[3] The corpus of maritime terminology, in progress at the ILC, consists of nearly 140,000 occurrences.

[4] A kind of evaluation is also planned.

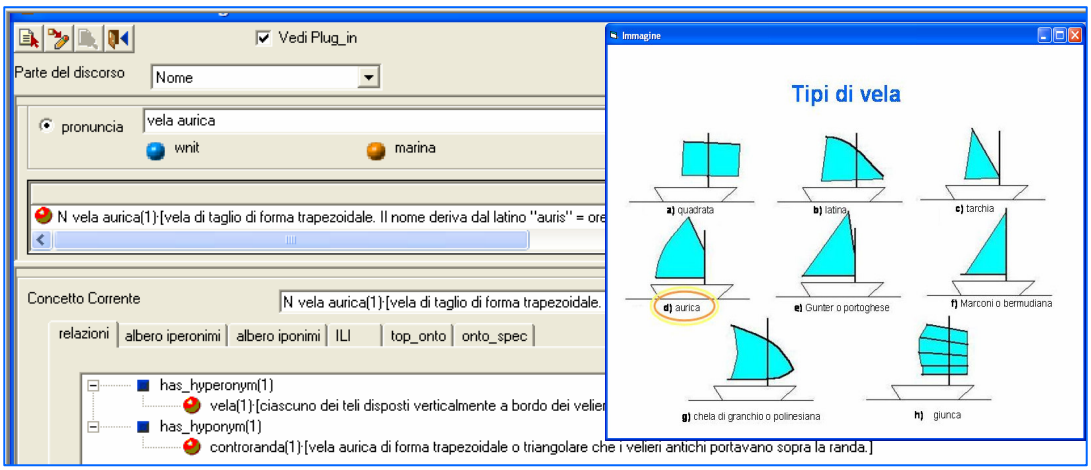

Figure 4. *Vela aurica* (fore and aft sail)

Termine: *Vela aurica*	
SOGGETTO:	Vela aurica
N. Inventario:	21
Fonte:	Accademia Navale di Livorno – Direzione Studi
Data:	18 - 03 - 2009
Tipo:	Disegno schematico
Descrizione:	Immagini di vele di cui una evidenziata
Caratteristiche tecniche:	Originariamente: 720 x 540 – 32,8k – jpg

Figure 5. Catalographic file of "*vela aurica*"

The initial set of terms to be illustrated, used as a prototype set of samples, will be improved, including terms which are at lower level in the taxonomic chains which will be given a higher degree of specialization, and will, therefore, be more useful. Interesting points to be investigated in the future will concern methods able to: i) make this resource available to answer the needs of various kind of communities: professionals and non-professionals alike, ii) to interact with illustrated semantic networks namely PicNet (Borman et al., 2005) and large scale image ontologies such as ImageNet (Deng et al., 2009).

6 Conclusion

The terminological resource Mariterm was enhanced increasing its lexical coverage, designing a domain modelling and improving the tool for the database management with the possibility of giving visual information showing an image of each term. In such a way the meaning of a term can be clarified and more exhaustively represented.

Image classification and selection criteria were delineated together with the use of a set of files that contain catalographic information. In this framework, a system is planned for image cataloguing with more accuracy, based on user friendly tools and effective indexing strategies. In this way, the delivery of more complete description and useful information is performed from different points of view; the system, enlarged and provided with new details, becomes a flexible dynamic structure where there is a connection with applicative and pragmatic processes.

References

Borman A., Mihalcea R., Tarau P. 2005. PicNet: Pictorial Representations for Illustrated Semantic Networks, *Proceedings of the AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors,* Stanford, CA.

Cabré Castelvì M. T. 2000. *La terminologia: representacion y comunicacion,* Barcelona: IULA.

Deng J., Li K., Do M., Su H., Fei-Fei L. 2009. Construction and Analysis of a Large Scale Image Ontology. *Vision Sciences Society (VSS).*

Gangemi, A. 2005. *Development of an Integrated Formal Ontology and an Ontology Service for Semantic Interoperability in the Fishery Domain*, CNR – ICST, OCM Group.

Marinelli R., Roventini A., Spadoni G. 2006. Using core ontology for domain lexicon structuring. *Proceedings of LREC 2006*. Paris, ELRA.

Marinelli R., Roventini A.. 2006. The Italian Maritime Lexicon and the ItalWordNet Semantic Database. In E. Miyares Bermúdez and L. Ruiz Miyares (Eds.), *Linguistics in the Twenty First Century*; Cambridge.

Marinelli R., Spadoni G. 2007. Modelling a Maritime Domain Ontology. *Proceedings of the X International Symposium on Social Communication*. Santiago de Cuba.

Rosch, E. Principles of Categorization. 1988. *Readings in Cognitive Science, a Perspective from Psychology and Artificial Intelligence*, Morgan Kaufmann, San Mateo-California.

Appendix A. Screen Dumps

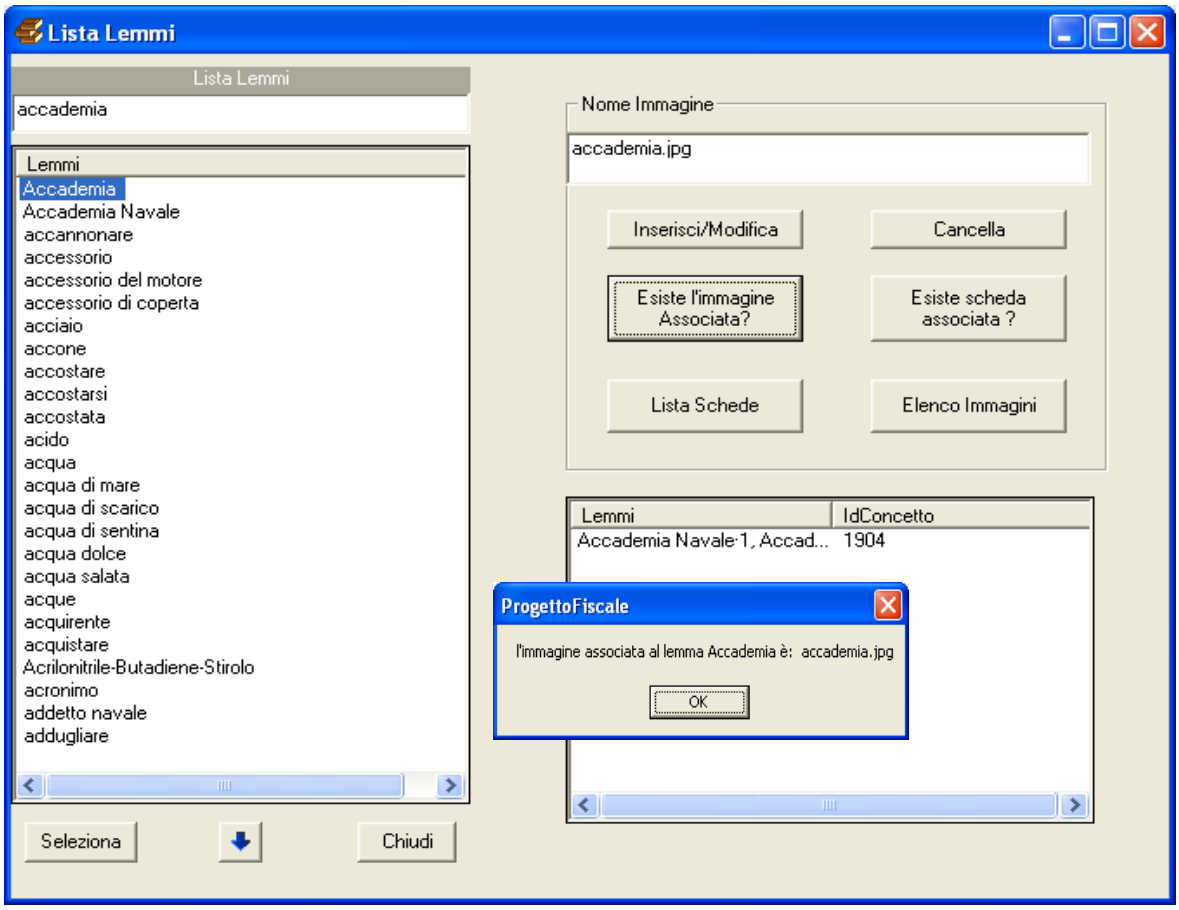

Figure 6. Screen Dump "Lista Lemmi".

The tool allows visualization of the image already inserted, e.g.:"*Accademia Navale*"(Naval Academy) is the lemma required; the name of the image and the identification number (id) of the concept are linked in a one-to-one correspondence.

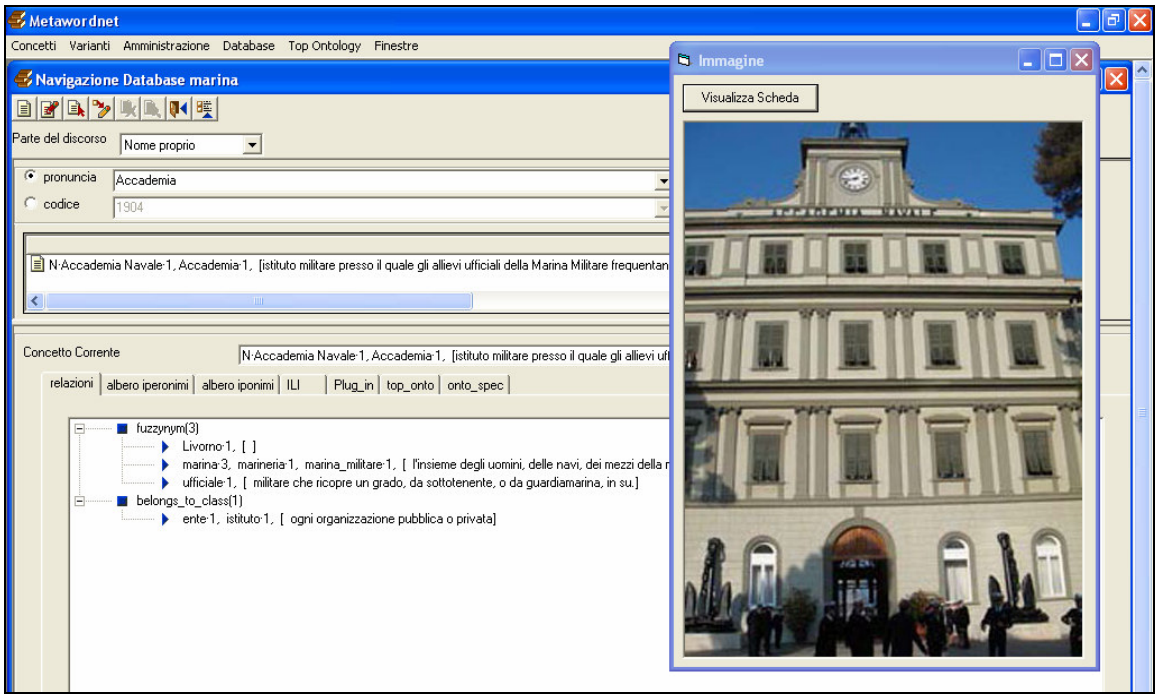

Figure 7. Screen Dump "Accademia"

The synset "*Accademia Navale*" (Naval Academy) is visualized together with the semantic relations and the image.

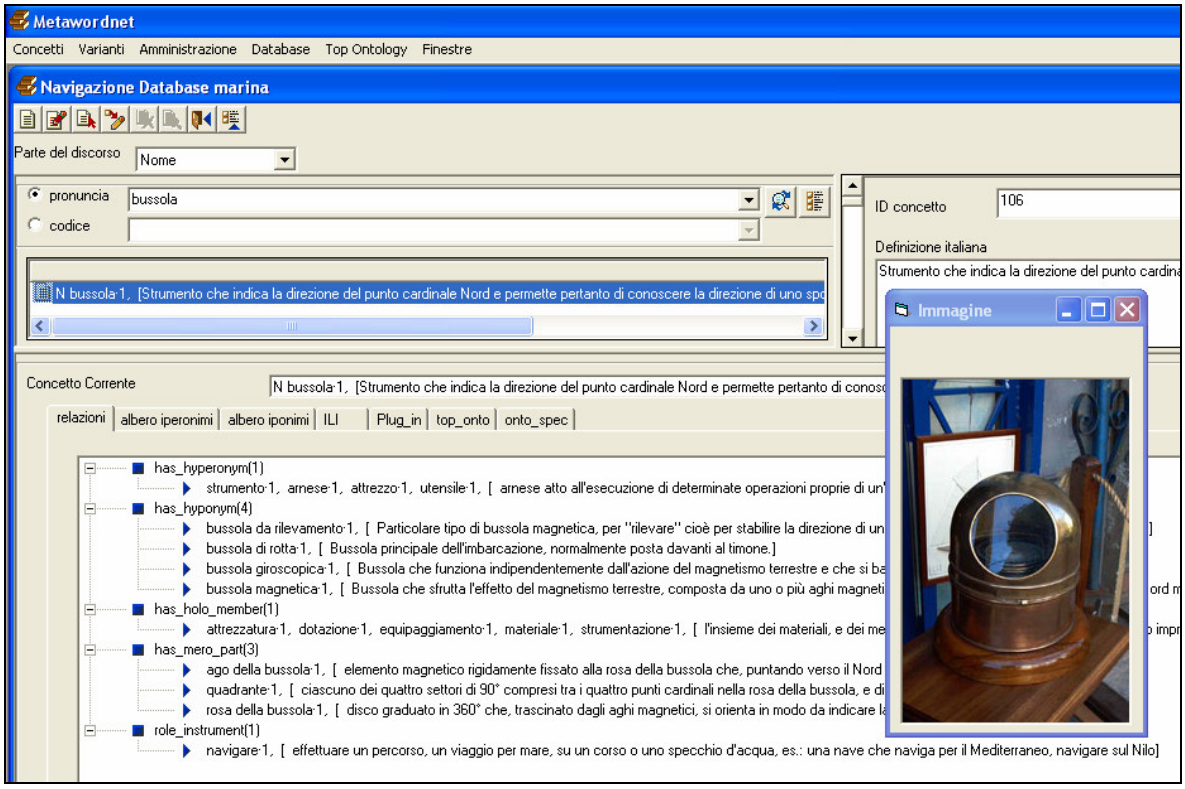

Figure 8. Screen Dump "Bussola"

The synset "*bussola*" (compass) as it appears with the semantic relations and the image.