RETRIEVAL AND EXPLORATION OF TERMINOLOGICAL KNOWLEDGE OVER THE WORLD WIDE WEB

George Vouros[*], *Konstantinos Kotis, Petros Tselios*
Department of Information and Communication Systems
University of the Aegean
Karlovassi, Samos
{georgev, kkot, tpe}@aegean.gr
http://www.samos.aegean.gr/icsd/Research/InCoSys/index.html

ABSTRACT

The objective of this paper is to report on the implementation of the *BIL*ingual *I*nformation *B*rowser (BILIB). BILIB in a greater extend than other systems allow users to retrieve terminological information, explore conceptual knowledge about terms and navigate transparently between a terminological database and a formal conceptual knowledge base. Major emphasis is given to the structure and design of the Knowledge Base, which has been designed and is being developed on the basis of EuroWordNet top ontology and base concepts.

Keywords: Terminology knowledge base, exploration, retrieval of terminological knowledge, Description Logics, EuroWordNet.

1. INTRODUCTION

Terminology is an interdisciplinary subject field involving the description and ordering of knowledge (cognitive level), and the transfer of knowledge (communicative level). Its central elements are *concepts* and *terms* [1].

As it is well known, terminology work involves among others [I1,2] the (a) identification of concepts and concept relations in a subject domain, (b) establishment of concept systems on the basis of concepts and their interrelations (c) assignment of preferred terms to concepts (d) recording of terms and their definitions.

Although concepts, the building blocks of knowledge, are the starting point of all practical terminological work and therefore, terminology is a very knowledge-intensive activity, only little concern has been given to the explicit representation and exploitation of formal concept systems of a subject domain in conjunction with linguistic data about the lexicalisation of concepts.

The aim of the project PROMETHEUS is to built a generic framework for exploiting terminology knowledge bases over the WWW, supporting users to explore the conceptual model underlying terminology in any domain, and providing rich information concerning the meaning and use of terms. Specifically, the aim of this project is twofold:

- The production of bi-lingual (Greek – English) terminological bases, and
- The development of a generic *BIL*ingual *I*nformation *B*rowser (BILIB) for searching, exploring and navigating in the terminology data and knowledge bases.

This paper describes the overall architecture of the *BIL*ingual *I*nformation *B*rowser (BILIB) and describes how the system supports users to retrieve terminological information and explore conceptual knowledge about terms. Major emphasis is given to the structure and design of the Knowledge Base, which has been designed and is being developed on the basis of EuroWordNet top ontology and base concepts.

The paper is structured as follows: Section 2 provides the key issues and motivating points of our work and introduces the technology utilized. Section 3 presents the requirements for BILIB, the overall architecture of the system, and a thourough description of the terminology data and knowlege bases utilized by the system. Finally, section 4 concludes the paper.

2. RELATED WORK & KEY ISSUES

One of the major aims that underlies most of the work in the context of the PROMETHEUS project, is to build a *generic framework* for developing multilingual "electronic encyclopaedias", which would *provide terms' translations* in dif-

[*] Contact Person

ferent languages, allow users to *explore the conceptual system* in any subject field, *retrieve terms* in any supported language based on their semantic relations with other terms and/or their linguistic characteristics, *understand the semantic relations between terms*, and view multimedia documents in which terms are referred, *supporting the proper use of terms* in the written and/or spoken language.

The fields of a typical terminological entry format [3] are organized in the following clusters:

- Acquisition data, such as language, country, date of term introduction and terminologist.
- Linguistic relevant data, such as part of speech, idiomatic expression, abbreviation.
- Explicatory data, such as definition, contextual use, comments.
- Deployment data, such as relations with other terms. These relations are the following:
  1. *hyperonym relation* : A is a hyperonym of B iff the term A has *broader* meaning than a term B. In this case, A lexicalizes a concept, which is a super-concept of (*subsumes*) the one that B lexicalizes.
  2. *hyponym relation* : A is a hyponym of B iff the term A has *narrower* meaning than a term B. In this case, A lexicalizes a concept, which is a sub-concept of (*is subsumed by*) the one that B lexicalizes.
  3. *meronym relation* : Term B is a meronym of A in case B lexicalizes a concept that plays a special *functional role* in the definition of the concept lexicalized by A.

Terminological data may be stored in database tables that may be organized according to this clustering of terminological information. Although recording of deployment data in databases may provide the semantic relations between terms, such a representation does not facilitate the explicit and formal specification of the lexicalised concepts and of their semantic relations. This prohibits multifunctionality of the terminological base, as well as its effective maintenance, presentation and exploitation [2,3,7]. Formal conceptual representations are important for describing concepts, describing semantic information about them and providing advanced reasoning services for exploitation and maintenance of the concept system. These issues lead to the explicit representation of deployment data in terminology knowledge bases, in conjunction with storing acquisition, linguistic and explicatory data about terms in terminology databases.

For the explicit representation of concept systems, this paper proposes the use of Description Logics (DLs). DLs provide an important, unifying framework for object-centred systems. DLs are suitable for describing generic concepts and individual objects using a subset of first order logic. They provide constructors for the description of concepts, objects and roles. Concepts represent classes of objects and roles represent binary relations between them. Concepts are taxonomized in a hierarchy. Concept classification is the task of placing a concept to the correct place in the concept hierarchy. Subsumption is the task of checking whether a concept is more abstract than another. Description Logics vary in the set of constructors offered, i.e., in their expressive power, and consequently in the complexity of determining class subsumption. Advanced reasoning services of DLs may facilitate:

- Maintenance and handling of inconsistencies [7]. For instance, to add a new concept in a concept system, the terminologist must decide its proper position within the system, check for inconsistencies with other concept definitions and record any reasoning implications that this addition may have in the concept system. Automatic classification of concept definitions, subsumption testing and role/attribute inheritance are advanced reasoning services of formal conceptual languages, such as Description Logics, towards solving these problems.
- Retrieval and exploration of terminological knowledge. Answering a query such as "provide all As that play a special functional role R in the definition of B" should provide all terms that lexicalize concepts that are more specific to the concept corresponding to A, and that play the functional role R in the definition of any concept that is more specific than the concept lexicalised by B. This requires advanced query-answer services, involving concept subsumption and classification.

In the context of PROMETHEUS the NeoClassic [4] DL has been used.

Since one of the major aims of the project is the development of a generic framework for the exploration and retrieval of terminological knowledge, a critical issue is the following:

- Use of a widely accepted semantic framework that would allow sharing and communicating conceptual and linguistic information about terms.

Towards this aim we have structured the conceptual knowledge base using the EuroWordNet (EWN) Top Ontology and Base Concepts. EWN top ontology offers a hierarchy of language-independent concepts, reflecting important semantic distinctions (Object and Substance, Part and Group, Static and Dynamic etc) [6]. Base concepts represent the most important, generic meanings and offer "the building blocks" for the construction and categorization of the subject domain concepts' definitions. Domain concepts offer an interlingua and provide the core of the multilingual (in the context of PROMETHEUS, bilingual) terminological database.

3. THE BILIB

3.1 The Overall System Architecture

The major requirements concerning BILIB are as follows:

1. Users must be able to search for any term, by specifying any combination of the following:
   a. The lexicalization of a concept in any supported language,
   b. Any combination of terminological (explicatory, linguistic, acquisition) data,
   c. A Concept description: This comprises semantic information about a term in any supported language.
2. Users must be able to navigate between terms and explore the concept system of the subject domain via widely used internet browsers. For instance, users must be able to retrieve all terms that are broader or narrower to a term, or those that play a special role in its definition (i.e. its "meronyms").
3. Users must be able to inspect terminological information in any of the languages supported by the system and must be able to explore and navigate through the terminological data and knowledge bases using any language.
4. Users must be able to browse not only between terminological entries but also between multimedia documents in which terms are referred.

Based on the above requirements, the overall architecture of the BILIB is as it shown in Figure 1:

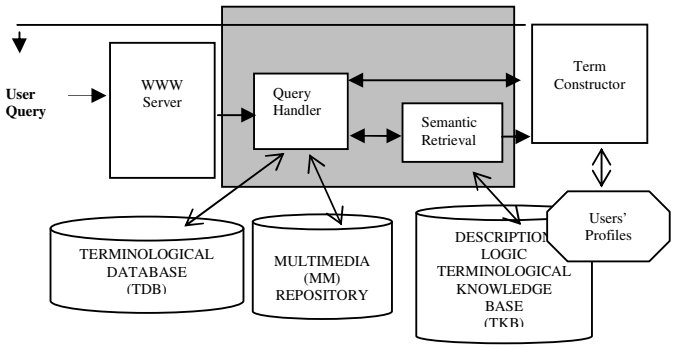

Figure 1. The BILIB overall architecture

The major system components are as follows:

a) **WWW server**: This is a widely used http server such as the IIS or Apache
b) The **user query** can be either:
- A query about a term (this can be formed by any combination of the three ways described above).
- A request for a set of documents.
- A special type of query such as "get a map of the concept system".

c) The **Terminological Data Base** contains terminological data (acquisition data, explicatory data, linguistic data) corresponding to each term as well as an indication of the formal concept that the term lexicalizes.
d) The **DL Terminological Knowledge Base** contains semantic information corresponding to each term encoded in NeoClassic Description Logic.
e) The **multimedia objects repository** contains multimedia objects in which terms are referred. These documents are linked to the corresponding terms.
f) The **user profiles** contain information on several types of users and the corresponding preferred views of information (e.g. projected fields, types of multimedia objects retrieved etc).
g) The **query handler** retrieves terminological data from the Terminological Data Base responding to user queries or to queries from the Semantic Retrieval module.
h) The **semantic retrieval** takes as input either a formal concept name, or a concept description and retrieves terminological knowledge from the Terminology Knowledge Base.
i) The **term constructor** takes as input (a) Terminological Data about terms and (b) Formal Concept Definitions, and constructs HTML pages that are sent to the http server.

3.2 Terminological Knowledge and Data Bases

Although in the overall system architecture provided above, data and knowledge bases consulted by the system appear to be separate resources, they are certainly interconnected. This section describes the structure of the major system bases and the way these bases refer to each other.

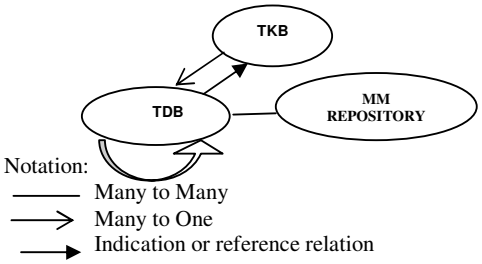

Figure 2. The overall BILIB bases architecture

a) As figure 2 indicates, each terminological record in the Terminological Data Base (TDB) is related with at most one concept and many multimedia objects stored in the MM repository.

b) There are separate terminological entries for each language, which are related among themselves via an equivalence relation. This happens through the concept, which the corresponding terms lexicalise.

c) Each concept is related with exactly one term for each language, and each object in the MM repository may be related with more than one term in each language (e.g. it may contain references for two or more terms).

e) Each terminological entry may be related with 0 or more non-preferred synonym terms via the "synonym" relation (this is indicated with the curved line showing to the TDB). Synonyms are used only for indexing purposes.

As already noted, in order to build a generic framework for the exploration and retrieval of terminological knowledge we have structured the conceptual knowledge base using the EuroWordNet (EWN) Top Ontology and Base Concepts.
Our objective is to provide a generic, domain independent and reusable ontology for classifying domain concepts. As Figure 3 shows, the conceptual knowledge base comprises three levels. Each level includes concepts that are subsumed by the top-level concepts:

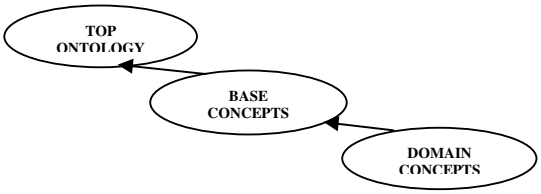

Figure 3. Terminology Knowledge Base layers

The Top Ontology level includes language - independent concepts that group Base Concepts (BCs) into coherent clusters, reflecting fundamental semantic distinctions. Top Ontology concepts include 1st and 2nd order types of Entities. Top-level concepts of these entity types (e.g. origin, form, situation-type etc) have been defined as Description Logic roles. This is evidenced by the fact that, as noted in [6], top concepts are more like semantic features than like conceptual classes.

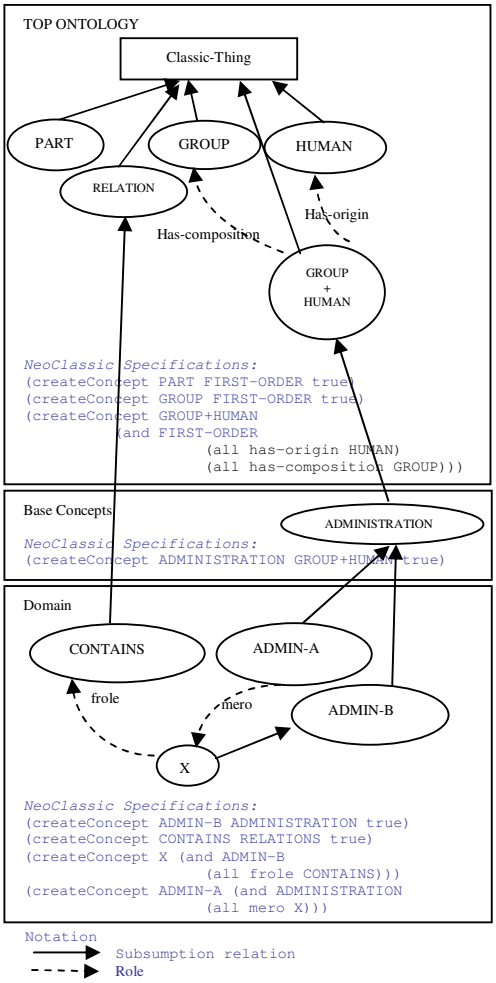

Figure 4. Terminology Knowledge Base

Finer distinctions of these top concepts are defined as primitive DL concepts (e.g. Part, Group,

Human, Object, Place, Building, Plant, Time, Usage, Social, Purpose, etc). Having made these specifications, as Figure 4 depicts, it is possible to define more complex top ontology concepts, such as Part+Group, in a comprehensive way, encoding semantic information about them.

BCs comprise the core of the multilingual database, classified under Top Ontology concepts. In the context of PROMETHEUS project, similarly to EWN, the importance of BCs has been measured in the content of their capability to function as an anchor to attach domain concepts. For instance, the concept ADMINISTRATION, classified under the concept GROUP+HUMAN is very important for one of our terminology bases.

All domain concepts are classified under Top Ontology and Base Concepts. The only roles defined at this level comprise: (a) The role "mero", which indicates the concepts that play a special functional role in the context of the defined concept (e.g. A is a mero of B, means that A plays a functional role in B's definition). (b) The role "frole" that indicates the type of functional role that is played by the "mero" concepts. For instance, the definition of ADMIN-A in Figure 4 specifies that ADMIN-B plays a special role in this definition (via the role "mero") and, specifically, that ADMIN-B is contained (as a physical part) in ADMIN-A. Notice, that to introduce such a definition we need special "place-holder" concepts (e.g. X) that denote the special role that is played by concepts in the context of defining more abstract concepts. In our example, X is a placeholder concept, denoting the special role played by ADMIN-B in the context of ADMIN-A (i.e., "ADMIN-B is contained in ADMIN-A"). ADMIN-B may, in a similar way, play other roles in the context of other concepts' definition.

Let us now for instance consider the following user query: *"Which terms are hyponyms of the term "administration" and have a meronymic relation of type "contains" with other "administrations"?"*.

To answer such a query, the system introduces the following concept specification

```
(createConcept Q (and ADMINISTRATION
                all mero (and ADMINISTRATION
                        (all frole CONTAINS)))
```

which is automatically classified by NeoClassic in the concept system. Lexicalisations of the subsumers of concept Q provide the answer to the user query. In our case, the answer is ADMIN-A.

4. CONCLUSION

This paper presented the architecture of the Bilingual Information Browser (BILIB) and introduced the generic framework utilized for exploiting terminological data and knowledge over the WWW.

BILIB is a generic tool for the exploration and exploitation of terminological resources over the World Wide Web using widely used web browsers and allowing users to explore and navigate in terminological data and knowledge bases in an integrated and transparent way, providing facilities for answering complex semantic queries about terms. BILIB has been implemented in C++ and is operational over the WWW.

Major emphasis has been given to the development of a formal conceptual system using a Description Logic language. This provides explicit specification of the conceptualisation of the subject domains. This further facilitates retrieval and exploration of terminological knowledge as well as maintenance and development of terminological resources.

5. REFERENCES

[1] The Localization Industry Standards Association, ISO TC / 37 Distribution, "Terminology Work - Principles and Methods", April 26, 1994.

[2] Ingrid Mayer, "Knowledge Management for Terminology - Intensive Applications: Needs and Tools".

[3] G.Vouros, V.Karkaletsis, C.Spyropoulos, "Documentation and Translation", in Software without Frontiers, P.A.V.Hall, R.Hudson, (Eds), Wiley, 1997.

[4] Peter F. Patel-Schneider, Merryll Abrahams, Lori Alperin Resnick, Deborah L. McGuinness and Alex Borgida. ``NeoClassic Reference Manual: Version 1.0." Artificial Intelligence Principles Research Department, AT&T Bell Labs, 1996.

[5] British Standard Guide to Establishment and Development of Multilingual thesauri, BS 6723: 1985.

[6] Piek Vossen (ed.), EuroWordNet, General Document 1, Final, July 19, 1999 University of Amsterdam, http://www.hum.uva.nl/~ewn.

[7] E. A. Karkaletsis, C.D.Spyropoulos, G.Vouros, "The Use of Terminological Knowledge Bases in Software Localisation", Machine Translation and the Lexicon, LNAI 898, 1994.