A Specialized Knowledge Base: from Distributed Information to the Specialized Dictionary Construction

M. Teresa Cabré, Rosa Estopà, Judit Feliu
Insitut Universitari de Lingüística Aplicada (Univesitat Pompeu Fabra)
La Rambla, 30-32; 08002 Barcelona, Spain
{teresa.cabre; rosa.estopa; judit.feliu@upf.edu}

Abstract
The main goal of this paper is to show the procedure of deriving linguistic and semantic information from a specialised knowledge base in order to build a terminographic entry. Thus, from the information contained in the knowledge base, which is organised into four different modules (corpus, terms, ontology and bibliographic references) we will be capable to extract all different kinds of information necessary for the construction of a terminographic register. This terminographic register will be the basis for the further development of a specialised dictionary.

1. Term data bases

From the computer development in the 70's up to the present moment, term databases have been an important resource for most terminological applications, among others, specialised translation, specialised writing, thesaurus and dictionary building. They provided terminological units accompanied with linguistic, conceptual and pragmatic information which was easily accessed.

However, term databases have suffered from different stages concerning their conceptualisation, organisation and goal, as stated by Sager (1990), Cabré (1992), Arntz y Picht (1995), and Tebé and Cabré (2002). These different stages reflect the computer science evolution and the new terminological needs that have arisen as time goes on.

Thus, the first generation of term databases, created in the 70's, were mainly oriented to store large amount of data. In this sense, they gathered a great number of terminological registers from a wide variety of domains (see, for example EURODICATOM, the Banque Terminlogique of Quebec o TEAM (Siemens)). The term base format was very similar to terminological files (Rondeau, 1981) and, even though they are conceived from the concept, they can be accessed from the terms or the domains containing all terms.

From the 80's, the specialists on the building up of term databases emphasised the quality and update of the data. Access and international transfer were also another priority. For this reason, smaller term databases were developed and the specialised domains concerned were very much delimited in order to attain user's needs. In the sake of international information transfer, the ISO Technical Committee 37 proposed several international standards with the aim to aid the transfer of terminological data (see, for example, MATER 1987 [MAGNETIC *Tape Exchange Format for Terminological/Lexicographical Records*], TEI/LISA/ISO-TIF [*Terminology Interchange Format*], MARTIF, and so on.

It is at the end of the 20th century when term databases have suffered from a process of specialisation and they have been designed according to professional needs. This evolution concerns both the design criteria (information organisation) of the databases and also the technical aspects (format, accessibility, information retrieval). An example of this new and more complex database is DOT (Martin and Heid, 2001).

Probably, the main changes on the term databases design affect diversity and dynamics. Thus, new data bases are designed according to their goals, their domain, their volume, the number of languages included and, for this reason, it is impossible nowadays to think about one single type of term database. They are no longer a *depot*, but also a way to store and classify linguistic and semantic information in a way oriented to the user's needs and the possible information retrieval tasks carried on by the future users.

2. Term databases for the terminographic work

Term databases have been traditionally used in a wide range of tasks but they have been an elementary starting point for the build-up of dictionaries. The term databases are essential for the units selection and in all other stages of the terminographic work oriented to create a terminological application.

Undoubtedly, term databases are necessary for any terminographic work but there are some cases in which there is no a database including all the information that could be useful for a particular resulting application. This enlarged term bases would be useful for the selection of a particular unit, for the creation of a definition, for the indication of usage restrictions and for the possible query of new usage contexts of a concrete terminological unit. It would also be very useful to have a conceptual structure, such an ontology, relating those terminological units referring to the same concept and establishing different kinds of conceptual relations among these units.

Thus, we would like to state that the knowledge bases, —defined as the automatic information organised that represents the whole knowledge of the specialists about a particular domain (Cabré, 1999)— have become the automatic applications that give a more adequate response to the user's needs. They integrate a large amount of information and they gather together the terminological information, the conceptual information by means of an ontology and they are directly linked to a specialised corpus and to bibliographic information about a particular specialised domain.

3. The design criteria of the specialized knowledge base

In the Institute for Applied Linguistics framework, it has been designed and developed the GENOMA-KB, a specialized knowledge base on the human genome domain (http://www.iula.upf.edu/).

This knowledge base has been conceived as a collection of information and data, organised in a distributed structure. Thus, it includes four separated modules: a textual database, a documental database, a terminological database and an ontology.

As for the textual module is concerned, the human genome corpus follows the IULA core project of LSP corpora building (Technical corpus of IULA). All texts included in the

corpus building process are selected and validated by a domain expert and classified according to a domain structure.

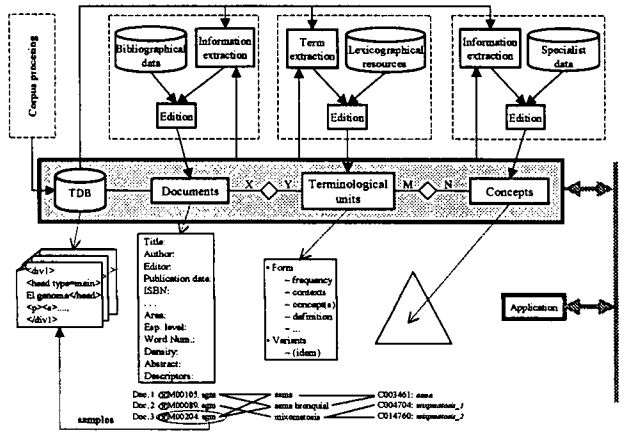
The documental database contains the references of publications and organisms concerned with the human genome domain. This module includes all the references to the documents contained in the corpus but also the indication of books, articles, papers, PhD dissertations and some relevant organisms and sites related to the human genome domain.

The ontology is very closely tied to the terminological database and, both modules are built up with the OntoTerm© terminological management system. The ontological module has been built with the aid of a domain expert who provided a core group of concepts that has been further enlarged. These concepts are the basis for the inclusion of any new term in the term database and they are linked among them by the traditional hyponymy relation but also by a set of other conceptual relations predefined in the editor. The concepts and their conceptual relations will become important semantic information for the construction of a term entry in a specialized dictionary.

Concerning the terminological module, it is constituted by a term database. Terms are included in three different languages (Catalan, Spanish and English) and each term is semantically described by its link to the ontology, on the one hand, and by a set of linguistic data categories, on the other hand. These linguistic data categories are, for each term in each different language, the part of speech; the number and gender assignment; the usage contexts and their sources; the lemmatised form of the entry and some administrative data, which are mandatory. Moreover, we are also free to include the definition and its source and some usage notes.

At the moment of this research, the Human Genome Knowledge Base includes 458 entities in the documental database, more than 7M tokens in the textual corpus, 1.206 concepts in the ontology and 2.378 terms. The general structure described above can be illustrated as follows:

Figure 1: The Human Genome Knowledge Base overview.



4. The terminological module: a source of linguistic information

The semantic and linguistic data collected in each one of the records of the terminological database represent a very important source of information for the future elaboration of a terminological entry in a specialised dictionary. The following illustration shows the ontology and the term editors that, as you can see, contain very valuable information which cannot be automatically transformed into a dictionary but can be, without any doubt, reused for the writing of an specialised dictionary.

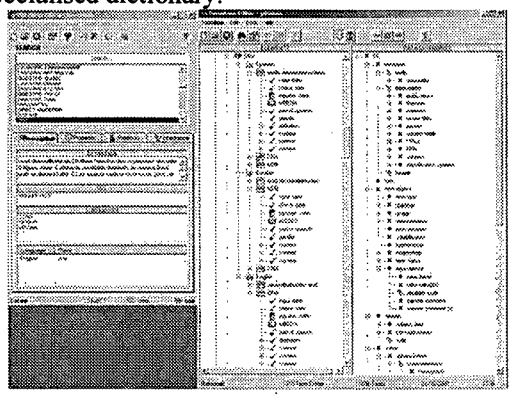


Figure 2: The ontology and the term base editors: the DNA concept.

In this figure, we can distinguish, on the left side, the ontology editor containing a part of the semantic information inherited by a term. On the right side, the figure shows the term base editor containing all terms in Catalan, Spanish and English, which refers to the previously selected DNA concept.

More concretely, on the ontology editor we have the concept DNA marked into blue colour. We can indicate a definition and, in the two following boxes below, the system shows the hyponymy relation. In this case, DNA is a NUCLEIC_ACID and its hyponyms are CDNA, CPDNA and MTDNA. Still in the ontology editor, the properties box will show the other kinds of conceptual relations indicated for this concept (DNA is generally associated with REPLICATION / DNA is located in CHROMATIN). All this information is not visible in the term base editor but when exporting the whole core of information it will also be retrieved.

In the case of the term base, we observe that before any term or linguistic information inclusion is possible, we have to select a concept previously introduced in the ontology. From the DNA concept, we have introduced all possible denominations for this concept in the three different languages already mentioned. All term entries are accomplished and described by the same data categories, which are the following: The first three categories (input date, check date and inputter) are administrative information. Next, the number appearing (m00216) indicates the number of the document in the textual corpus where the term unit has been documented. Following this information, we include the part of speech indication (and the number and gender for those entries in Catalan or Spanish). The optional definition is included in this case and the three mandatory contexts have been retrieved from the corpus module, as the numbers of the documents indicate.

All the information commented on until now can be consulted and exploited in two different ways. The term base editor of the OntoTerm© applications foresees a way of exporting all conceptual entities into individual .html pages. All the information contained in the ontology and in the term base appear grouped into a single html register and all different registers are interrelated by means of the conceptual relations browse links.

The other way to consult the information contained in the Human Genome Knowledge Base is a web site, which is still under development. This web site will present three different kinds of research: simple, complex and combined. At present, the simple search option is already available. The main aims of the GENOMA-KB web site is to unify the access to all the information collected in the different four modules of the knowledge base. The integration of the information is showed in the search option. Searches will be on each one of the modules individually and the system expands the query to the other modules in the case of direct information relation.

5. Towards the construction of a term entry in a specialised dictionary

Undoubtedly, the information contained in the term base and the ontology, but also the linguistic and non-linguistic data derived from the corpus and the document databases, will help us in a decision-making and construction of a specialised dictionary. If we have decided to build up a multilingual terminological dictionary about a specific topic of the human genome domain, we dispose of a large amount of semantic and linguistic information, which will be directly derived into the body of the dictionary. For example, the information about the part of speech and gender for all term entries introduced in its lemmatised form.

As for the definition is concerned we have two different possibilities. We can draw the same definition from the database and indicate the source or, in the case of absence or disagreement, we can write a new definition with the aid of the three mandatory contexts and the conceptual relations that this concept, and also its corresponding terms, keeps with the other concepts of the dictionary. In this sense, the real usage contexts would help us to write a completely new definition with a high degree of reliability concerning the use of these units from the domain experts.

The semantic information contained in the ontology will help us, without any doubt, to classify the term entries of the multilingual dictionary according to a pre-established semantic classification and the different relations among terms will be an useful tool in order to construct definitions and to refine the use of some other terms also defined in the dictionary. Finally, the usage contexts can be directly included as the examples of a particular term entry taking into account that this information has been collected with the aim to reflect the maximum different points of view on a particular terminological unit.

6. Conclusions

In this paper, and after having designed the framework for databases work, we have depicted the main characteristics of a specialized knowledge base containing four different information modules. Our aim has been to show how a knowledge base with diverse information and distributed structure could be useful for applications such as the construction of a specialized dictionary, which would be created on the basis of the linguistic and, particularly, semantic information contained in the specialized knowledge base.

Thus, this knowledge base will become the starting point for a future work consisting on the writing of term entries that will include specialized semantic information in different languages. The working process will be parallel for these three languages and the final resulting terminological application will have followed a well-defined methodological strategy.

7. Acknowledgements

This research has been partially funded by a public project: TEXTERM 2: Fundamentos, estrategias y herramientas para el procesamiento y extracción automática de información especializada (BFF2003-2111).

References

Arnzt, R. and Picht, H. 1995. *Introducción a la terminología.* Madrid: Fundación Germán Sánchez Ruipérez, Pirámide, Biblioteca del libro.

Condamines, A. and Rebeyrolle, J. 1998. 'CTKB: A corpus-based approach to a Terminological Knowledge Base' in: D. Bourigault, Ch. Jacquemin, and M. Cl. L'homme (eds.) *Computerm '98. First Workshop on Computational Terminology. Proceedings of the Workshop.* COLING-ACL '98, 15th August 1998. Université de Montréal. Montreal, Quebec, Canadà, p. 29-35.

Condamines, A. 1999. 'Approche sémasiologique pour la constitution de Bases de Connaissances Terminologiques' in: V. Delavigne and M. Bouveret (eds.) *Sémantique des termes spécialisés.* Rouen: Université de Rouen, col. Dyalang, p.101-118.

Cabré, M. T. 1992. *La terminologia. La teoria, els mètodes, les aplicacions.* Barcelona: Empúries.

Cabré, M. T. 2003. 'Sobre l'organització de bancs de coneixement' in: *Ciclo de conferencias: Semántica cognitiva y tratamiento automático del español*, Universitat Autònoma de Barcelona. Barcelona, 18th september de 2003 [oral presentation].

Martin, W. and Heid, U. 2001. 'Frame-based definitions and the selection of multiword term candidates in DOT' in: *TIA 2001, Actes des quatrièmes rencontres Terminologie et Intelligence Artificielle*, (Nancy: INIST/CNRS), p. 55-65.

Meyer, I. 1990. 'Computer-assisted Concept Analysis for Terminology Work' in: *Proceedings of the Nordic Post Graduate Course in Terminology* (Mariehamn, Finland, Sept. 1990). Estocolm: Tekniska nomenklaturcentralen, p. 193-212.

Meyer, I. 1998. *The COGNITERM Project.* [on-line resource: http://aix1.uottawa.ca/~imeyer/research.htm.

Meyer, I. 2001. 'Extracting knowledge-rich contexts for terminography' in: D. Bourigault, Ch. Jacquemin and M. Cl. L'Homme *Recent Advances in Computational Terminology.* Amsterdam/Filadèlfia: John Benjamins Publishing Company, p. 279-302.

Sager, J-C. 1990. *A Practical Course in Terminology Processing.* Amsterdam / Philadelphia: John Benjamins Publishing Company.

Tebé, C. and Cabré, M. T. 2002. 'Hacia un nuevo modelo de bancos de datos terminológicos' in: M. Correia (ed.) *Actas del VI Simposio Iberoamericano de Terminología: Terminología, desarrollo e identidad nacional.* Lisboa: Ediçoes Colibri, p. 851-864