

# The GENOMA-KB project: a concept based term enlargement system

Judit Feliu, John Jairo Giraldo, Vanesa Vidal, Jorge Vivaldi, M. Teresa Cabré

Institute for Applied Linguistics  
La Rambla, 30-32; 08002 Barcelona, Spain  
{judit.feliu; john.giraldo; vanesa.vidal; jorge.vivaldi; teresa.cabre}@upf.edu

## Abstract

The GENOMA-KB knowledge base includes four independent modules: a textual database, a factual database, a terminological database and an ontology. We will briefly introduce in this paper the main features concerning each one of the modules, and we will highlight the process of enlarging both the term base and the ontology.

## Introduction

In the framework of the GENOMA-KB project, a special relevance has been given to the interaction between the ontology and the terminological information organized into modules. The first goal of this paper is to describe the process of enlarging and updating the term base module both for the linguistic information and the terms' direct link to the ontology. The second aim is to propose a combination methods strategy oriented to retrieve terms and term candidates for their inclusion in the term database. Some final conclusions and future research lines will be drawn at the end of this paper. A detailed discussion about the GENOMA-KB project is presented in a separated paper.

## The GENOMA-KB description

In the genetics domain, there are some databanks publicly available, such as, Gene Ontology<sup>1</sup>, LocusLink<sup>2</sup>, Gene Ontology Browser<sup>3</sup>, and GeneCards<sup>4</sup>. These resources include high-specialized data and they are an important assistance for domain researchers. However, their exploitation is difficult for other kind of users as terminologists, translators and scientific journalists.

From the terminology point of view, it is worth mentioning the attempts for the integration between terminological units and some conceptual information found at (Meyer et al., 1992; Condamines et al., 2000).

The GENOMA-KB integrates four modules: a textual database that contains specialised texts of this particular domain; a factographic and documental database containing the meta-information about the tagged texts in the corpus; a terminological database including the linguistic units transferring specialized knowledge, and a human genome ontology, which will be the basis for establishing a conceptual link between terminological units and the concepts they transfer. Figure 1 shows the tight relation between the four modules that take part in this knowledge base:

- Textual database: it contains actual documents directly related to the human genome domain. We collect texts in three languages: Catalan, Spanish and English.

- Document and factographic database: it registers bibliographic information about the texts in the textual database and metadata related to the genome domain.
- Terminological database: specialized knowledge units extracted from texts are introduced in this database and they are linked to concepts in the ontology.
- Ontology: concepts and their corresponding knowledge units entered at the terminological database appear in a knowledge organization based on a set of both hierarchical and non-hierarchical conceptual relations.

The term base and the ontology modules are closely tied due to the theoretical approach and its derived design decision of the software application to be used for managing the terminological and the conceptual information. The term base and the ontology editors are part of the OntoTerm<sup>5</sup> terminological management system. One of the application requirements is the previous development of the ontology, which will be the basis for the inclusion of any new term. This obligatory assignment has some theoretical consequences such as biunivocity between concepts and terms. One of our concerns has been to overcome this restriction by enlarging the number of conceptual relations and allowing multiple inheritance.

The ontology is built upon a initial set of concepts that have to be used as the starting point. These top concepts mainly concern the basic semantic categories, that is, events, objects, relations and properties. At present, the Human Genome ontology gathers more than five hundred concepts linked by a wide range of conceptual relations different from the traditional hyponymy one. We are currently applying a set of relations (equivalence, location, space, causality, meronymy, etc.) derived from a more general running research on this subject. Also properties can be attributed to a concept.

<sup>1</sup> <http://www.geneontology.org/>

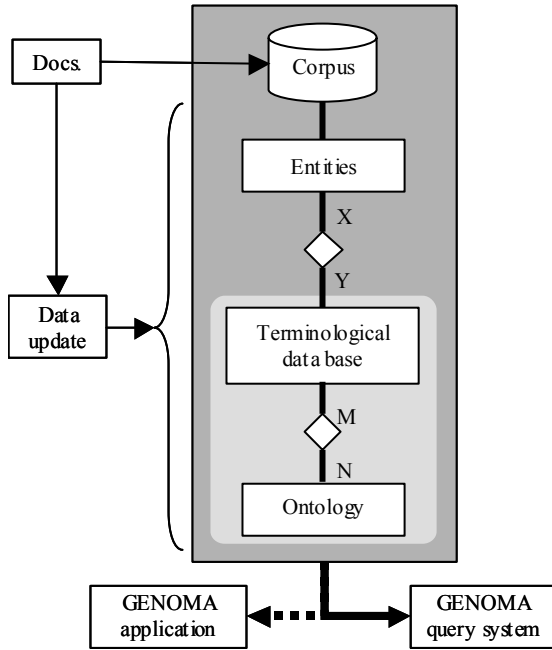
<sup>2</sup> <http://www.ncbi.nlm.nih.gov/LocusLink/>

<sup>3</sup> [http://www.informatics.jax.org/searches/GO\\_form.shtml](http://www.informatics.jax.org/searches/GO_form.shtml)

<sup>4</sup> <http://bioinfo.weizmann.ac.il/cards/index.html>

<sup>5</sup> OntoTerm is a terminological management system built by Antonio Moreno, from the Universidad de Málaga. More information available at: <http://www.ontoterm.com>.

Figure 1: General structure of the GENOMA-KB



### The mutual need between the Term Base and the Ontology

After having reviewed the available resources for terminological management and ontology building, we have decided to use OntoTerm (Feliu; Vivaldi; Cabré, 2002a and b). This tool is based on a conceptual structure previous to the term base creation. In this sense, the ontology building is the previous stage before the construction of the term base. Given this design philosophy, we present firstly the ontology and, secondly, the main characteristics of the term base directly linked to the ontology.

The core ontology was built with the aid of a domain expert who has provided its initial structure for the conceptual structure building. Thus, new concepts have been added to a previous list of base concepts necessary for the system performance. As a matter of fact, the system includes 21 base concepts (ALL, OBJECT, EVENT, PROPERTY, etc.). In addition, the domain expert has proposed a list of about 100 concepts used in the human genome domain that have been integrated to the initial list.

Once the concepts recommended by the domain expert were included in the ontology, the following step consisted on filling in the term base. For each term (related to a given concept) and each language it has been provided with the following specific information:

- The term itself
- Part of speech
- Number and gender assignment
- Contexts (each entry term has to have at least one context)
- Context's source
- Lemmatized form (from the morphological point of view)
- Administrative information

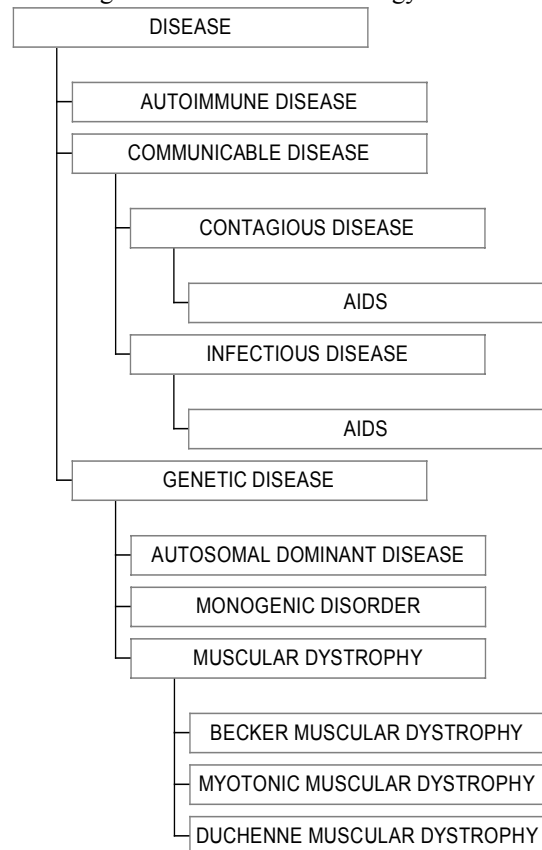
All the information above mentioned is mandatory. Apart from it, there is some additional information that is

optional: the term definition as well as its source and some usage notes.

The term base methodology consists of selecting *a priori* some tied related concepts forming a category. For example, under the category DISEASE it can be found different types of diseases such as: AUTOIMMUNE DISEASE, COMMUNICABLE DISEASE, GENETIC DISEASE, etc. Similarly, deriving from a given type of disease it can also be found its corresponding subordinate concepts. Thus, for GENETIC DISEASE will appear concepts such as AUTOSOMAL DOMINANT DISEASE, MONOGENIC DISORDER and MUSCULAR DYSTROPHY. Under the term MUSCULAR DYSTROPHY will appear concepts like BECKER MUSCULAR DYSTROPHY, DUCHENNE MUSCULAR DYSTROPHY, MYOTONIC MUSCULAR DYSTROPHY, etc. See Figure 2.

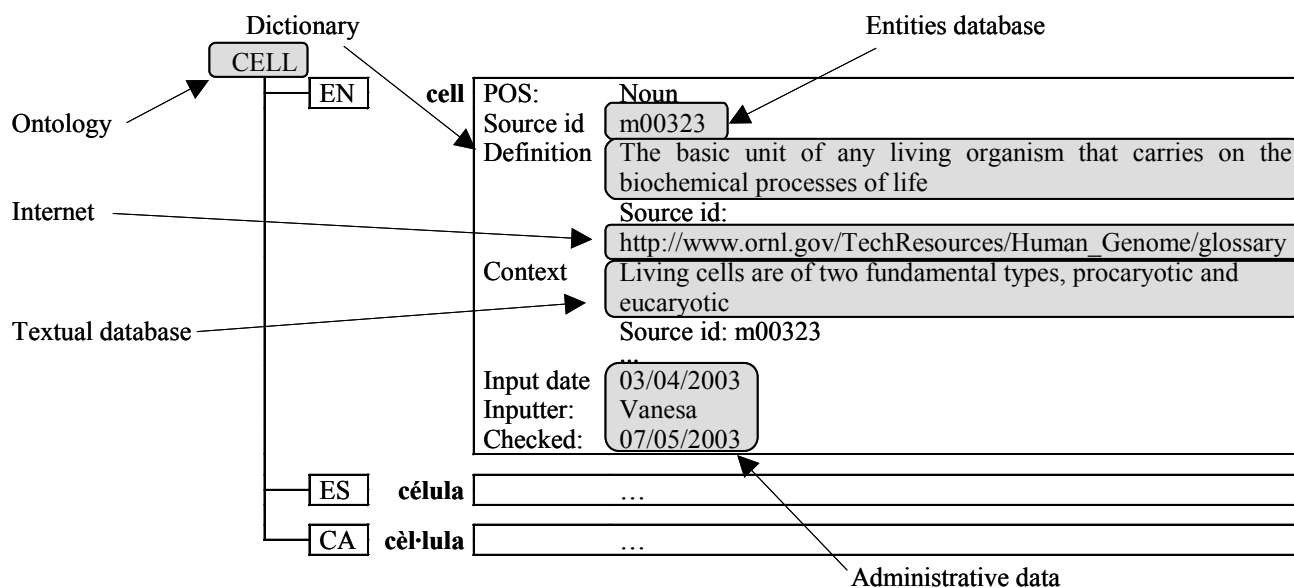
Another important ontology feature is the multiple inheritance. As usual, for multiple inheritance we understand that one concept might have two or more ancestors. This fact can be clearly observed in the case of the concept AIDS, which is at the same time a CONTAGIOUS DISEASE and an INFECTIOUS DISEASE. We think that a robust ontology should account for this kind of phenomena.

Figure 2. Extract from ontology tree



From the terminologist viewpoint, working on conceptual categories, as for example all concepts related with DISEASE, has two main advantages. On the one hand, it permits to gain efficiency and internal coherence since it becomes a systematic work. On the other hand, it also allows getting cognitive competence in the subject matter. Now, we describe the criteria applied to retrieve relevant data from the corpus module, which will be the source of information for many of the fields characterizing each term.

Figure 3. Data sources for the term database



As mentioned above, the work is done in a systematic way. Consequently, it is divided into two main steps. Firstly, equivalent terms and their variants are assigned to each ontology concept. Secondly, all the information related to each term entry is added.

In regard to the first step, it is composed of the following phases:

1. Introduction of equivalent terms for Catalan, English and Spanish
2. Selection of variants for each term. The definitions and specially the contexts are the main source for obtaining these variants. The following is an example that illustrates well the way we detect term variants.

- “**Duchenne's muscular dystrophy**, the most common and severe type of pseudohypertrophic muscular dystrophy; chronic and progressive, it begins in early”. Called also *Duchenne's d.*, *Duchenne's* or *Duchenne-Griesinger disease*, *Erb's atrophy* or *Erb's d.*, and *Zimmerlin's atrophy*. Cf. *Becker's muscular d.*”.
- “*Some communities have begun screening for **Duchenne muscular dystrophy (DMD)** by measuring creatine kinase levels in newborns*”.

It is important to remark that not any variant is prioritized. Hence, our work is not prescriptive but descriptive. In fact, our departure point is the text.

Once the first step has been accomplished, it is necessary to enter the remaining linguistic information as well as the administrative information. The procedure is systematic and is reflected in the following stages:

- 1) *Part of speech* (adjective, noun, verb).
- 2) *Gender* (specified only for Spanish and Catalan).
- 3) *Number* (only for lexicalised plurals).

- 4) *Lemma form*. It is used for establishing the link between the term and the units contained in the textual database.
- 5) *Definition and its source*. It is taken from both analogical and digital specialized dictionaries.
- 6) *Contexts and their source*. They are taken from the IULA's technical corpus as well as Internet. In the last case, the context's source documents must accomplish some requirements. That is, the information must come from relevant journals, and/or public and private research centers web sites. In addition, these sites must contain all the data dealing with the author, place and year of publication.
- 7) *Source of the term*. The source comes from the most representative context.
- 8) *Notes*. They are used generally for remarking preferred uses or irregular forms. For example, *pulmonary alveolus*. (*Note: usually is documented in plural form. The plural form is "pulmonary alveoli"*).
- 9) *Inputter*. The name of the person responsible for the entry.
- 10) *Input date*. The date in which the entry was created.
- 11) *Check date*. The date in which the entry has been revised.

Figure 3 above shows all the just described information for the English term of the record corresponding to CELL concept.

### The Term Base enlargement

As for the second aim is concerned, two different tools, *Mercedes* and *YATE* will be presented. They have been developed in the Institute for Applied Linguistics, and they are used to retrieve terminological units contained in specialised texts, more specifically, in human genome domain texts.

The first tool, *Mercedes*, is a term detector used to retrieve terminological units from texts. Essentially, the system compares the units of a particular domain contained in an

internal term database with all the units of a given text. This database has been built from public domain specialised dictionaries and can be easily adapted to any new domain.

*YATE* is a term candidate extractor based on the combination of different (linguistic and stochastic) strategies. It has also been applied to the same texts in order to obtain a different list of single and multiword terminological units.

Next step of the working process consists of the straight inclusion to the term base of the list of terms containing the coincident units derived from both tools. Following this procedure is mandatory to check that the concept it represents is already included in the ontology. Otherwise the conceptual structure must be updated.

For the remaining list of terms proposed just by *YATE*, it envisages to check their termhood by looking to relevant resources combined with a domain expert consultation. Table 1 shows a contingency table sample of the term candidates proposed by both tools and for each one separately. This information, as already mentioned, will be used for updating both the Term Base and the Ontology.

Table 1. Term Candidates comparison

	Mercedes	non-Mercedes
Yate	<i>genoma</i> <i>gen</i> <i> cromosoma</i>	<i>mioglobina</i> <i>apolipoproteína</i> <i>mucina</i>
non-YATE	<i>hebra</i> <i>hélice</i> <i>secuencia</i>	<i>microarray</i> <i>melanocortina</i> <i>retropseudogen</i>

## Conclusions and Future Research

In this paper, we have described the main features of the GENOMA-KB, understood as a multi-module knowledge base integrating the corpus, the bibliographic data, the term base and the directly related ontology. It has been highlighted the methodology followed in order to enlarge the term base and the ontology and the reuse of the results of a term detector and a term extractor to ease the updating task.

In order to reuse the maximum information obtained from these tools it is foreseen, in a running research, to take profit from the textual fragments containing terms (both validated and candidates) linked by a conceptual relation expressed by a verbal form (Feliu, 2004).

## Acknowledgements

The authors acknowledge the contribution of Eva Valero in the ontology building.

This research and the development of the query system have been carried on under the two finished public funding projects *TEXTERM: Textos especializados y terminología: selección y recuperación automática de la información* (BFF2000-0841), lead by M. T. Cabré; and *RICOTERM: Sistema de recuperación de información con control terminológico y discursivo* (TIC2000-1191), lead by M. Lorente. And also under the current public funding of *TEXTERM2: Fundamentos, estrategias y herramientas para el procesamiento y extracción*

*automática de información especializada* (BFF2003-2111), lead by M. T. Cabré.

## References

- Condamines, A; Rebeyrolle, J. (2000). Construction d'une base de connaissances terminologiques à partir de textes: expérimentation et définition d'une méthode. In: Charlet, J. et al. (eds.) *Ingénierie des Connaissances, évolutions récentes et nouveaux défis* (pp. 225-242). Paris: Eyrolles.
- Feliu, J. (2004). *Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica*. PhD dissertation. Universitat Pompeu Fabra. Barcelona.
- Feliu, J.; Vivaldi, J.; Cabré, M. T. (2002a). «Towards an Ontology for a Human Genome Knowledge Base». *LREC2002. Third International Conference on Language Resources and Evaluation. Proceedings* (pp. 1885-1890). Las Palmas de Gran Canaria, may 2002.
- Feliu, J.; Vivaldi, J.; Cabré, M. T. (2002b). *Ontologies: a review*. Working Paper, 34. Barcelona: Institut Universitari de Lingüística Aplicada.
- Martin, L.E. (1990). *Knowledge Extraction*. In *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 252--262). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Meyer, I.; Bowker, L.; Eck, K. (1992). *Cogniterm: An Experiment in Building a Terminological Knowledge Base*. *Proceedings 5<sup>th</sup> Euralex International Congress on Lexicography*. Tampere, Finland.
- Vivaldi, J. (2001). *Extracción de candidatos a término mediante combinación de estrategias heterogéneas*. PhD dissertation. Universitat Politècnica de Catalunya. Barcelona.