The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities

M. Teresa Cabré, Carme Bach, Rosa Estopà, Judit Feliu, Gemma Martínez, Jorge Vivaldi

Institute for Applied Linguistics
La Rambla, 30-32; 08002 Barcelona, Spain
{teresa.cabre; carme.bach; rosa.estopa; judit.feliu; gemma.martinez; jorge.vivaldi}@upf.edu

Abstract

In the past twenty years much efforts have been devoted to the development of ontologies and term bases for different fields. All this work has been done separately or with slight integration. The GENOMA-KB is a project whose main aim is to integrate, at least, both resources. In this paper, most relevant aspects of the project are presented. Each module is individually described and the links among them are highlighted. Finally, a query system to interrogate the knowledge base is briefly introduced.

Introduction

Since the 90's, the ontologies have acquired quite relevance in the knowledge organization field. Thus, some general and specialized ontologies have been developed for different tasks such as documents classification, general and specialized information retrieval and computer-assisted translation. The best-known ontologies are Cyc (Lenat *et al.* 1990), µKosmos[1], UMLS[2] and WordNet/EuroWordNet (Fellbaum, 1998; Vossen, 1998).

Recently, given the Internet progress, the ontologies are becoming a key element for a more accurate information retrieval. The World Wide Web Consortium[3] have issued a number of standard closely related to ontologies (XML, RDF, OIL, DAML, etc.) and some institutions have developed different tools in order to facilitate their implementation (Protegé, Ontoprise, OilEd, etc.).

In the genetics domain, there are some information databanks publicly available, such as, Gene Ontology[4], LocusLink[5], and GeneCards[6]. These resources include high-specialized data and they are an important assistance for domain researchers. However, their exploitation is difficult for other type 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 (Meyer et al., 1992; Condamines et al., 2000).

The main goal of this paper is to show the general structure of the GENOMA-KB. It differs from the resources mentioned above by a wider and easier access to the specialized information contained. Firstly, the motivations and the general structure of this knowledge base will be introduced. Secondly, each of the modules that constitute the whole knowledge base will be described. Finally, we will discuss some aspects about the query system that is used to retrieve information from the GENOMA-KB. Future research and some final conclusions will be issued at the end of the paper.

The GENOMA-KB description

The GENOMA-KB is the result of a ongoing research project at the Institute for Applied Linguistics. This project aims to establish the main theoretical ground basis and some refined strategies in order to improve terminological units retrieval, the emergence of cognitive nodes from texts and the detection of conceptual relations among terms in a semiautomatic way. The long-term goal is to improve the specialized information retrieval systems. Thus, the GENOMA-KB becomes an essential resource for the information retrieval with terminological control.

The GENOMA-KB is built upon four different and independent modules: the ontology module, the term base module, the corpus module and the entities module. All these modules are interrelated as shown in Figure 1.

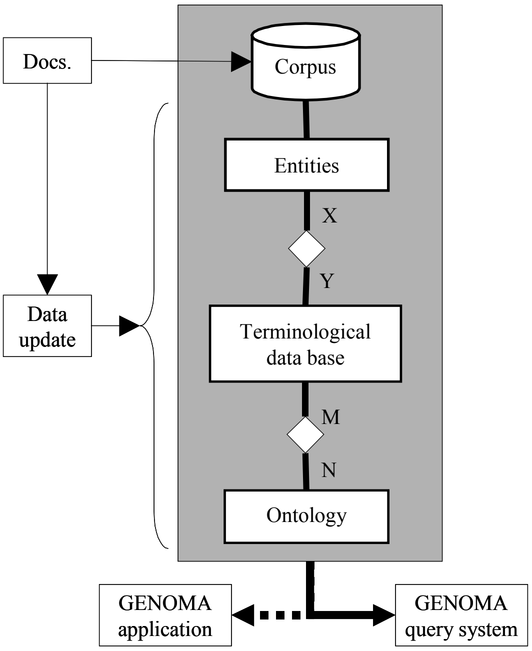


Figure 1: General structure of the GENOMA-KB

In this paper, the main relations among the modules will be depicted and the tied relation between the ontology and the term base will be detailed. At present, the GENOMA-KB gathers more than 1.000 concepts on the ontology,

[1] http://crl.nmsu.edu/Research/Projects/mikro/ (not publicly available)
[2] http://umlsinfo.nlm.nih.gov
[3] http://www.w3.org
[4] http://www.geneontology.org/
[5] http://www.ncbi.nlm.nih.gov/LocusLink/
[6] http://bioinfo.weizmann.ac.il/cards/index.html

2.000 terms in the term base, 7.6 M tokens in the corpus and near 450 registers in the entities database.

The Ontology module

From our theoretical point of view, terminology work deals with the specialized knowledge nodes represented by the lexical units in a specialized text (Cabré, 2002). These terminological units have clear-cut borders in each specialized domain and they correspond to a concept placed in an ontology. From this approach, we adopted a tool for managing together the terminological units used in the genetics domain and their correspondent concepts in the ontology.

After having reviewed the available resources for terminological management and ontology building, we have decided to use OntoTerm[7] (Feliu; Vivaldi; Cabré, 2002b). 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.

A core ontology was built with the aid of a domain expert who has provided its initial structure for the conceptual structure building. New concepts have been added to a previous list of base concepts necessary for the system performance. More precisely, the system includes 21 base concepts (ALL, OBJECT, EVENT, PROPERTY, etc.). 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.

The actual creation of the ontology has included a set of starting concepts together with a number of glue concepts to appropriately integrate the specialized ones. New concepts introduced in the ontology have been obtained from textual specialized information and with the aid of specialized linguistic resources. Concepts are represented in the ontology by means of English labels. The terms corresponding to these concepts in the term base module have been retrieved into three different languages (Catalan, Spanish and English) and each term is semantically described by is conceptual relations established in the ontology.

In the Ontology Editor, the user can introduce for each individual concept a brief description (concept definition) and the conceptual relation hyponymy-hyperonymy that is the fundamental relation in order to introduce each new concept. Concepts are fully described with the use of other conceptual relations, properties and the inherited information from parent concepts. The system organizes the information on the basis of the concepts, their attributes and the conceptual relations among them. Attributes and relations can be locally assigned or they can be inherited. Talking about inheritance, the system allows multiple inheritance and, in this sense, a concept can receive some properties and relations coming from more than one parent concept.

It is worth mentioning the research carried on in order to fulfil the ontology with the real use of the concepts in specialized texts by means of the inclusion of a new list of conceptual relations (Feliu, 2004 and Feliu; Cabré, 2002) used to link concepts introduced in the ontology:

- *Similarity*: is similar to, is different to
- *Hyponymy*: is hyponym of
- *Place and time sequenciality*: is located in, goes to, occurs with, occurs after
- *Causality*: causes
- *Instrumentality*: is used for
- *Meronymy*: is part of (is component of, is member of, is portion of, is material of, is stage of, is place of)
- *Association*: generally associated with, is specially associated with

Figure 2 shows the different conceptual relations established among the concept 'cell' and some other concepts that give account of the semantic information related with this concept.

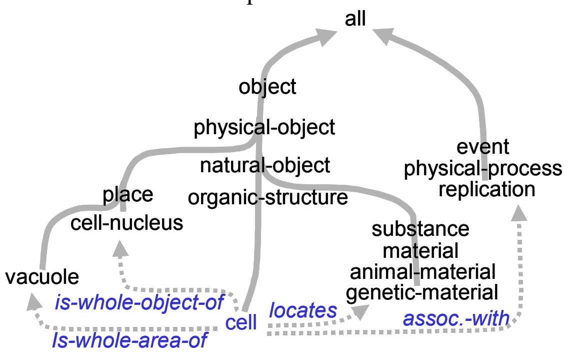


Figure 2. Conceptual relations assignment to the 'cell' concept.

Main comments on Figure 2 concern the diversity of conceptual relations used in order to describe the 'cell' concept. This kind of information will be very useful for a domain expert but mainly for lexicographers and terminologists interested in reusing this type of information in order to create a terminological application containing, for example, term definitions.

The Term Base module

As mentioned above, the term base and the ontology modules are closely tied both by the theoretical approach and also by the design decision of the application used for managing the terminological and the conceptual information. Therefore, no term entry is possible if its corresponding concept has not been previously introduced in the ontology.

The information given for each terminological unit in the term base (see Figure 3) is the concept expressed by the term; the term itself in Catalan, Spanish and English; the part of speech; the number and gender assignment; the usage contexts and their sources; the lemmatised form, and some administrative data, which are mandatory. The term definition and its source and some usage notes are optional.

As shown below, there is a particular source for each data category. The "source id" indicates an occurrence of the term in the Corpus; the full bibliographical data is located in the Entities module. The non-mandatory definition comes from specialized dictionaries. Finally, term contexts (up to three) are retrieved from the Corpus module when available; otherwise, from Internet.

[7] OntoTerm is a terminological management system built by Antonio Moreno, from the Universidad de Málaga. More information available at: http://www.ontoterm.com.

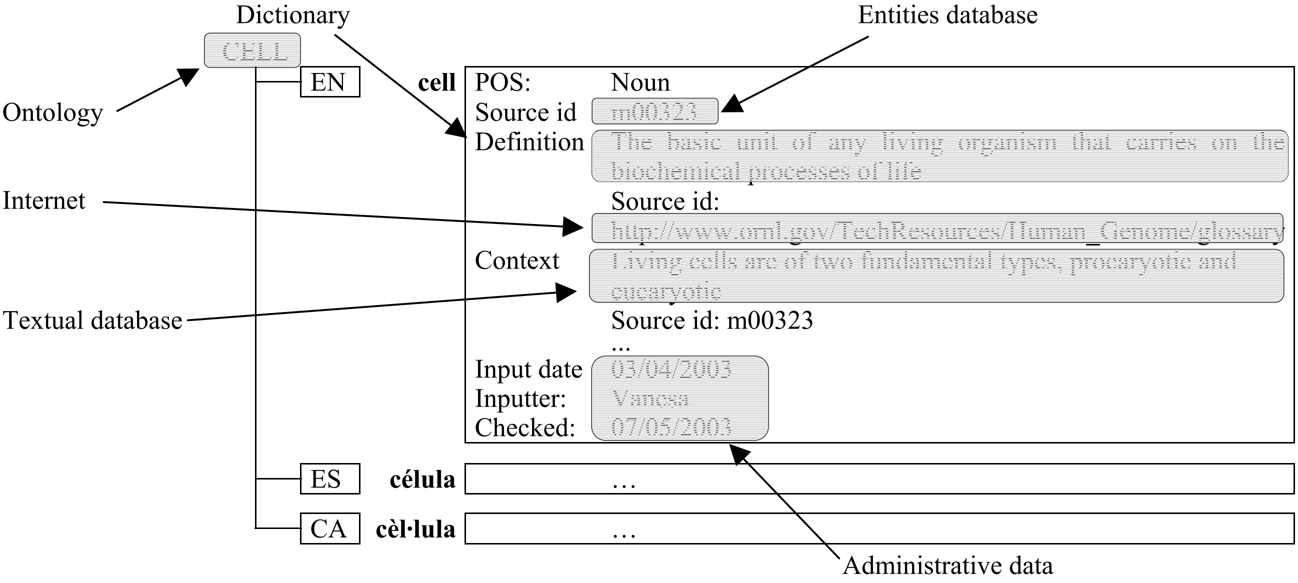

Dictionary
Entities database
CELL
Ontology
EN
cell
POS: Noun
Source id m00323
Definition The basic unit of any living organism that carries on the biochemical processes of life
Internet
Source id:
http://www.ornl.gov/TechResources/Human_Genome/glossary
Context Living cells are of two fundamental types, procaryotic and eucaryotic
Textual database
Source id: m00323
...
Input date 03/04/2003
Inputter: Vanesa
Checked: 07/05/2003
ES
célula
…
CA
cèl·lula
…
Administrative data

Figure 3. Data sources for the term database

The Corpus module

Following the revival of the empiric methods usage in NLP research, IULA has been compiling an LSP corpus (Badia et al. 1998). The texts belonging to this corpus have been linguistically processed as usual in NLP applications. All documents have been classified according with a conceptual structure validated by some domain experts.

One of the branches of such corpus concerns the genomic domain. All the documents included in this subcorpus are registered in the Entities database module. Moreover, specialized knowledge fragments are included in the term base as the term entry usage context.

At present the genomic subcorpus contains texts into Catalan, Spanish and English. Near one million tokens per language have been collected.

The Entities database module

The presence of this module in the GENOMA-KB is specially significant as it complements terminological and textual corpora information in an unusual way in these type of projects. The module is organised in two different parts: a bibliographic module and a factographic module.

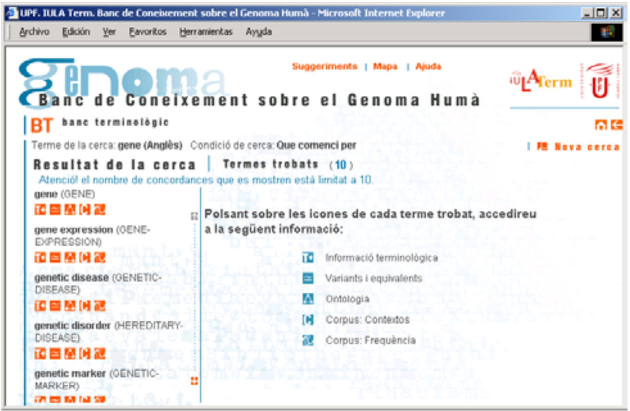
The main focus of the bibliographic module is to compile full references of all information sources used in the term base and corpus modules in order to complement their contents. It includes basically monographs, dictionaries, bibliographies, proceedings, articles from specialized journals and Ph dissertations. The formats of theses sources are paper, electronic resources and microfilms. The language of the references are also Catalan, Spanish and English, as for the case of the corpus and term base modules. At present, the number of registers contained in the bibliographic database are: more than 300 articles, and near 80 monographs, 30 specialized journals and 30 Ph dissertations.

And finally the factographic module collects continuously updated data about relevant research centres such as universities, enterprises and public funded organisations and domain experts (around 80 registers at this moment).

The Query System

Concerning the specialized information visualization of the ontology, the term base, the corpus and the entities database we have developed GENOMA an on-line query system called. The web site foresees three different query levels: simple, complex and combined. The first option is already accessible and will be soon available from the IULA's website[8]. Complex and combined queries are still under development.

In the case of the simple query, information is retrieved from each of the different modules that are linked among them. Thus, if we search the term 'gene' into English, the system proposes all the linguistic information contained in the term base module, the link between the term and its corresponding concept and the position of this concept in the conceptual structure, that is, its position on the ontology and all the conceptual relations and attributes assigned to this concept. Moreover, the user has access to the IULA's technical corpus and it is possible to retrieve the contexts and the frequency of the selected term. Lastly, it is also possible to visualize the full



UPF. IULA Term. Banc de Coneixement sobre el Genoma Humà - Microsoft Internet Explorer

genoma Banc de Coneixement sobre el Genoma Humà

BT banc terminològic

Terme de la cerca: gene (Anglès) Condició de cerca: Que comenci per

Resultat de la cerca | Termes trobats (10)

Atenció! el nombre de concordances que es mostren està limitat a 10.

gene (GENE)

gene expression (GENE-EXPRESSION)

genetic disease (GENETIC-DISEASE)

genetic disorder (HEREDITARY-DISEASE)

genetic marker (GENETIC-MARKER)

Polsant sobre les icones de cada terme trobat, accedireu a la següent informació:

- Informació terminològica
- Variants i equivalents
- Ontologia
- Corpus: Contextos
- Corpus: Freqüència

[8] GENOMA-KB is available at the following url: http://brangaene.upf.es:8080/genoma/index.jsp

bibliographical references of the textual contexts containing the term ‘gene’.

Figure 4 shows the result of querying the system with all the terms that start with the string “gene”. For each of the found terms, the system proposes five icons giving access to the following information:

- T Term database content
- ≅ Ontology information
- △ Term variants and equivalents in other languages
- C Context found in the corpus
- ∿ Frequency count in the corpus

Figure 4. GENOMA-KB query system interface

Conclusions and future research

In this paper, we have presented the main characteristics of a resource combining semantic information (ontology) and terminological information (term base) together with actual contexts of use and bibliographical information. Thus, ontology and term base modules are related with the textual database (specialized multilingual corpus) and the entities database.

However, this resource is not a final goal in itself. It has been designed and developed in order to be queried and used by researchers working on different knowledge domains, that is, human genome domain experts, linguists (terminologists and lexicographers) and linguistic mediators (translators, reviewers and journalists) that require some kind of specialized knowledge.

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 public funding of 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.

References

Badia (1998) Badia, T.; Pujol, M.; Tuells, A.; Vivaldi, J.; de Yzaguirre, L. and Cabré, T. (1998) “IULA’s LSP Multilingual Corpus: compilation and processing”. ELRA conference, Granada, 29-31 May. http://www.iula.upf.es/corpus/corpubca.htm

Cabré, M. T. (2002) «Análisis textual y terminología, factores de activación de la competencia cognitiva en la traducción». In: Alcina Caudet, A. et S. Gamero Pérez (eds.) La traducción científico-técnica y la terminología en la sociedad de la información (pp. 87-105). Castellón: Publicacions de la Universitat Jaume I.

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.

Fellbaum, C. (ed.) (1998) WordNet: An Electronic Lexical Database. Cambridge: MIT Press.

Feliu J. (2004). Relacions conceptuals i terminologia: anàlisi i proposta de detecció semiautomàtica. Barcelona : Institut Universitari de Lingüística Aplicada. [Ph. Dissertation]

Feliu, J.; Cabré, M. T. (2002) «Conceptual relations in specialized texts: new typology and an extraction system proposal». TKE2002. Terminology and Knowledge Engineering Proceedings. 6th International Conference. 28th-30th August 2002. Nancy, p. 45-49.

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.

Lenat, D. ; Guha, R. (1990) «Building Large Knowledge Base Systems : Representation and Inference in the CYC Project ». Addison Wesley.

Meyer, I.; Bowker, L.; Eck, K. (1992) Cogniterm: An Experiment in Building a Terminological Knowledge Base. Proceedings 5th Euralex International Congress on Lexicography. Tampere, Finland.

Vossen, P. (ed.) (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht: Kluwer Academic Publishers.