Towards an Ontology for a Human Genome Knowledge Base

Judit Feliu, Jorge Vivaldi, M. Teresa Cabré

Institute for Applied Linguistics, Universitat Pompeu Fabra
La Rambla 30-32, 08002 Barcelona, Spain
judit.feliu@iula.upf.es; jorge.vivaldi@info.upf.es; teresa.cabre@trad.upf.es

Abstract

Ontology, usually understood as a particular representation of a given domain, will become an essential item in the information retrieval system we aim to build. Our research activities are developed on the communicative terminology framework, that is, we mainly deal with units effectively contained in specialized discourse. Bearing in mind this theoretical approach, we consider essential to establish a link between the specialized knowledge units appearing in specialized texts and the concepts organized in a particular ontology. Having the specialized knowledge units closely linked to a conceptual organization will lead us to propose an information retrieval system based on a Human Genome Ontology that should perform better than the current state-of-the-art systems.

1. Introduction

The aim of this paper is three fold: first, we will briefly describe two of the actual projects carried on by the IulaTerm group[1] in the Institute for Applied Linguistics (IULA). Second, we will summarize the results obtained from an analysis of some already existent ontologies. Lastly, we will briefly introduce main features of the ontology we are currently building.

As for the ongoing projects (section 2), we are working with the final objective to build a Human Genome Knowledge Base integrating the following four modules: a textual database that contains specialised texts of this particular domain; a factographic and documental database containing the metainformation 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. This paper will focus on the process of building the ontology module.

At the moment of starting to work on the ontology module, we realised that it was necessary to review some of the main existent ontologies in order to analyse the characteristics that we have to consider before the design of our ontology. For this reason, we established the main parameters to take into account and we compared several ontologies in order to decide whether it was possible to reuse and/or to expand an already built ontology or if it was more useful for the project purposes to build a new ontology. In section 3, we summarize some results obtained from our analysis.

In section 4, we point out some of the characteristics of the ontology we are building and we emphasize on the way this ontology can be expanded in a future in order to cover some other specialized domains.

2. Project Description

The Institute's ongoing project, the so-called Human Genome Knowledge Base Project, is carried on within the framework of two public funded projects[2]: TEXTERM and RICOTERM. The TEXTERM project aims to go a step forward in discourse, grammar and semantic analysis of specialised texts. It is more specifically devoted to the characterization of the lexical (simple or complex) and phraseological units, which constitute the terminology of those domains, with the final purpose of building an automatic detection system of the cognitive underlying structures in specialised texts. The main goal of this first project is to provide a sound theoretical basis for computer-aided unit detection and semi-automatic mapping of cognitive nodes and conceptual relations. It is foreseen that our working methodology —oriented to improve information retrieval (IR) systems— would combine strategies both from the cognitive sciences and linguistics. We will also resort to indexation strategies and thesaurus building standards, coming from information science, and some other linguistic engineering working lines, such as natural language processing and statistical analysis.

Traditionally, most information retrieval systems have been based on strategies of formal string detection, complemented with the statistical analysis of text properties. These systems have some constraints because they do not use the semantic and pragmatic information associated to these strings and their context. For this reason, the main objective of the RICOTERM project is to build an IR system prototype, capable of improving current systems using terminological control. We hope to reach such an objective by taking profit of the grammatical, semantic and pragmatic information associated with the units that convey specialised knowledge.

The methodology to be used should integrate a tool for natural language processing (NLP), which includes structural mark-up, morphological and syntactic analysis, disambiguation, and a terminology extraction system based on heterogeneous strategies combination and lexical ontologies (Vivaldi et al., 2002). Ground criteria will be refined by standards for the identification and mark-up of semantic and pragmatic elements within a restricted domain.

The two projects briefly described above are carried on bearing in mind one general goal : the construction of the

[1] The research activities of this group follow the theoretical approach to terminology established by Cabré (1999).

[2] 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.

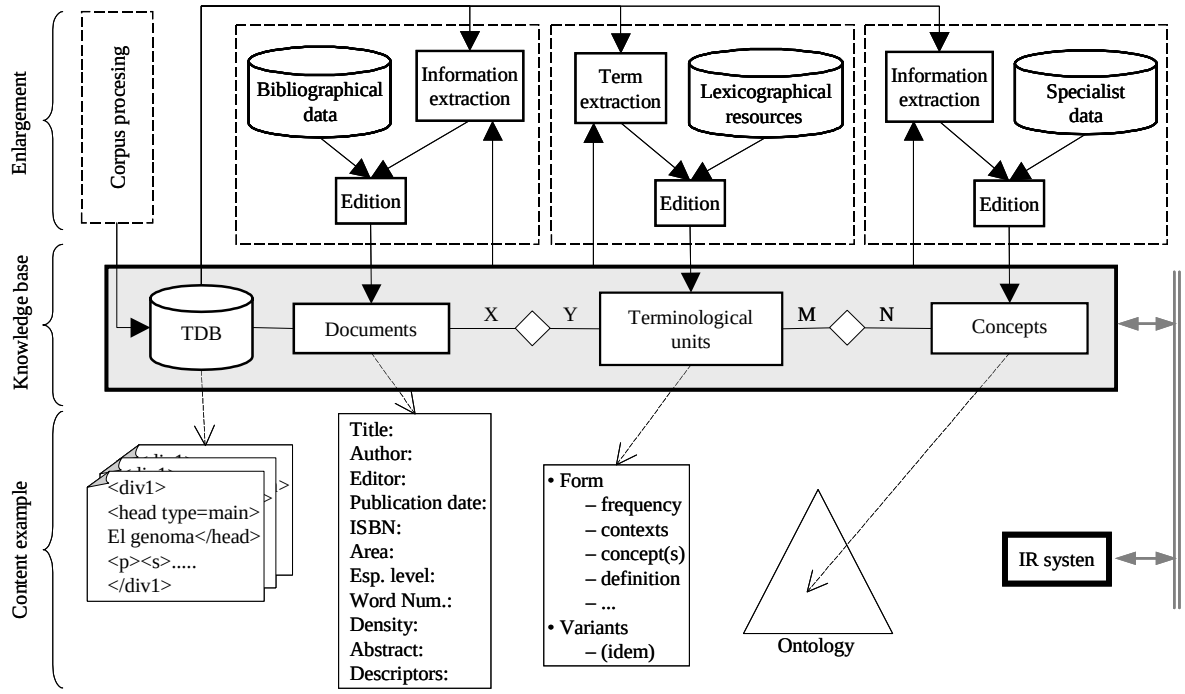

Figure 1: Human Genome Knowledge Base Project: an overview

Human Genome Knowledge Base, whose architecture is shown in Figure 1.

In this figure, we show 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.

For each of the above mentioned modules, we foresee different enlarging procedures. As shown in Figure 1, a particular corpus processing module is responsible for obtaining and processing specialized texts. At this stage, we rely on the experience gained in the compilation of an LSP corpus[3]. Concerning the three other modules mentioned, we are currently working on their developement in cooperation with domain specialists.

The organization proposed for the knowledge base allows a complete interaction between the four modules. Thus, the terms registered in the terminological database will be linked to both the ontology and the documents from where they have been retrieved. The resulting set of knowledge will be used for different tasks, such as, document indexation and summarization, machine translation support, etc.

3. Main already existent ontologies review

In our theoretical approach, terminology is defined as a set of specialized knowledge units used in a given domain. And a term is conceived as a set of relations centered in a lexical or lexicalised unit. For this reason, it seems plausible to use ontologies in order to map specialised knowledge contained in a domain-specific corpus and, consequently, to describe specialised knowledge units transferred by linguistic units (or terms) and the relationships among them.

The first step to be taken was to check if it was possible or not to reuse an already built ontology. As a working starting point, we set main parameters to be taken into account to evaluate and create an ontology (section 3.1) and, on the basis of these parameters, we compared five existent ontologies (section 3.2).

3.1. Design criteria

Specialised knowledge mapping is a hard task and most efforts must be directed in the ontology design. Ground principles become an essential working point and some of the basic decisions to be taken concern:

a) The coverage required to the ontology: i.e., number of concepts collected.
b) The end purpose of the application that will use the ontology: i.e., the characteristics of the ontology (domain, coverage, node representation, etc.) behind a particular tool will be determined by the application constraints. The requirements for an ontology used in a semantic web or in a

[3] See the following URL for details about this corpora: http://www.iula.upf.es/corpus

machine translation system would be very different.

c) Top nodes of the ontology. Traditionally, top nodes of ontologies have been entities, properties and relations. However, in some cases the number of top nodes may increase and differ (for example, WN uses eleven tops and it does not include relations among them).

d) The conceptual relations allowed in the ontology. "Is-a" is the basic relation of any ontology but some other conceptual relations are also possible and even necessary for some applications. Then, the number of conceptual relations is not a closed list. Some general relations such as meronymy are generally used while specific relations such as "affects" are more present in specific domain ontologies (see UMLS for medicine). Enlarging the number of relations enriches the ontology but it makes it difficult to maintain consistence.

e) Use of inheritance. Inheritance is a general mechanism to add information to a particular node in a compact and easy to maintain way. According to this mechanism, the corresponding node and all its hyponyms share such information. The "simple monotonic inheritance" is the simplest mechanism. It means that each node inherits properties only from a single ancestor and the inherited value cannot be overwritten at any point of the ontology. This inheritance method has problems to manage real situations (like exceptions handling). This situation may be overcome by using "multiple inheritance" (each node may inherit properties from one or more ancestors) and/or "default inheritance" (a node may locally overwrite the value of an inherited property). Contradiction arises when a node inherits incompatible values for a single property coming from different ancestors. Mechanisms mentioned above do not solve this problem but some solutions mainly based on a deep control of the hierarchy have been proposed. For example, the "orthogonal inheritance" suggests gathering the data and allowing multiple inheritance from different groups only.

f) Node representation. Concepts may be indicated by means of a label (case letter, numbers, etc.) or structured information (feature structure).

Traditionally, and besides the above mentioned criteria, ontologies are usually classified from different points of view:

- general (i.e., WN [Fellbaum *et al.*, 1998]) or domain specific (i.e., ULMS [NLM, 1998]),
- generic (i.e., EWN) or built for a particular application (i.e., µKosmos),
- episodic or encyclopedic ontologies (i.e., Cyc [Lenat *et al.*, 1990]),
- lexical (i.e., EWN) or conceptual (i.e., Cyc) ontologies.

The analysis of the above mentioned ontologies has allowed us to isolate the main characteristics concerning their criteria design and general structure. Our analysis[4] covers in depth the following five well known ontologies: Cyc, EuroWordNet, µkosmos, SIMPLE[5] and UMLS.

3.2. Comparative analysis

In this section, and having reviewed most outstanding features of the five former selected ontologies, we will analyse some of the key parameters that have to be taken into account in order to evaluate an ontology. It has to be pointed out that since they are very different ontologies, a direct comparison is a very difficult task.

However, some characteristics are in fact comparable. In this sense, the elements reviewed in the comparative analysis are the following: management facilities, expressiveness, application field, ontology type and size, granularity and completeness. We want to explicitly mention at this point that, from now on, all information given about µKosmos has been extracted through the management tool OntoTerm®, which has allowed us to have access in depth to the ontology organization.

3.2.1. Management facilities (enlargement and modification)

A very important aspect in developing an ontology is the availability of tools helping to keep consistency in the whole system. This section reflects the tools that could be used to update each resource. As far as we know, the available tools are the following:

Cyc. No indication has been found about the existence of management tools.

EuroWordNet. For Spanish and Catalan versions of EWN, there are some management tools, mainly designed to enlarge the ontology. There is also a browser in Internet.[6]

µKosmos. The tool used for this evaluation is OntoTerm®, an ontology management application. It provides a user friendly interface for adding concepts, relations and lexical entries.[7]

SIMPLE. In Bel *et al.* (2000) it has been mentioned the existance of some tools at least for the Spanish and Catalan language.

UMLS. The only tool included in the UMLS distribution is MetamorphoSys, a systems that allows to customize and create subsets of the UMLS Metathesaurus in order to better meet the user needs.[8]

3.2.2. Expressiveness

All ontologies analysed present very different types of formalisms. One of the main distinctive parameters in order to evaluate these ontologies is the concept and the expression of relations in each of these formalisms.

[4] See Feliu et al. (2001) for a more detailed description of each ontology.

[5] SIMPLE is not oriented to build an ontology as it is understood in this paper. In contrast, it represents an attempt to encode lexical semantics information for an important number of languages.

[6] The browser is reachable at the following url: *http://nipadio.lsi.upc.es/cgi-bin/public/wei2.html*.

[7] A demo version of this tool is available at: *http://www.ontoterm.com*.

[8] Some possible user needs are to exclude vocabularies as required for License Agreement, to exclude non useful vocabularies, to personalize the resource, and so on.

A brief comment on these characteristics is indicated below:

Cyc: It uses CycL, a representation language, which is essentially a form of First Order Predicate Calculus with some additional features such as: equality, augmentations for default reasoning, skolemization, and some second-order features (e.g., quantification over predicates is allowed in some circumstances).

EuroWordNet: It describes concepts (called synsets) as a set of variants. There are a finite number of relations and its management tool is restrictive about the type of relations included. It defines a top ontology according to main lexical-semantics[9] principles. Semantic information for each concept is inherited from its ancestors except for the cases where some parts of this information are redefined.

µKosmos: Concepts are described by their position in the ontology and by the indication of their properties and values.[10] Relations are not restricted in number but it is required to define, for each direct one, the corresponding inverse relation. µKosmos allows multiple inheritance which, using the management tool, can be visualized as exclusive or cumulative for every concept.

SIMPLE: Each lexical unit is described using a system of types organized through the principles of orthogonal inheritance (according to Pustejovsky 1995). All semantic information is added to refine linguistic information (i.e. semantic types for each kind of argument, relations between semantic units).

UMLS: Each concept placed in the semantic network is just described by a denominative tag. Concepts are related among each other by a rich set of medical-specific, controlled number of relationships. UMLS presents *a priori* a simple inheritance mechanism but it is possible to block this process when needed.

Two different groups can be differentiated from the analysed ontologies. A first group with ontologies that have hierarchies and information associated to each node of the hierarchical structure (i.e.: EWN, µKosmos and UMLS). A second group, constituted by the other two ontologies mentioned (Cyc and SIMPLE), where the information is quite differently organised and represented.

However, all ontologies include some kind of definition for the concepts contained. The expression of definitions in natural language is given in a number of different ways: formal definition, glossa, examples, explanatory context, and so on.

3.2.3. Application field

Most of the ontologies analyzed in this paper are not domain-specific. Keeping aside UMLS, which is devoted to the medicine domain, all other ontologies cover general information. In spite of the later consideration it has to be mentioned that the general ontologies do not have all the domains equally developed. Thus, µKosmos has considerably developed those branches of the ontology concerned with the joint-venture domain. Finally, EWN has asymmetrically developed the different domains it tackles.

3.2.4. Ontology Type

Talking about the ontology type, it is important to notice that both EWN and SIMPLE are conceived from the point of view of the lexicon, that is, they are lexical ontologies. Conversely, µKosmos, UMLS and Cyc may be classified as conceptual ontologies. Except for the later, information is represented by concepts which are expressed with different labels containing all information required (see expressiveness above) in order to convey their meaning.

3.2.5. Size, granularity and completeness

The size of all the resources analysed is very different. Table 1 shows the global sizes for each resource in the different languages considered.

Resources	Ontology	Size for each language		
		English	Spanish	Catalan
Cyc	3.000	14.000	0	0
EWN	∅	90.000	50.000	20.000
µKosmos[11]	4.800	0	0	0
SIMPLE	∅	?	3.000	3.000
UMLS	134	800.000	30.000	0

Table 1. Analysed resources: size comparison

4. Genome Ontology Building Process Outline

In this section, we make some brief comments on the decision to build a new ontology taking profit of some information and design parameters of the ontologies reviewed. Finally, we also present some relevant aspects of the new ontology we are still working on.

Aiming at a general ontology that allows enlargement, we had to leave aside Cyc, UMLS and SIMPLE. In the case of Cyc, it is produced by a private company and it is not publicly available. Moreover, and according the information retrieved from literature, it seems difficult to deal with. As for UMLS, it is a domain-specific ontology about medicine. SIMPLE is oriented to add lexical semantics information to a dictionary and it can not be considered as an ontology itself. However, UMLS has been an important source to enlarge the selected ontology for our project and SIMPLE has also been useful in order to complete and refine linguistic information for NLP.

Both remaining ontologies, µKosmos and EWN, are general domain ontologies that satisfied the basic

[9] These semantic principles of the lexicon are established following Pustejovsky (1995).

[10] A natural language definition of most concepts can be visualized using the management tool OntoTerm®.

[11] There are a number of lexical modules (English, Japanese and Spanish) for µKosmos but the number of lexical entries is not indicated. In OntoTerm® implementation, the system provides a tool for including all lexical information for many languages (the system provides a picking-list of ISO language codes) related to a particular concept. Lexical information is organised according to the languages concerned using a previously designed template.

requirements of the IULA's ongoing project. In spite of this, we found important differences between such resources. Table 2 indicates the most salient parameters of both resources.

Parameter	μKosmos (OntoTerm®)	EWN
Type	Conceptual	Lexical
Completeness	Medium	High
Coverage (Medicine)	Low	High
Implementation OS	Windows	Unix

Table 2. Main characteristics of μKosmos and EWN.

It is obvious that any ontology is tied to its management tool. So, the final selection had to take into consideration both aspects: facility and adequacy of the ontology and an appropriate management environment for its development.

The complete analysis and comparison lead us to follow the μKosmos design adopted by OntoTerm® because it is conceived on the basis of two separated modules: the ontology (ontology module) and the lexicon (terminological module). We believed that this approach would fulfill our project requirement. However, some of the reviewed ontologies are being used as a reference for enlarging our ontology, mainly UMLS and EWN. EWN has also become a pattern to follow in the specification of conceptual relations and in the treatment of some non-nominal units such as verbs and adjectives. Another essential information source is the support given by a specialist from this domain.

OntoTerm® is the only available management tool related with the ontologies reviewed that allowed us to build an ontology integrating, at the same time, the ontology and a lexical resource, that is, a terminological database. See Moreno et al. (2000) for a complete description of this tool.

We have used OntoTerm® to create the ontology and we have already introduced about 260 concepts. This tool provides 21 preestablished core nodes which are a system constraint derived from the implementation of basic top nodes of μKosmos (i.e., all; event; object; property [top nodes], etc.).

A brief example of conceptual organization in the ontology is depicted in Figure 2. It shows for the concept 'transcription' the 'is-a' relation, as well as some other conceptual relations kept with other nodes of the ontology. The non-hierarchical conceptual relations defined in this ontology directly derive from those presented at Feliu (2000). A part from the hierarchical relation, we indicate the meronymy relation ('stage-of') as one of the six types of the 'part-of' relation. Moreover, it is also indicated the instrument relation by means of 'used-for', and finally, the associative relation, expressed by 'associated-with'. The latter relation deserves a particular comment because it will be useful in order to determine one of the possible senses of a concept used in more than one domain. Imagine the concept 'transcription' which can be at least linked to 'DNA' and 'book' concepts. In our approach, we indicate as many associations as possible and, in a further step, the proper sense of 'transcription' will be emerged according to the domain instantiation of the concept related. We hope the use of this mechanism will allow us to introduce some other domains in our ontology. In this case, there will be a number of concepts common to all domains and some other which would be differentiated. This working methodology should allow us to use the same ontology in an information retrieval system covering one or more domains.

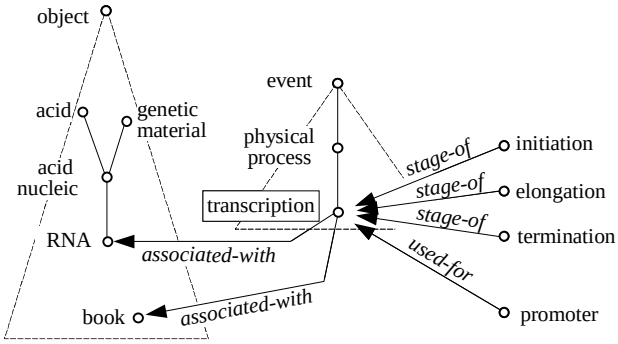

Figure 2. Sample of the concepts organisation

5. Conclusions and future work

In this paper we have described our research work carried on in order to build an ontology. After the mandatory previous review of the existent ontologies, we have presented main design criteria for the selection of a particular ontology and its management tool.

We are aware it is the first step of a long-term project. Future work will follow two main research lines. On the one hand, considerable efforts will be devoted to enlarge and to link the four different modules of the Human Genome Knowledge Base. Terms will be extracted from the textual database and linked to the ontology, which will be updated when necessary. On the other hand, incipient work will be carried on the final application of the project. Thus, it will be necessary to design main lines of the information retrieval system using this ontology. Once this application will be finished, we hope to be able to confirm the initial hypothesis that an IR system using terminological control performs in a more refined way.

6. Acknowledgements

We would like to thank Antonio Moreno for all his indications and interest in order to facilitate us the access to OntoTerm®. We also want to specially mention Marie-Claude L'Homme for her interest and valuable discussions on the subject.

The authors acknowledge the invaluable contribution of Eva Valero, for her help in the ontology construction.

7. References

Bel N. and M. Villegas (2000) "An introduction to SIMPLE". IULA's Workshop, December 2000.

Cabré, M. T. (1999) *La terminología. Representación y comunicación. Elementos para una teoría de base cognitiva y otros artículos*. Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada. Barcelona.

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

Feliu, J. (2000) *Relacions conceptuals I variació functional: elements per a un sistema de detecció automática.* Ph. Dissertation draft, non published. Barcelona: Institut Universitari de Lingüística Aplicada.

Feliu, J.; J. Vivaldi and M. T. Cabré (2002). *Ontologies: A Review*. Research Working Paper. Barcelona: Institut Universitari de Lingüística Aplicada.

Lenat D. y R. Guha (1990) «Building Large Knowledge-based systems: Representation and Inference in the CYC project». Addison Wesley.

Moreno Ortiz A. y C. Pérez Hernández (2000). "Reusing the Mikrokosmos Ontology for Concept-Based Multilingual Terminology Databases". Proceedings of 2nd International Conference on Language Resources and Evaluation (LREC-2000). Atenas, 31 de mayo-2 de junio, pp. 1061-1067.

NLM (1998) "UMLS Knowledge Sources". National Library of Medicine. U.S. Dept. of Health and Human Services, 8th edition.

Pustejovsky, J. (1995) *The Generative Lexicon*. Cambridge, Massachusetts / London, England: The MIT Press.

Vivaldi J. and H. Rodríguez (2002). "Improving Term Extraction by combining Different techniques". Terminology, Vol. 7-1. John Benjamins Publishing Co. pp. 31-47.