Aguado de Cea G, Montiel-Ponsoda E

Benefits of Ontologies to Multilingual Needs

6th International AELFE Conference
Editors: Maria Kuteeva, Helder Fanha Martins.
September 13th-15th, 2007.
Lisbon, Portugal.
ISBN: 978-972-98646-1-2. Pages: 519-531

Benefits of Ontologies to Multilingual Needs

Aguado de Cea, G.; Montiel-Ponsoda, E.

Abstract

The way in which multilingual information is organized and presented, accounts for its usefulness or adequacy for a specific purpose. As globalization becomes more pervasive, people all over the world need to make use of information in different languages in their everyday work. Bilingual or multilingual dictionaries have been, and still are, relevant resources for facing up multilingual issues. However, the multilingual information collected in dictionaries remains insufficient for those people who need to gain a general view of a specific parcel of knowledge in two or more languages. In LSP (Language for Specific Purposes) this issue becomes even more relevant. In the case of translation of specialized texts, specialization on source and target subject matter becomes imperative in order to get a complete understanding of the source text and transfer that knowledge to the target reader. Ontologies may come to solve this knowledge acquisition problem, since they offer a multilingual conceptualization of a specific parcel of knowledge by organizing the information according to the different and various relations between concepts. In this way, translators are able to gain the required domain knowledge, as well as the type of equivalence relations between concepts in the different languages, and their context of use. Thus, pursue of this paper is to give an overview of the benefits of multilingual ontologies to the multilingual information retrieval.

Multilinguality, multilingual dictionaries, multilingual ontologies, translation of specialized texts

1. Introduction

Though accepting that the translation of specialized documents requires the sufficient specialized knowledge on the domain, specific subject fields can always be approached from different perspectives, or they change or develop in other directions, and translators need to be constantly training in new subjects. This requires a great effort and a non worthless amount of time, if we consider that they need to repeat this training in all their working languages. As Wilss states (1996: 58): *Whether translators (…) understand an LSP text depends, apart from familiarity with the respective terminology, upon their knowledge of the respective domain. This may be a simple truth, but simple truths may imply consequences which are far from being simple of trivial*. This is why translators, apart from having an excellent command of the terminology of a domain, they need to be in continuous process of training in new specific topics. However, existent lexical resources are of little use in helping translators to obtain the required multilingual knowledge.

The possibility of offering translators the specific domain information they need in different languages, with the guarantee of being previously agreed by experts in the field, and with an explicit reference to relations between concepts, could only be performed by a machine-readable resource modelling a parcel of knowledge, i.e. by an ontology. As the most quoted definition of **ontology** in Artificial Intelligence literature says: an ontology is “an explicit specification of a conceptualization” (Gruber 1993). Studer and colleagues (Studer et al. 1998: 185) widened this idea to state that:

> *Conceptualization* refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon. *Explicit* means that the type of concepts used, and the constraints on their use are explicitly defined. *Formal* refers to the fact that the ontology should be machine-readable. *Shared* reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted group.

On the order hand, and according to the Merriam-Webster Online Dictionary, a dictionary is “a reference source in print or electronic form containing words usually alphabetically arranged along with information about their forms, pronunciations, functions, etymologies, meanings, and syntactical and idiomatic uses”. Bilingual dictionaries offer as well an explanation or translation of the headword in another language.

We can state therefore that ontologies differ from dictionaries in various aspects, as summarized in Table 1. The most evident one is that dictionaries organize terms following an alphabetical order, and that ontologies arrange their concepts (and not terms) in a hierarchical way around a unique concept, superordinated to the rest. Moreover, the established relationships among the concepts of the ontology capture the semantics of the domain, which have been previously accepted by experts on the field. In the case of dictionaries however, semantics between lexical units are limited to the implicit information offered by definition, etymologies or usage examples. And in fact, in dictionaries only the human user is able to find out those relations among concepts, whereas in ontologies those relations have been formalized, i.e., cannot only be interpreted by humans, but also by machines (Arano, 2005).

CLASSIFICATION CRITERIA	DICTIONARY	ONTOLOGY
Organization	alphabetical order	semantically related lexical entries
Semantic information	definition + pos + etymologies + derivation + usage examples in NL	explicitly defined hierarchy relationships around a unique concept
Physical format	paper + electronic format	electronic format (readable also by machines)
Domain of knowledge	general + specific	general + specific (agreed by domain experts)

Table 1: Comparison of dictionary and ontology

After this brief comparison between dictionaries and ontologies, we now aim at carrying out a deeper analysis of a representative example of both types of resources. In what follows in this paper we will compare the authoritative multilingual on-line dictionary of the European Commission (EC), EURODICAUTOM, to the multilingual knowledge base GENOMA-KB, developed at the Universitat Pompeu Fabra, in Barcelona, Spain, in order to determine the benefits of the organization and presentation of multilingual information in ontologies against the classic representation offered by dictionaries.

2. Short description of Eurodicautom

Eurodicautom was first set up in 1973 and it was the result of the cooperation work of terminologist, translators and computer science experts of the EC. Currently, Eurodicautom covers twelve languages, eleven official languages (Danish, Finnish, Greek, Portuguese, Dutch, French, Italian, Spanish, English, German and Swedish) and Latin, containing about five million terms and two hundred thousand abbreviations. This dictionary aims at meeting the demands of the EU of conferring the same recognition to each language.

Eurodicautom is particularly rich in technical and specialised terminology related to EU policy. Entries are classified into 48 subject fields, as for example, medicine or public administration, and each of them constitutes a technical dictionary.

2.1 Steps in the development of Eurodicautom.

1) The first steps were carried out by the translators themselves, as they used to elaborate technical cards of every technical term they came across. Source languages were mainly French and English, since these are the languages in which the EU documents are first drawn up.

2) Two lexical tools were merged to become the foundations of the Dictionary, DICAUTOM (launched in 1962) and EUROTERM (1964) both in French, German, Italian and Dutch.

3) The Terminology Bank of the University of Montréal, Canada, put at the disposal of the EC 80,000 bilingual cards (English-French).

4) Other glossaries and resources were as well merged (Goffin 1997).

5) In 1976 Eurodicautom was finally launched as a multilingual automatic dictionary, and when more countries joined the EU, the dictionary had to be enlarged continuously by a team of terminologists, specialized in the task.

   - The enlargement was made mainly *manually* by terminologists. Multilingual information was extracted from the multiple publications of the EC, especially from the Official Journals (manually at the beginning, *semi-automatic* in the recent years). This work was supervised by experts in the corresponding domain. Translators also contributed to the task by delivering computerized terminological cards, whenever new terms were introduced and translated within the EU institutions.

   - In 1995, with the introduction of the *Euramis* project (*European Advanced Multilingual Information System)*, a series of applications enabled translators to a more effective management of terminology, and provided access to a variety of services in the field of Natural Language Processing (NLP) - *translation memories* (Trados Multi Term), *mass processing of linguistic data, machine translation* (Systran), and *workflow automation*- and to the store and management of term bases.

3. Short description of GENOMA-KB

The Human Genome Knowledge Base (GENOMA-KB) is an ongoing research project started in 2001 at the Institute of Applied Linguistics (IULA) of the Universitat Pompeu Fabra, with the objective of developing a biomedical knowledge base for the human genome.

According to the developers, the resulting set of knowledge can be used for different tasks, such as, document indexation and machine translation support. Main target users are translators, terminologists and lexicographers; information science experts; researchers and scholars; linguists. The languages involved in this project are English, Spanish and Catalan.

3.1 Steps in the development of GENOMA-KB.

In order to understand the construction process of this resource, we need to describe its architecture in the first place. As shown in Figure 1, the knowledge base is divided in 4 interrelated modules:

- **Ontology module**. This module was developed using OntoTerm® (Moreno & Pérez, 2000), a terminological management tool that allows the construction of the ontology, integrating at the same time the ontology and the terminological database. This tool provided a core ontology with the 21 basic concepts from Mikrokosmos (ALL, OBJECT, EVENT, PROPERTY, etc.). A list of 100 concepts was then added to the initial ones, which were proposed by experts in the human genome domain. The rest of the concepts were recovered form textual specialized information with the aid of lexical resources. Concepts were fully described with the use of conceptual relations, properties and the inherited information from parent concepts.

- **Term base module**. The information gathered in this module is: the term accompanied by part of speech (*pos*), number and gender, context and its sources, lemmatised form and administrative data.

- **Corpus module**: text corpus of the genomic domain selected and validated by experts, and processed using NLP applications.

- **Entities module**: references of the information sources used in the Term and Corpus bases.

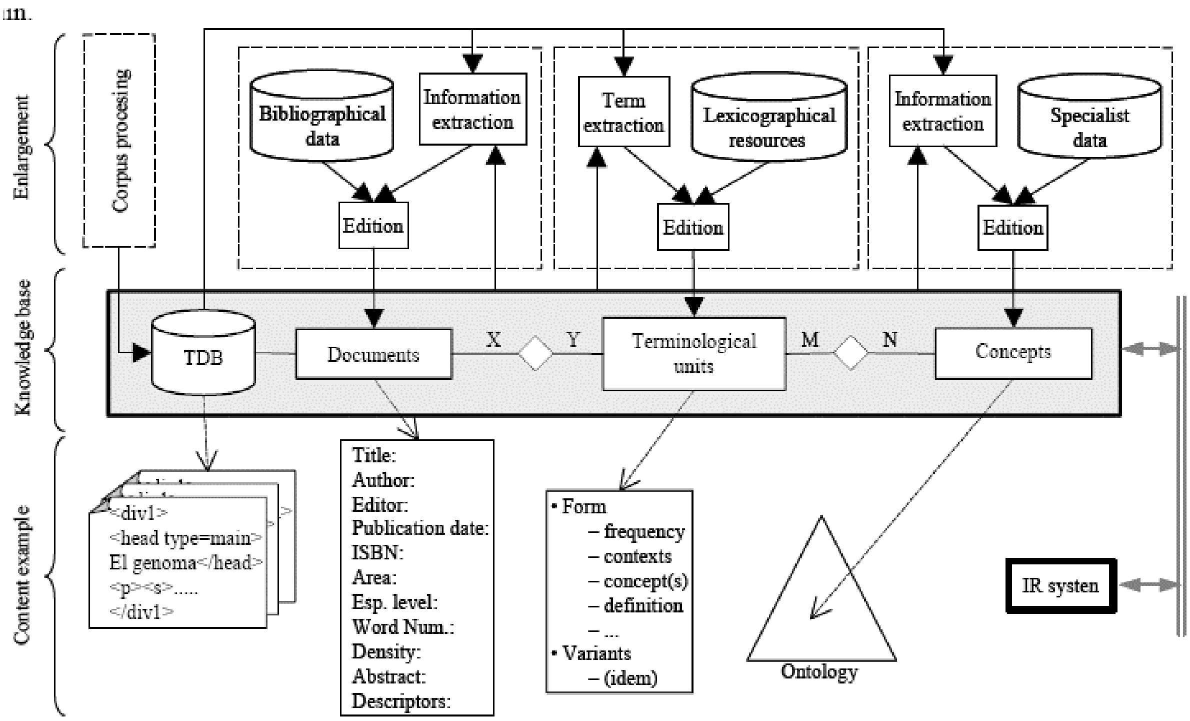


Figure 1: Knowledge Base architecture (Feliu et al. 2002)

For the development of each module, the following steps were followed:

1) Use of Mikrokosmos to develop the **Ontological module** based on ontology concepts and their relations.
2) Representation of the ontology labels in English
3) Compilation of the **Corpus module** with genomic domain documents selected and validated by experts.
4) Development of the **Term base module** with knowledge units extracted **semi-automatically** from the specialized corpora and other on line lexical resources (LRs). Terms are then mapped to the ontology. Inclusion of non-mandatory definitions and three contexts for each term.
5) Addition of full bibliographical data in the **Entities module**

4. Main implications of the development

According to the steps followed for the creation of the aforementioned resources, we observe the following aspects, which may have an influence on the suitability of the resource for the translation of specialized texts:

- In the Eurodicautom case, the development starts with the merging of LRs as glossaries, thesauri, etc. However, GENOMA-KB starting point consists of a core of main concepts of the field proposed by experts and structured according to relations among them.
- Multilingual information for enlarging the Eurodicautom dictionary is then extracted by terminologists and translators from EU documents, and supervised

by experts. In a similar way, experts in the genome field compile specialized texts, and concepts are then retrieved by terminologists, translators, linguists, or the experts themselves.

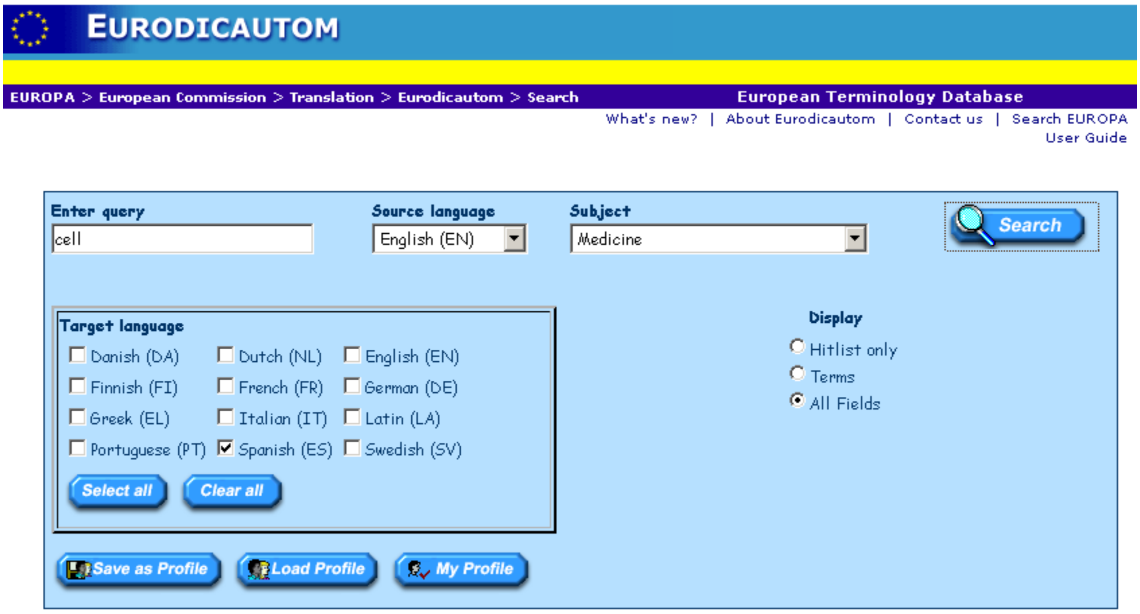
- In the last years, NLP and translation supporting tools have been introduced in the EU. Whereas, for GENOMA-KB, the retrieval task has always counted on such resources.

Therefore, if in both cases experts, terminologists and translators work together in order to extract concepts from specialized documents with the support of NLP and translation tools, the main difference between them remains the organization and presentation of the information, as we analyze in the next sections.

5. Multilingual information retrieval in Eurodicautom and GENOMA-KB

This section provides a description of the process of retrieval of multilingual information in both resources, and the contributions of the obtained information for the purposes of a specialized translator. We will take as example for the search the term *cell*, which is common to both resources.

In the initial search for *cell* in Eurodicautom, the user can select the source language (*English*), the subject (*Medicine*), the target language or languages (*Spanish*), and the amount of information to be displayed (*All fields*), as can be seen in Figure 2.


EURODICAUTOM
EUROPA > European Commission > Translation > Eurodicautom > Search
European Terminology Database
What's new? | About Eurodicautom | Contact us | Search EUROPA
User Guide
Search
Select all
Clear all
Save as Profile
Load Profile
My Profile

Enter query: cell

Source language: English (EN)

Subject: Medicine

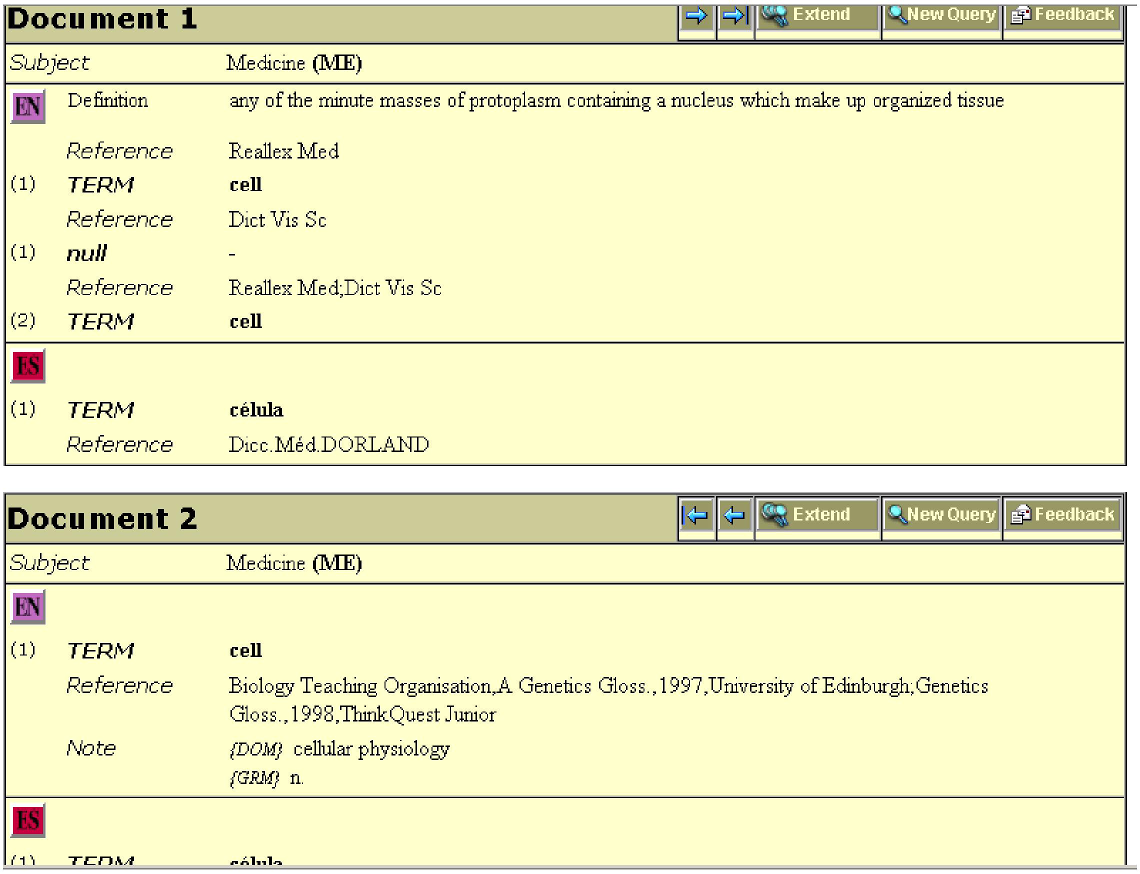
Target language

□ Danish (DA)	□ Dutch (NL)	□ English (EN)
□ Finnish (FI)	□ French (FR)	□ German (DE)
□ Greek (EL)	□ Italian (IT)	□ Latin (LA)
□ Portuguese (PT)	☑ Spanish (ES)	□ Swedish (SV)

Display
○ Hitlist only
○ Terms
● All Fields

Figure 2: Eurodicautom interface

The information obtained is displayed as shown in Figure 3. Each **Document** section represents a different entry and it is supposed to be related to a different sense of the term. The **Subject** sub-section (*Medicine*, in this case) refers to the domain in which the term is used, and the **Reference** (*Reallex Med.*) to the source where the term has been obtained, which can give us a hint about the reliability of the results. The **Definition** field is not compulsory, and is eventually to be found for the source language as well as for the selected target languages.



Document 1

Subject Medicine (ME)

EN Definition any of the minute masses of protoplasm containing a nucleus which make up organized tissue

Reference Reallex Med

(1) TERM **cell**

Reference Dict Vis Sc

(1) null -

Reference Reallex Med;Dict Vis Sc

(2) TERM **cell**

ES

(1) TERM **célula**

Reference Dicc.Méd.DORLAND

Document 2

Subject Medicine (ME)

EN

(1) TERM **cell**

Reference Biology Teaching Organisation,A Genetics Gloss.,1997,University of Edinburgh;Genetics Gloss.,1998,ThinkQuest Junior

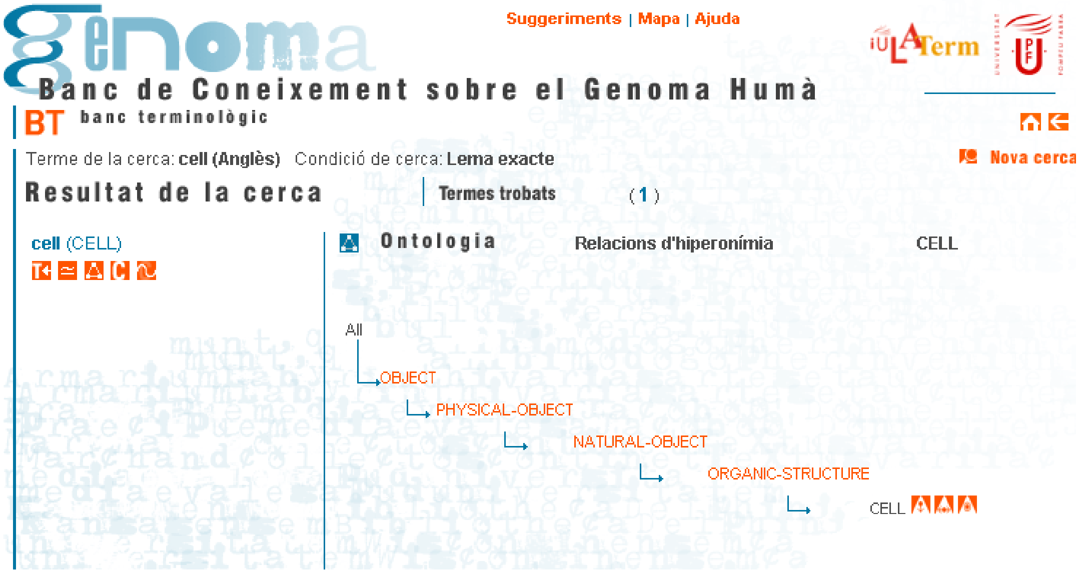
Note *{DOM}* cellular physiology
{GRM} n.

ES

(1) TERM **célula**

Figure 3: Results for *cell* in Eurodicautom

The web page of the GENOMA-KB offers the user multiple search possibilities. One can choose to consult amongst the Ontology/Term base module, the Corpus module, the Bibliographic module or the Factographic module. For the purposes of our paper, the Ontology/Term base module and the Corpus module will offer the most interesting results. The following results are displayed for *cell* depending on the type of relations of this concept to its “neighbour concepts”. Not all relations are displayed at once, but one has to look for each kind of relations at a time. In the following figures (4 to 7), we show how hyperonymy, hyponymy, co-hyponymy and *ad hoc* relations for the term *cell* are presented. Next to the searched term there are always links to the other types of relations, which can easily be consulted.



genoma
Banc de Coneixement sobre el Genoma Humà
BT banc terminològic

Suggeriments | Mapa | Ajuda

IULATerm

Nova cerca

Terme de la cerca: cell (Anglès)   Condició de cerca: Lema exacte

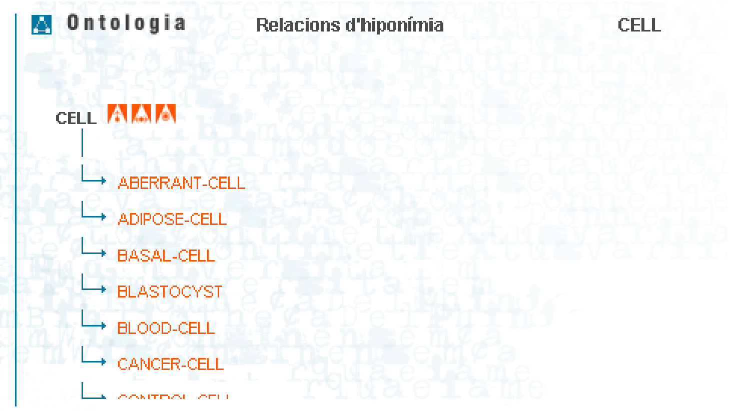
Resultat de la cerca | Termes trobats (1)

cell (CELL)

Ontologia   Relacions d'hiperonímia   CELL

All
└→ OBJECT
  └→ PHYSICAL-OBJECT
    └→ NATURAL-OBJECT
      └→ ORGANIC-STRUCTURE
        └→ CELL

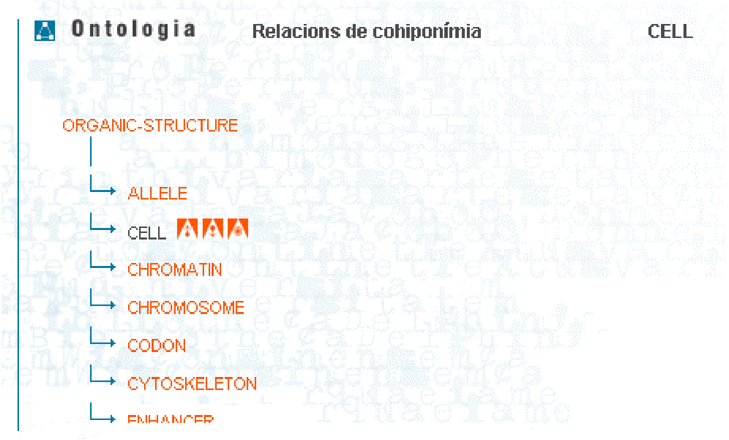
Figure 4: Hyperonymy relations for *cell* in GENOMA-KB



Ontologia   Relacions d'hiponímia   CELL

CELL
└→ ABERRANT-CELL
└→ ADIPOSE-CELL
└→ BASAL-CELL
└→ BLASTOCYST
└→ BLOOD-CELL
└→ CANCER-CELL
└→ CONTROL-CELL

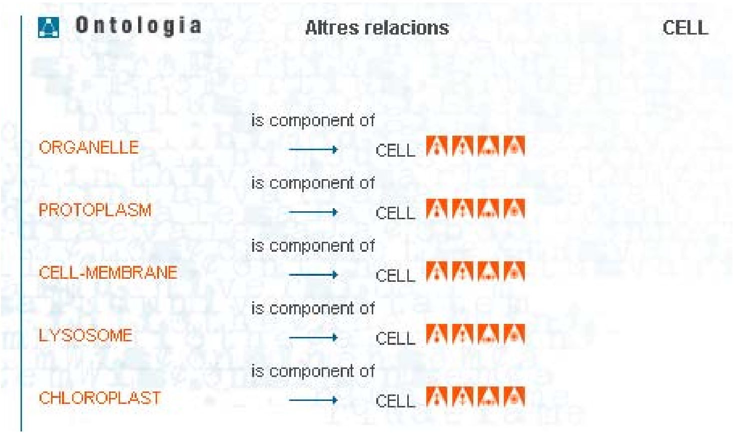
Figure 5: Hyponymy relations for *cell*



Ontologia   Relacions de cohiponímia   CELL

ORGANIC-STRUCTURE
└→ ALLELE
└→ CELL
└→ CHROMATIN
└→ CHROMOSOME
└→ CODON
└→ CYTOSKELETON
└→ ENHANCER

Figure 6: Co-hyponymy relations for *cell*

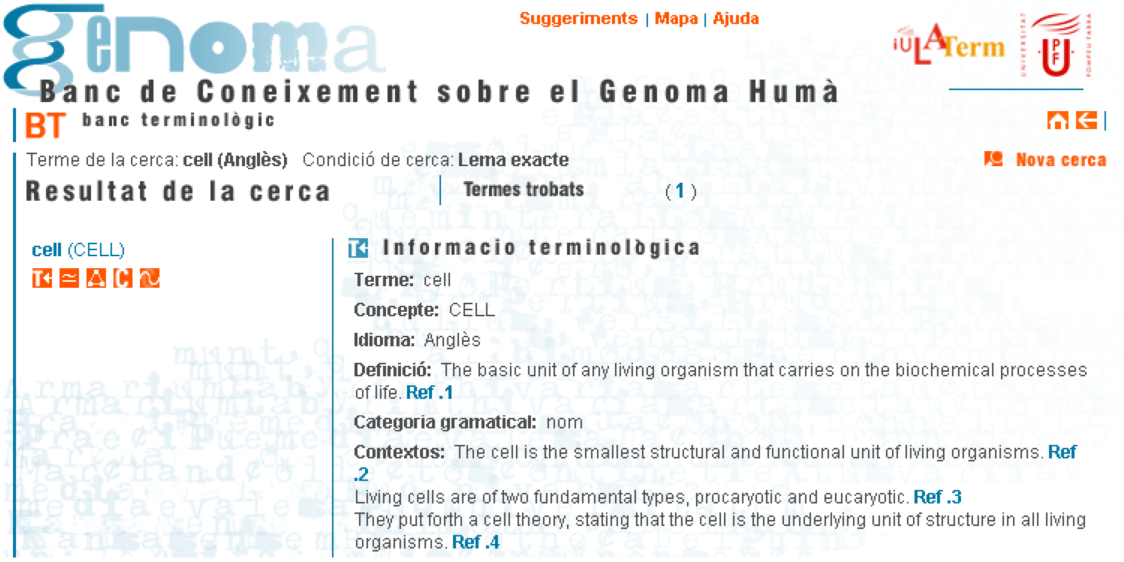


Ontologia Altres relacions CELL

ORGANELLE	is component of →	CELL
PROTOPLASM	is component of →	CELL
CELL-MEMBRANE	is component of →	CELL
LYSOSOME	is component of →	CELL
CHLOROPLAST	is component of →	CELL

Figure 7: *Ad hoc* relations for *cell*

In Figure 8 we can get information about pos, gender, definition and meaningful use contexts for the concept *cell*. Next to definition and context there are links to the extraction sources (*Ref.*), which are always an important source of knowledge.



genoma
Suggeriments | Mapa | Ajuda
iULATerm
UPF
Banc de Coneixement sobre el Genoma Humà
BT banc terminològic
Nova cerca

Terme de la cerca: cell (Anglès) Condició de cerca: Lema exacte

Resultat de la cerca | Termes trobats (1)

cell (CELL)

Informacio terminològica

Terme: cell

Concepte: CELL

Idioma: Anglès

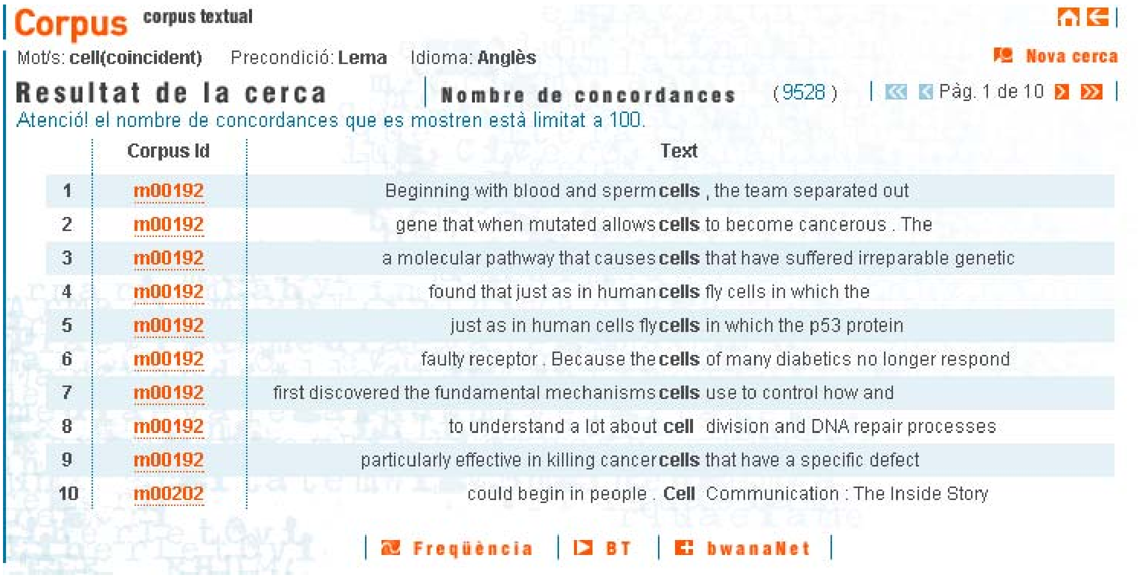
Definició: The basic unit of any living organism that carries on the biochemical processes of life. Ref .1

Categoria gramatical: nom

Contextos: The cell is the smallest structural and functional unit of living organisms. Ref .2
Living cells are of two fundamental types, procaryotic and eucaryotic. Ref .3
They put forth a cell theory, stating that the cell is the underlying unit of structure in all living organisms. Ref .4

Figure 8: Terminological information for *cell*

Finally, we could as well access the Corpus module and visualize the concordances of the term *cell* in the corpus (cf. Figure 9).



Mot/s: **cell(coincident)** Precondició: **Lema** Idioma: **Anglès**

Resultat de la cerca | **Nombre de concordances** (9528) | Pàg. 1 de 10

Atenció! el nombre de concordances que es mostren està limitat a 100.

	Corpus Id	Text
1	m00192	Beginning with blood and sperm **cells** , the team separated out
2	m00192	gene that when mutated allows **cells** to become cancerous . The
3	m00192	a molecular pathway that causes **cells** that have suffered irreparable genetic
4	m00192	found that just as in human **cells** fly cells in which the
5	m00192	just as in human cells fly **cells** in which the p53 protein
6	m00192	faulty receptor . Because the **cells** of many diabetics no longer respond
7	m00192	first discovered the fundamental mechanisms **cells** use to control how and
8	m00192	to understand a lot about **cell** division and DNA repair processes
9	m00192	particularly effective in killing cancer **cells** that have a specific defect
10	m00202	could begin in people . **Cell** Communication : The Inside Story

Figure 9: Concordances of *cell* in the Corpus module

6. Main implications of quantity and presentation of multilingual information

After this brief comparison of the different stages in the development of both resources and the actors that intervene, as well as the multilingual information we can retrieve in a search, we are able to conclude that the form in which multilingual information is organized and presented in the ontology offers the following benefits for the translation of specialized texts against to the information offered by electronic dictionaries:

1) *Domain knowledge* acquisition thanks to:

   a. conceptual relations between concepts in the field, which capture the semantics of the domain, as for example relations of hyperonymy or meronymy
   b. encyclopaedic information gathered in definition, meaningful context and corpus sections

If the translator knows that *cell-membrane*, *lysosome* and *chloroplast* are components of a *cell* (cf. Figure 7), he or she will be much closer to identifying those terms accurately and finding the equivalent terms in the target language than if the translator just knows that those terms belong to the *Medicine* field. Sometimes, the real problem for translators is “the inability to accurately describe or delimit the foreign concept in the source language in the first place” (Bonnono, 2000:660), which can be solved with the domain information ontologies offer. Dictionaries, however, rely on definitions and subject sections to offer information of the domain. Even considering specialized dictionaries, it is complicated for a dictionary definition alone to offer a complete picture of a specific parcel of the domain knowledge in question.

2) *Linguistic and terminological knowledge* acquisition derived from:

a. lexical relations between terms that enable translators to identify which concepts are at the same level (polysemy) and can be used in the same context
b. concordances in the corpus module

In order to enrich his or her linguistic and terminological knowledge on the field, a translator using a dictionary will have to resort to specific documents, corpora or other linguistic resources, to check concordances, synonymy uses, etc, whereas the ontology is able to offer all these data in a visual and practical way.

3) *Multilingual information across cultures*, since the linguistic information that accompanies ontology concepts is in two or more languages. The translator can easily check the correspondence between terms in different languages or find out those nuances or variations between the different conceptualizations. In the case of dictionaries, we mostly find a quantity imbalance of multilingual information.

4) *Reliable knowledge and multilingual information* in the domain agreed by authoritative experts from different languages and cultures. This is one of the main characteristics of ontologies, and the one which differentiates it form the most LR, then it is compulsory to reach consensus on conceptualizations during the development of the resource by the experts in the field.

Besides that, references to definitions and context not only provide critical information about a term, but also enable the translator to validate that information in the data corpus.

As a consequence of the reasons previously listed, we can state that the quantity of multilingual domain knowledge information and the way it is organized and presented in ontologies could be of a great contribution to the translation of specialized texts. Firstly, ontologies offer domain knowledge at a glance, enabling translators to save time in searching in additional resources. And, secondly and more important, ontologies can conceptualize very specific parcels of knowledge by obtaining the previous agreement from authoritative people on the field.

7. Acknowledgements

The research described in this paper is supported by the European Commission's Sixth Framework Programme (IST-2005-027595) under the project name: *Lifecycle support for networked ontologies* (*NeOn*). We would also like to thank Asunción Gómez-Pérez and Mari Carmen Suárez-Figueroa for their help with some ontological aspects.

8. Bibliography

Arano, S. (2005). "Thesauruses and ontologies" [on-line], in *Hipertext.net, num. 3, 2005*. http://www.hipertext.net [Consulted: 15th September, 2006]

Bononno, R. (2000). "Terminology for Translators – Implementation of ISO 12620". Meta, XLV, 4: 646-669.

Cabré, M. T, C. Bach, R. Estopà, J Feliu, G. Martínez, & J. Vivaldi (2004). "The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities", in *LREC 2004*, Lisboa: European Languages Resources Association, 87-90.

Directorate-General for Translation of the European Commission, (2005). *Translation Tools and Workflow* [on-line], in http://ec.europa.eu/dgs/translation/index_en.htm [Consulted: 13th September, 2006]

Feliu, J., J. Vivaldi & M.T. Cabré (2002). "Towards an Ontology for a Human Genome Knowledge Base", in *LREC2002*, Las Palmas de Gran Canaria, 1885-1890. ISBN: 295-1740-808.

Goffin, R. (1997). "EURODICAUTOM. La banque de données terminologiques muiltilingues de la Commission européenne (1973-1997)" in *Terminologie et Traduccion 2*, 30-73.

Gruber, T.R. (1993). *A translation approach to portable ontology specification.* Knowledge Acquisition 5(2):199-220.

Hernúñez, P. (2000). "Las bases de datos terminológicos de la Comisión Europea. EURODICAUTOM", in Gonzalo García, C. & V. García Yebra (eds.), 2000: *Documentación, Terminología y Traducción,* Madrid: Síntesis, Fundación Duques de Soria: 97-107.

Moreno Ortiz A. & Pérez Hernández (2000). "Reusing the Mikrokosmos Ontology for Concept-Based Multilingual Terminology Databases" in *LREC-2000*, Athens, 1061-1067.

Studer, R., V.R. Benjamins, & D. Fensel (1998). *Knowledge engineering: principles and methods*, in IEEE Transactions on Data and Knowledge Engineering 25(1-2): 161-197.

Wilss, W. (1996). *Knowledge and Skills in Translator Behaviour*. Amsterdam: Benjamins.

URLs:

Euramis: http://ec.europa.eu/translation/reading/articles/pdf/1998_01_tt_blatt2.pdf#search=%22euramis%22

Eurodicautom: http://ec.europa.eu/eurodicautom/Controller

GENOMA-KB:http://genoma.iula.upf.edu:8080/genoma/index.jsp

Institut Universitari de Lingüística Aplicada, UPF: http://www.iula.upf.edu/

Merriam-Webster Online Dictionary: http://www.m-w.com/dictionary/

Mikrokosmos Ontology: http://crl.nmsu.edu/Research/Projects/mikro/htmls/ontology-htmls/onto.index.html

OntoTerm: http://www.ontoterm.com/

Systran: http://www.systransoft.com/index.html

Trados Multi Term: http://www.trados.com/