Ontologies: A Review

J. Feliu, J. Vivaldi, M. T. Cabré

Sèrie Informes, 34

Barcelona
Universitat Pompeu Fabra. Institut Universitari de Lingüística Aplicada
2002

Direcció de les Publicacions de l'IULA: M. Teresa Cabré

Coordinació de les Publicacions de l'IULA: Elisabet Solé, Gemma Martínez

Primera edició: maig de 2002

La Rambla, 30-32
08002 Barcelona

Dipòsit legal: B-34.232-2002

Ontologies: A Review[♣]

Judit Feliu
judit.feliu@iula.upf.es

Jorge Vivaldi
jorge.vivaldi@info.upf es

M.Teresa Cabré
teresa.cabre@trad.upf.es

Institut Universitari de Lingüística Aplicada
Universitat Pompeu Fabra
Barcelona

En aquest paper, analitzem les principals ontologies amb la finalitat de dibuixar un panorama general d'una de les eines més utilitzades en l'estructuració del coneixement. En primer lloc, presentem una àmplia descripció de les cinc ontologies més difoses entre la comunitat científica dedicada a la gestió de la informació. Seguidament, repassem breument algunes de les eines de gestió que s'utilitzen per crear i actualitzar ontologies. I, finalment, presentem algunes conclusions en relació a la selecció d'una ontologia i d'un sistema de gestió per a la seva utilització en el marc dels projectes vigents del grup IULATERM.

In this paper, we analyse five main ontologies used in the knowledge organisation field. First, we present a detailed overview of these ontologies which have been widely used among the scientific community working in the information management domain. Second, we briefly summarize the main characteristics of some management tools used for creating and enlarging ontologies. Lastly, some conclusions are drawn concerning the selection of an ontology and its corresponding management tool in order to be reused in the IIULATERM's ongoing projects.

[♣] This working paper has been developed under the framework of the two following public funded projects: TEXTERM: *Textos especializados y terminología: selección y recuperación automática de la información* (BFF2000-0841) and RICOTERM: *Sistema de recuperación de información con control terminológico y discursivo* (TIC2000-1191).

Table of Contents

Introduction[*]

The aim of this working paper is to evaluate some of the existing ontologies in order to select the most appropriate one, as a starting point, for an Institute's ongoing project, which is briefly described in the following section.

The need for using an ontology in the project framework leads us to make a review of most known ontologies.

After an outline of the project main objectives (section 1), the report will review some general considerations on ontologies (section 2). Then, five well-known ontologies will be described and analysed (section 3). We also briefly comment on some other ontologies which have been the basis of the existing ones or which are already used in some documented projects. A few brief remarks on a cooperative ontology building system have been added.

Following this ontologies review, we present a comparative analysis (section 4) between the five main ontologies evaluated, according to the following parameters: availability, management facilities (enlargement and modification), expressiveness, application field, ontology type, and size, granularity and completeness.[1]

Very recently, many efforts have been devoted to the application of ontologies to the organisation of the web. Most worth mentioning advances obtained in this field are described in section 5. The paper ends with some final remarks about ontology design and the management tools available in order to determine whether they can be integrated in our project (section 6).

1 Genome Project

It is widely known that new information technologies provide different tools used for the management and the transfer of large amounts of documents. Considering the increase in data, one of the main goals of information management is to access and retrieve appropriate documents. Present information retrieval (IR) systems show a limited effectiveness because they mostly use statistical information and only in some cases ground level linguistic data. They do not usually have access to any form of semantic information.

[*] We would like to thank Marie-Claude L’Homme (Université de Montréal) for the revision and kind comments on this working paper.

[1] Ontologies can be further classified according to other criteria, such as fields covered, types and number of concepts and relations implemented, etc. We have not considered all these parameters in this first attempt to evaluate ontologies. In addition, this information is not always available.

The Institute's ongoing project, the so-called Genome Project, is carried on within the framework of two public funded projects: 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 characterisation 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, semi-automatic mapping of cognitive nodes and conceptual relations, and the algorithm and protocol designs. It is foreseen that our working methodology —oriented to improve information retrieval systems— would combine strategies from the cognitive sciences and from 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 (NLP) and statistical analysis.

Traditionally, most information retrieval systems have been based on strategies of formal strings 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, 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 combine a tool for natural language processing, which includes structural mark-up, morphological and syntactic analysis, disambiguation, and a terminology extraction system based on formal patterns and lexical ontologies. 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 Genome Knowledge Base, whose architecture is shown in Figure 1.

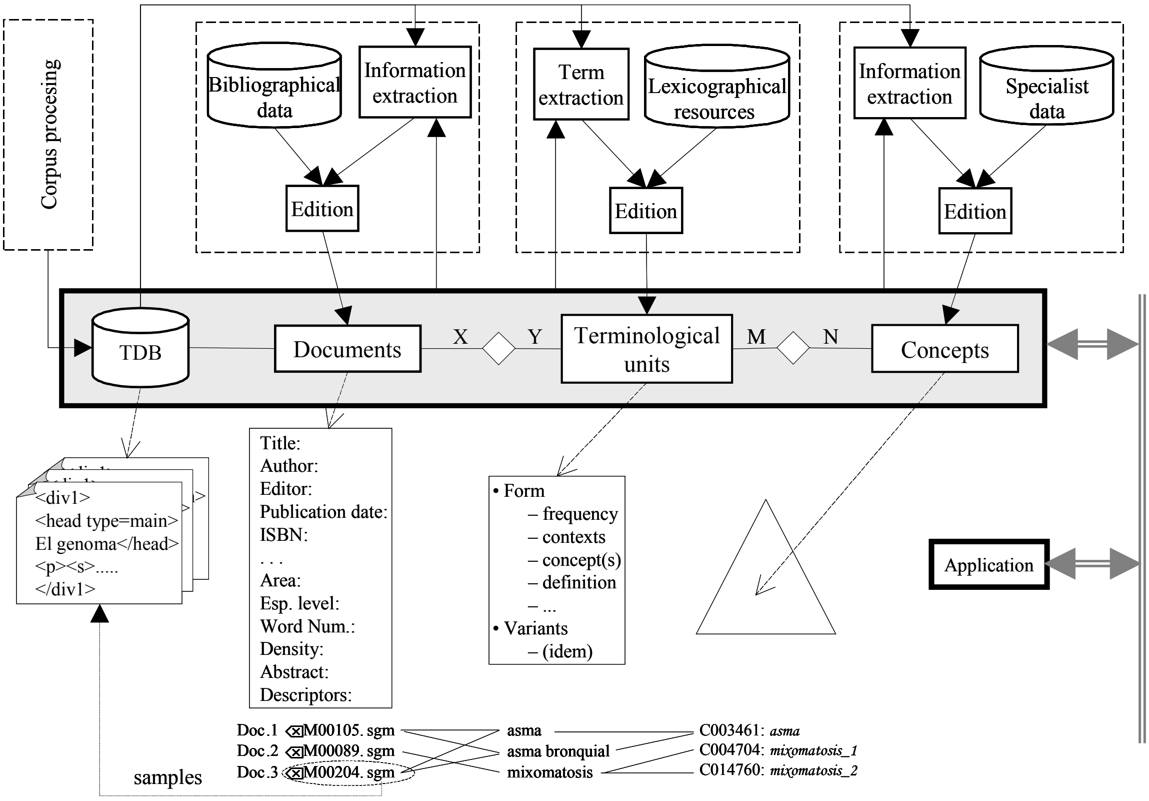


Figure 1. Genome project: overview

Figure 1 shows the tight relation between the terminological database, the specialised knowledge units, the concepts and the documents related to the Human Genome domain which will constitute the core of the project. The ontology, directly related to concepts, will be used in order to classify and structure specialised knowledge drawn from the corpus. In order to ease such task, the documents included in the corpus are previously morphologically and syntactically tagged.

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 summarisation, machine translation support, etc.

2 Ontologies: General Considerations

> *"An ontology is an explicit specification of a conceptualization. The term is borrowed from philosophy, where an ontology is a systematic account of Existence. For knowledge-based systems, what "exists" is exactly that which can be represented. When the knowledge of a domain is represented in a declarative formalism, the set of objects that can be represented is called the universe of discourse. This set of objects, and the describable relationships*

> *among them, are reflected in the representational vocabulary with which a knowledge-based program represents knowledge. Thus, we can describe the ontology of a program by defining a set of representational terms. In such an ontology, definitions associate the names of entities in the universe of discourse (e.g., classes, relations, functions, or other objects) with human-readable text describing what the names are meant to denote, and formal axioms that constrain the interpretation and well-formed use of these terms."*
>
> (T. Gruber, 1993)

Following T. Gruber, an ontology is an explicit specification of a conceptualisation. The author talks about "what exists" which, in our approach, will become what we find in specialised texts; that is, concepts represented by specialized knowledge units (or terms), conceptual relations and other kinds of linguistic information useful for terminological purposes. This knowledge material should occupy a place in the ontology. The set of specialised texts saved in our corpus will configure our *universe of discourse*.

The word ontology has been borrowed from philosophy and it has been extrapolated to Artificial Intelligence (AI). AI explores the use of formal ontologies as a way of determining content-specific agreements for the sharing and reuse of knowledge among software entities. Formal ontologies are viewed as designed artifacts, formulated for specific purposes and evaluated against objective design criteria. From AI, the role of ontologies is to give support to knowledge sharing activities and for this reason, researchers in this field try to establish a set of criteria to guide the development of ontologies oriented to knowledge representation, sharing and retrieval.

In a knowledge representation system, the term 'ontology' refers to all the concepts in a particular domain. Moreover, this term is frequently used to describe the concepts, the relationships and the restrictions taken into account in order to obtain the modelisation of a domain[2]. In this case, an ontology can be seen as a formal representation of a particular specialised field containing its own terminology.

Terminology is defined as a set of terms in a 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.

Specialised knowledge mapping is a hard task and most efforts must be directed to design the ontology. Criteria for its design becomes 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.

[2] For more information about the relationship between knowledge representation and terminology, see the following url: ***http://www.biomath.jussieu.fr/~pz/Publications/ZweigenbaumISIS99/isis99.html***

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 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, WordNet [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 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, such information is shared by the corresponding node and all its hyponyms. 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 a 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 based mainly in a deep control of the hierarchy have been proposed. For example, the "orthogonal inheritance" suggests to gather the data and allow 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, 1998]) or domain specific (i.e., ULMS [NLM, 1998]),
- generic (i.e., WN) or built for a particular application,
- episodic or encyclopedic ontologies (i.e., Cyc [Lenat *et al.*, 1990]),
- lexical (i.e., WN) or conceptual (i.e., Cyc) ontologies.

Finally, there also exist the so-called metaontologies; that is, the particular formalisms oriented to ontology building and reusability (i.e., *Ontolingua*).

Before analyzing the most commonly used ontologies, it is important to bear in mind that higher or upper level ontologies intend to represent all sort of things existing in the world and relations among them (LE3-4244, 1999). Although NLP applications in different domains seem to require domain-specific conceptualisations, there is some hope that a common upper level of domain-independent concepts and relations may be agreed. In this way, the workload necessary for each individual application should be in some way reduced. This kind of representations may be useful for a variety of natural language (NL) understanding and generation tasks such as syntactic disambiguation, co-reference resolution, inference based on world knowledge and language independent meaning representations for text generation and machine translation. This kind of ontologies are not lexical resources as they use a specific conceptualisation language. For this reason, it is necessary to provide some mapping between the ontology and the lexicon (see for example the lexical modules of Cyc, μKosmos and UMLS.

3 Ontologies Analysis

As we have already mentioned, we will analyse the most commonly used ontologies in order to determine the main characteristics concerning their criteria design and general structure. Our analysis covers the following five well known ontologies:

- Cyc
- EuroWordNet
- μKosmos
- SIMPLE
- UMLS

It is interesting to notice that except from UMLS, which is domain-specific, the four other ontologies are not restricted to a specialised domain, even though μKosmos is further developed for the economic field (to support a knowledge based machine translation [KBMT] system).

For each ontology we include a description of the resource, a sample and a final analysis[3]. For the sample we have chosen the concept "cell", whose representation has been made available to us.

As it has been mentioned in the introduction, our aim is to integrate the selected ontology in our project. For this reason, we would like to examine the management tool available for each ontology dealt in this paper. Unfortunately, most ontologies analysed do not offer a description for its corresponding management tool. It is only in the case of μKosmos that this kind of information is available and described correspondingly in a subsection.

[3] It should be noted that we do not have the same level of information for all the analyzed ontologies: EWN and UMLS are publicly available (and browsable), μKosmos has been examined through OntoTerm but also some description is available in Moreno (2000), Cyc has released a sample obtainable in ***http:www.cyc.com*** and SIMPLE has been described in Bel *et al.* (2000).

In order to give a thorough overview, in section 3.6 we also introduce a brief outline of some particular ontologies: $(KA)^2$, EDR, EngMath, Enterprise Ontology, Generalized Upper Model, Pennman Upper Model and PhysSys. All of them were developed for different purposes and they are not equally available. For this reason, our description is brief and limited to their designers, their project or enterprise orientation and their application domain. Finally, we have dedicated a section to introduce Ontolingua, an ontology building system, which offers some already designed and tested ontologies that can be reused in order to build a new one.

3.1 Cyc

This is a high level ontology developed by Cycorp, a private company participated by the major U.S. software houses. Its development started in the early 80s. Over the past years the Cyc team has added a huge amount of fundamental human knowledge to the knowledge base: facts, rules of thumb, and heuristics for reasoning about the objects and events of modern everyday life.

Unfortunately, there is only generic knowledge about this resource and scarce amount of information about the details of the architecture of this system. It is not until very recently that Cycorp has released a set of approximately 3000 terms capturing the most general concepts of human consensus reality. It is known as the “upper Cyc ontology” and, according to Cycorp, it may be the core of any ontology because it satisfies two important criteria: it is universal (any imaginable concept can be linked to this ontology) and articulated (the distinctions made in the ontology are both necessary and sufficient for most purposes). Anyway, it is a partial release because it includes just a small part of the full knowledge base. It does not include neither the (English) lexicon and the mappings to the knowledge base nor any components of their NLP system (parser and semantic interpreter).

Cyc is universally considered as the prototype of a high level ontology. As such, it uses CycL, a specific knowledge representation language, to represent concepts. It is a formal language whose syntax derives from first-order predicate calculus (the language of formal logic) with equality, augmentations for default reasoning, skolemisation, and some second-order features (e.g., quantification over predicates is allowed in some circumstances). The vocabulary of CycL consists of terms that may be: semantic constants, non-atomic terms, variables, numbers, strings, etc. Terms are combined into meaningful CycL expressions, ultimately forming meaningful closed CycL sentences. Each term has at least an “*is-a*” link to the superclass of which it is an instance and a ‘*genls*’ links to superclasses of which it is a subclass. Two of the most important Cyc classes are collections and relations (predicates and functions).

Each concept in the knowledge base is represented as a constant. A constant can represent a collection, an individual object, a word in a natural language, a quantifier (such as 'there exists'), a relation (a predicate, function, slot, attribute, etc.), and so on.

The knowledge base is organised as a collection of lattices where the nodes are the Cyc terms and the edges are different kinds of relations.

Cyc claims to have its own NL system but the only concrete data about such system is the English lexicon that currently contains about 14,000 entries (stems and idioms), containing “the usual sorts of linguistic information”. Each entry contains also a link to the appropriate Cyc concept.

3.1.1 Sample

In this section we show the information made available by Cycorp for the “cell” concept, which does not seem to be very complete.

#$Cell

The collection of living cells; a subset of #$BiologicalLivingObject. Each element of #$Cell is one of the basic structural units of nearly all living things, consisting (at least) of cytoplasm bounded by a cell membrane. Only the living structures viruses, mitochondria, and plastids are not composed of cells.

isa: #$ExistingObjectType

genls: #$BiologicalLivingObject

some subsets: #$SingleCellOrganism #$EukaryoticCell #$ProkaryoticCell, #$Protozoan (plus 2 more public subsets, 22 unpublished subsets)

Figure 2. Representation of the “cell” concept in Cyc.

3.1.2 Analysis

The main characteristic of Cyc is its attempt to collect all kinds of “common sense” knowledge and to provide a “deep” layer of understanding that can be used by other programs to make them more flexible. Unfortunately, this is not a public resource and the part that has been made available is very limited. This public release does not include the lexicon nor the NL system, together with a reduced subset of the full knowledge base.

The only existing lexicon has been developed for English and no mention is made on the possibilities to extend it to other languages.

3.2 EuroWordNet (EWN)

EWN (Vossen, 1999) is a general-purpose multilingual lexical database based on Princeton WordNet (Fellbaum, 1998) covering Spanish and other European languages[4]. Each language has its own wordnet structure[5], all wordnets are linked by

[4] EWN was a project funded by the European Union that initially covered Spanish, Dutch, English and Italian. Later, the project was extended to French, German, Estonian and Czech. Extensions, locally supported, are Catalan, Basque, Greek, among others.

[5] The internal organisation of each language is mainly based in WordNet version 1.5.

means of some common structure. In spite of being a closed project, WN is very active in the NLP area[6].

A wordnet is structured in lexical-semantic units, or synsets, that are linked according to basic semantic relations. A synset is a set of synonymous words –in the sense of Princeton WordNet[7]– that can be interchanged in certain contexts. Synonymy and hyponymy are the basic relations for both WN and EWN. The former is used to create the synsets while the latter defines the very basic relation between synsets.

A typical synset example (taken from WN 1.5) is the following set of words {car, auto, automobile, motorcar}. Any of them can be used in a given sentence without changing its basic meaning. This synset is related to some other synsets as illustrated in Figure 3.

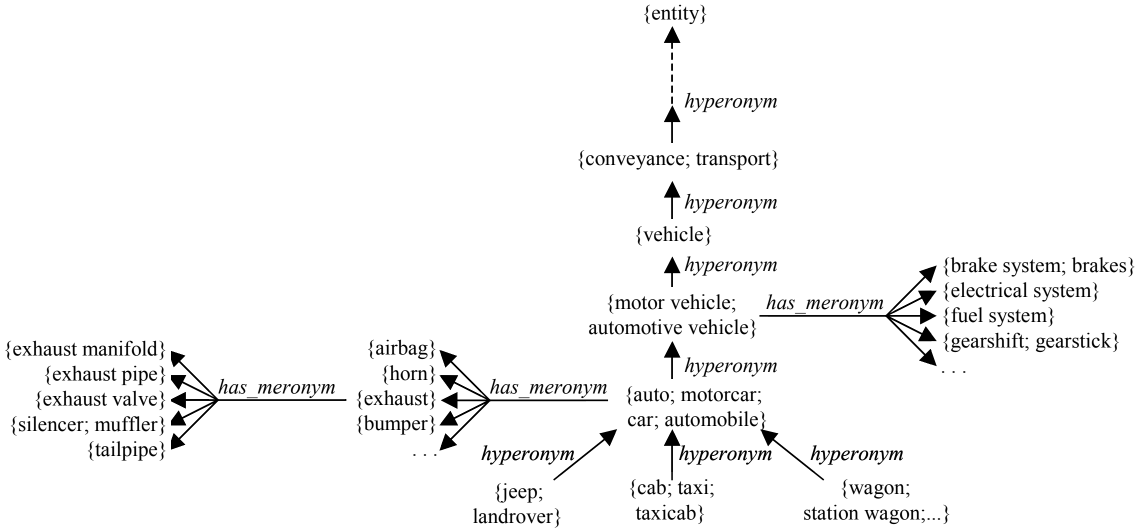


Figure 3. Synsets related to the first sense of the word 'car' (Princeton WN v. 1.5).

Both EWN and WN divide the full set of nouns into several hierarchies, each of them having its own beginner or "top". Each hierarchy corresponds to a relatively distinct and autonomous semantic field as for example: {act, activity}, {artifact} and {cognition, knowledge}. Inside the hierarchies, each synset is linked to its hyperonym forming a sort of chain. A lexical hierarchy may be built following the hyperonymical relationships as for example: {bronchus} → {cartilaginous tube} → {tube} → {body structure} → {body part} → {part} → {entity}. The symbol "→" may be read as a relation "*is-a*" or "*is-a-kind-of*" allowing to go from specific to generic items. In this

[6] Several international workshops are held discussing issues exclusively related to this project and its applications. See, for example, the workshop held in Canada in 1998 (in parallel with the ACL), or the NAACL 2001 Workshop (WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Carnegie Mellon University, 3 and 4 June, 2001). The web site *http://www.cis.upenn.edu/~josephr/wn-biblio.html* collects a large number of papers presented in several workshops and congresses and the web site *http://www.cogsci.princeton.edu/~wn/links/* identifies a number of projects closely related to WordNet.

[7] "two expressions are synonymous in a linguistic context C if the substitution of one for the other in C does not alter the truth value" (Miller G.A. *et al.*, 1993).

way, a given synset inherits all the information given to its hyperonyms. From the examples above, the inheritance mechanism allows to say that a 'bronchus' is a 'body part' and that a 'station wagon' is a kind of 'car'. Both concepts have an 'exhaust valve' and an 'electrical system' (among a lot of other things that are inherited from the hyperonyms of 'car').

The development of the various areas of knowledge is not homogeneous, causing the size and granularity of the hierarchies to be very different. Also, this classification produces a difference between both WN and EWN projects, in the first one there are 25 unique beginners that are almost doubled in some versions of the second[8].

The nouns form the basic and more developed hierarchy for both WN and EWN but they also include some organisation for other parts of speech (POS): adjectives, verbs and adverbs. Both projects define a series of relations like meronymy, holonymy, etc. However, while WN keeps a strict division between different POS, EWN defines a largest set of relations between synsets belonging to the same or different POS as, for example[9], those showed in Table 3-1.

Relation	POS	Example

XPOS_NEAR_SYNONYM	noun – verb
	verb – adjective
		{injection} – {to inject}
		{to alive} – {alive}

XPOS_NEAR_HYPERONYM	noun – adjective
	noun – verbs
		{age} – {old}
		{election} – {to vote}

CAUSES	noun – noun	{microorganism} – {health problem}
PERTAINS_TO	adjective – noun	{pulmonary, pneumonic} – {lung}

Table 3-1. Some new relations of EWN.

All the above description applies, with some differences, to both EWN and WN projects but the former presents some additional characteristics that will be briefly described below.

One of the most important differences between EWN and WN is multilinguality or the fact that the former takes multilinguality into account. It is achieved by adding an equivalence relation between synsets of every language. Such equivalence is obtained by means of the interlingual index (ILI) that may be defined as an unstructured list of meanings whose only purpose is to link synsets from all the language-specific wordnets. Every synset in each language has at least one direct or indirect link with a record in the ILI list. This structure allows each language to be organised autonomously but keeping the possibility to indirectly see any language specific organisation.

[8] Both Spanish and Catalan EWN fully reflect the organization of WN keeping the same number of tops.

[9] See (Vossen P., 1999) for the full list (and their description) of the available relations between synsets belonging to the same or different POS.

Some structuring is indirectly provided through two language independent ontologies which may be linked to ILI records:

- *the Top Concept Ontology*: it is a hierarchy of language-independent concepts that reflects important semantic distinctions, e.g., substance and objects, natural objects and artifacts, dynamic and static, etc. (see LE2-4003b for details).
- *a hierarchy of domain labels*: these labels form a knowledge structure that groups synsets according to their domains, e.g., traffic, sports, medicine, etc.

A possible existing link between an ILI record and the above-mentioned hierarchies would allow this data to be applied to the corresponding synsets (and its hyponyms) of all languages. Figure 4 shows the global architecture of the EWN database.

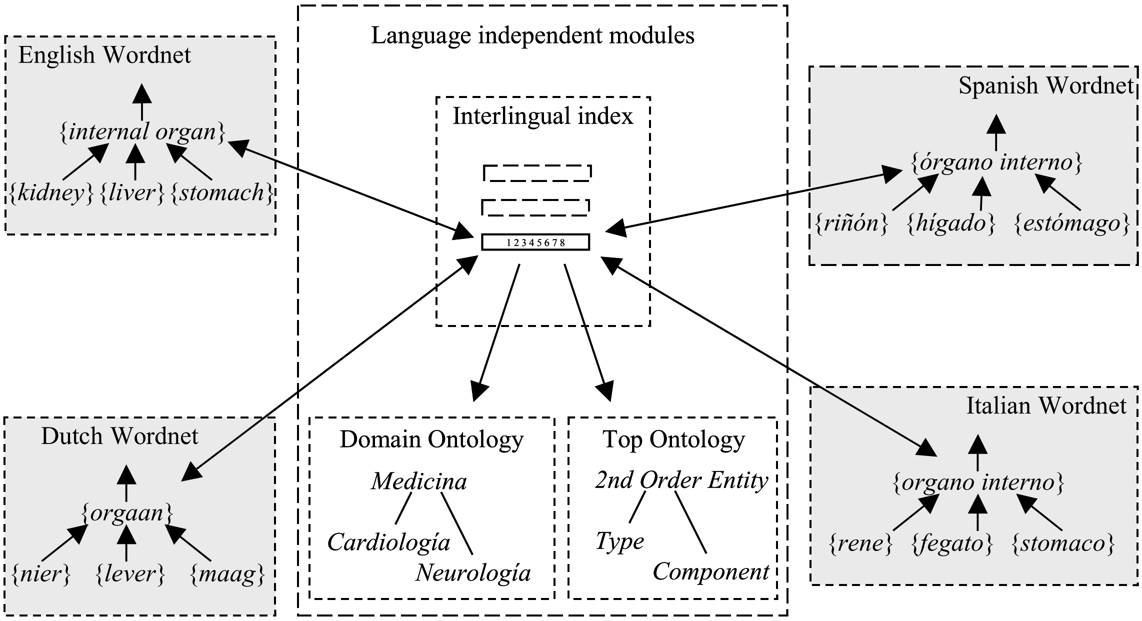


Figure 4. Global organisation of the EWN database.

Regarding adjectives, both EWN and WN (following Miller *et al.*, 1993) classify them into two major classes: descriptive and relational.

A descriptive adjective highlights —or gives value to— a particular attribute of the noun it modifies. Relational adjectives conversely do not qualify the noun but establish a connection with external entities or domains and classify the nouns (Soler, 1997). Of course, it is possible to find adjectives that may act as descriptive or relational, depending on the context, and this situation represents a difficult problem to solve. The solution adopted in this research is to give priority to relational adjectives.

For example, given the sequence *asma infantil* (‘infantile asthma’), the adjective *infantil* is acting on a specific property of this kind of asthma: the age of the patient; in this case it means that this variant of asthma affects young people[10]. In this case, the adjective *infantil* acts as a descriptive adjective. But considering the term *asma bacteriana* ('bacterial asthma'), the adjective *bacteriana* is informing that the patient has the disease 'asthma' produced by a bacterial infection, so it is relating the disease to its origin (relational adjectives).

Conversely, descriptive adjectives are represented in EWN —and WN— as a bipolar structure where each side of the structure has a nucleus and some satellite synsets. Taking for example the property *edad* ('age'), the values that it may take can be considered as having two main —and opposite— values: *joven* ('young') and *viejo* ('old'). Both values may be considered the nucleus of each side of the structure. This antonym relation is coded as 'near_antonym'. Other adjectives related with each nucleus are linked to it with the 'near_synonym' relation. In other words, the full graduation list of the values associated to the property is structured by adding the appropriate links. The relation "xpos_near_hyperonym"[11] is used to link both nucleus adjectives to the corresponding property. The resulting structure is represented in Figure 5. Every single domain selects which adjective must be used from each group in a particular situation .

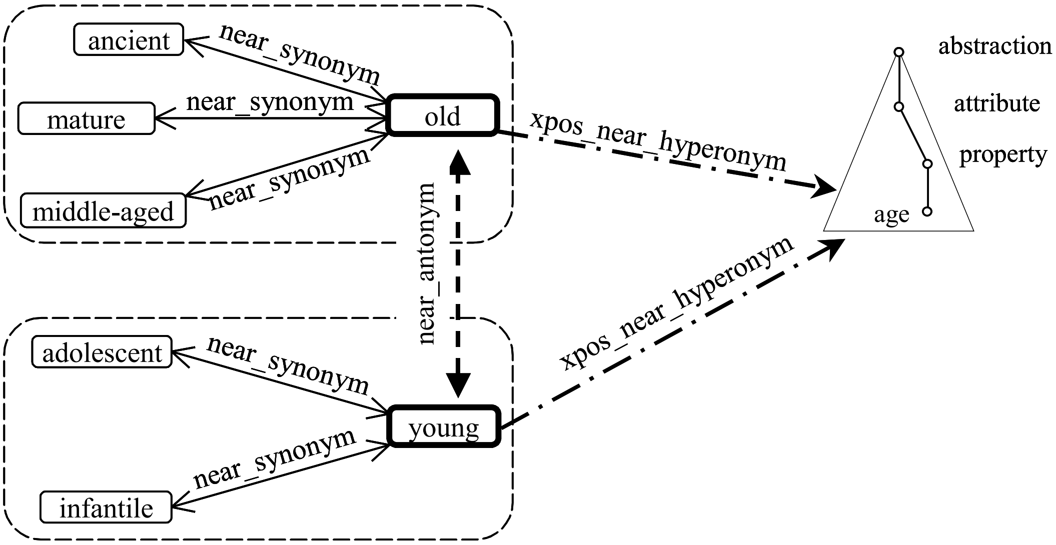


Figure 5. Descriptive adjective representation in EWN.

The organisation for the relational adjectives is simpler than for the descriptive ones. It is only necessary to relate the adjective with the corresponding

[10] The full definition found in *Diccionari Enciclopèdic de Medicina*, 1994 is the following: "*asma infantil: La que, en forma d'accessos, es presenta en els infants. Les crisis no són tan ben delimitades com a l'edat adulta, van sovint acompanyades de febre, tenen tendència a la insuflació pulmonar i s'associen a la presentació de raneres humides. Sempre hi ha una predisposició constitucional; diversos factors poden estar implicats en la seva aparició: rinofaringitis, causes psíquiques, al·lèrgens, etc. En general l'asma infantil desapareix en arribar l'infant a la pubertad*"

[11] The relation "xpos_near_hyperonym" may be paraphrased as follows: X (adjective) is an attribute value for the property Y (noun).

noun. For example, the adjective *bronquial* ('bronchial') must be related to the noun *bronquio* ('bronchus'). Figure 6 graphically shows the treatment of this kind of adjectives.

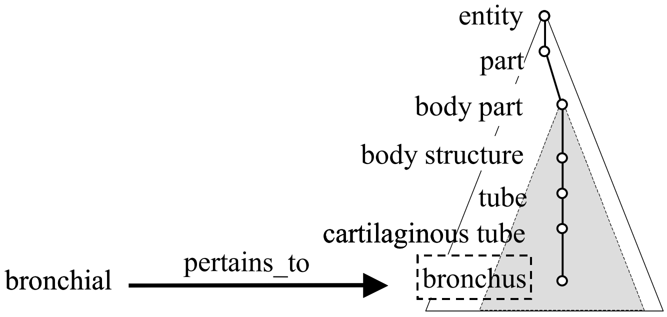


Figure 6. Relational adjectives representation in EWN.

Not the whole system has been defined and implemented at the same level, as for example the noun and verb hierarchies. They have been much more developed that those of adjectives and adverbs[12]. Similarly, many links have been defined but not effectively applied to the whole database.

3.2.1 Sample

The EWN lexical database provides three entries for "cell". Figure 7 only shows the information concerning the sense related to the genome domain.

[12] The efforts for each POS strongly depend on the organisations responsible for each language.

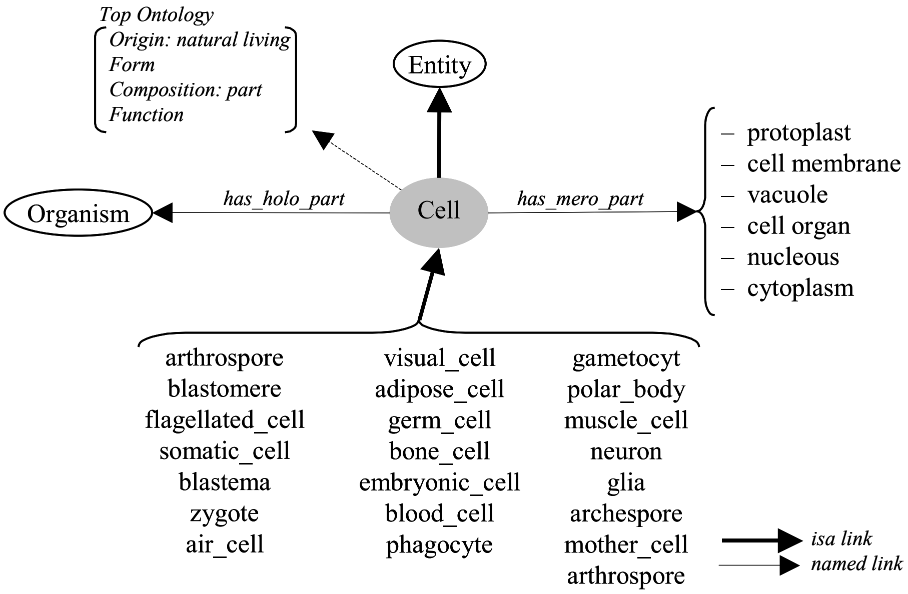


Figure 7. Representation for “cell” in EWN.

3.2.2 Analysis

EWN, as its ancestor WN, is a lexical-based ontology. Words are at the base of its concepts organisation. For such reason, it is very often not considered as an actual ontology. In spite of this, it is currently being used in a large number of projects mainly because this is an existent resource with a relatively large coverage of main European languages.

One of the points where WN (and consequently EWN) has received much criticism is regarding its asymmetrical granularity. It means that different areas of the ontology, reflecting different domains, are not equally developed.

The fact that it is a general knowledge resource produces that the same synset includes specialised and non specialised variants. For example, the synset that includes the variant *hematoma* (a medical term) includes also the variants *morado* and *cardenal* among some other variants belonging to a different specialisation level.

3.3 μKosmos

The MikroKosmos project, developed jointly by researchers at New Mexico State University, Carnegie Mellon University and various U.S. government agencies, presents a comprehensive study of a variety of computational linguistic microtheories which become central to the support of a KBMT system. The ultimate objective is to define a methodology for representing the meaning of natural language texts in a language-neutral interlingual format (Text Meaning Representation, TMR). The TMR is the result of the analysis of some texts, in any of the languages concerned in the machine translation process, and it serves as input to the generation process. The meaning of the texts is derived through the analysis of their lexical, syntactic, semantic, and pragmatic information. This information is represented in the TMR as elements which have to be seen in terms of an independently motivated model of the world, that is, an ontology. The lexicon provides the link between the ontology and the TMR because the meanings of open-class lexical items are defined and established depending on their mappings into ontological concepts and their relationship with the TMR structure. From a pragmatic approach, the project deals with a concrete linguistic field: Japanese and English joint-ventures corpora. The analysis was also extended to Spanish and French, using similar texts.

In the μKosmos project, an ontology is defined as «a computational entity and a resource containing knowledge about what "concepts" exist in the world and how they relate to one another. A concept is a primitive symbol for meaning representation with well defined attributes and relationships with other concepts. An ontology is a network of such concepts forming a symbol system where there are no uninterpreted symbols (except for numbers and a small number of known literals)[13].» Moreover, from a knowledge-based approach to machine translation, the interlingual meaning representation is extracted from representations of word meanings in computational lexicons and from representations of world knowledge captured in the ontology. The ontology becomes a language-independent knowledge source containing the set of symbols and the possible relationships among them.

Thus, an ontology created for NLP purposes becomes a body of knowledge about the world (or a specific domain) that:

- gathers together primitive symbols used for knowledge representation;
- organises these symbols in a hierarchical way;
- interconnects these symbols by means of a system of semantic relations.

In the μKosmos design, an ontology is a database with information about:

- the categories (or concepts) existing in a world or in a particular domain;
- the properties of these categories;
- how they relate to one another.

In a MT oriented project, it is important to use an ontology in order to:

- provide the basis for the TMR;
- enable lexicons for different languages to share knowledge;
- enable source language analysers and target language generators to share a specific knowledge.

It is worth mentioning that in every language, the lexicon is organized in superentries identified using the form of each dictionary entry. Within a superentry, individual lexemes are represented in each language by frames. Each lexical unit has different information levels, such as:

- CAT (syntactic category)
- ORTH (orthography, abbreviations)

[13] For a more detailed description of the μKosmos design criteria, see: *http://crl.nmsu.edu/Research/Projects/mikro/* by Kevi Mahesh, 1996.

- PHON (phonology)
- MOPRH (irregular forms or class information)
- SYN (syntactic features)
- SYN-STRUC (dependency features and subcategorisation)
- SEM (lexical semantics and semantic representation)
- LEXICAL-RELATIONS (collocations)
- PRAGM (pragmatic information)
- ANNOTATIONS (user or lexicograph information, examples, etc.).

The ontology developed in the µKosmos framework has, in some cases, up to ten subtrees. The Top-Level (*all*) has three first levels, *object*, *event*, and *property*. This last category is the one including *attributes* and *relations* as seen in Figure 8.

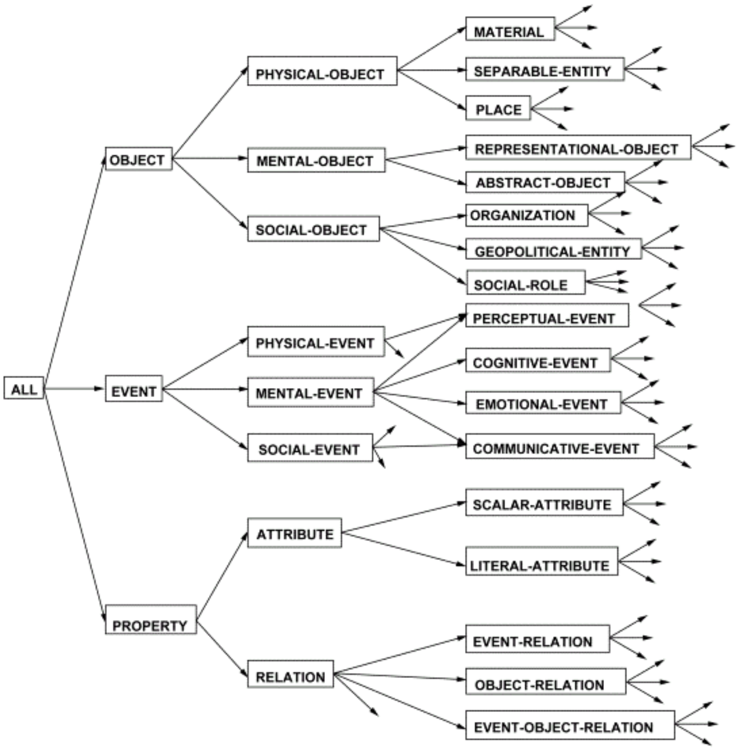


Figure 8. Top levels of the hierarchy.

As we have already mentioned, the µKosmos NLP architecture contains four main modules for textual analysis:

- the lexicon
- the ontology
- the interlingual representations
- the microtheories

which are interrelated as showed in Figure 9.

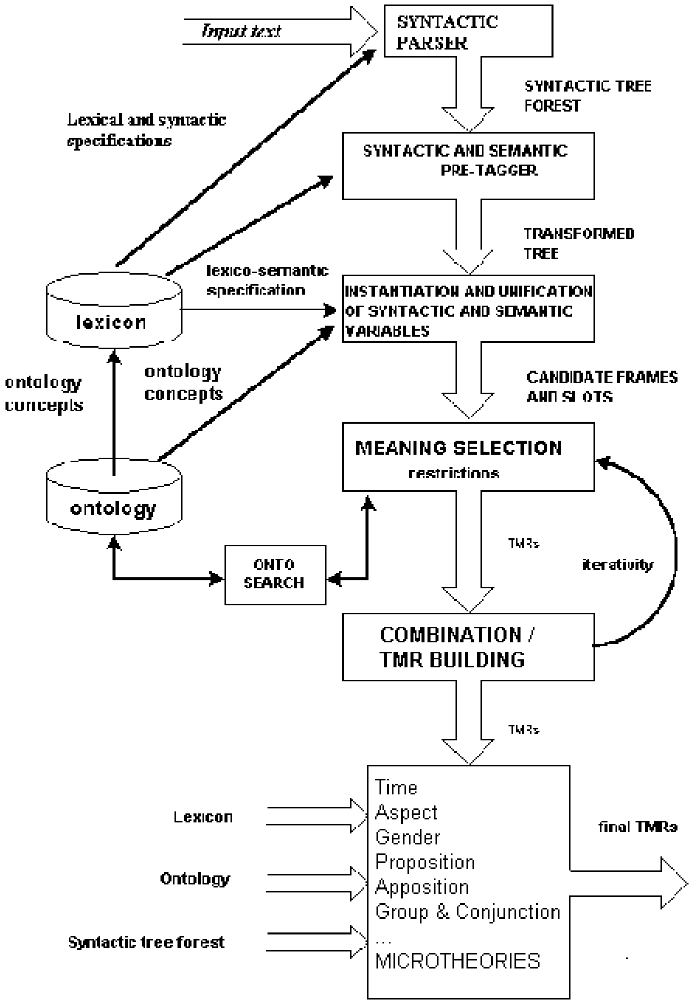


Figure 9. The µKosmos NLP architecture

3.3.1 Sample

In Figure 10 we show the information drawn from the ontology. It should be noted that such figure shows only the attributes, local relations of the "cell" concept or those relations available by inheritance[14].

[14] This information has been obtained through the use of OntoTerm. This tool implements the µKosmos ontology through the use of a relational database for microcomputers. Given this framework, OntoTerm probably does not seem to fully implement all features of the original µKosmos design.
No data is available from MikroKosmos web site in spite of their claim that any user may consult (?) the ontology or obtain a license for academic purposes.

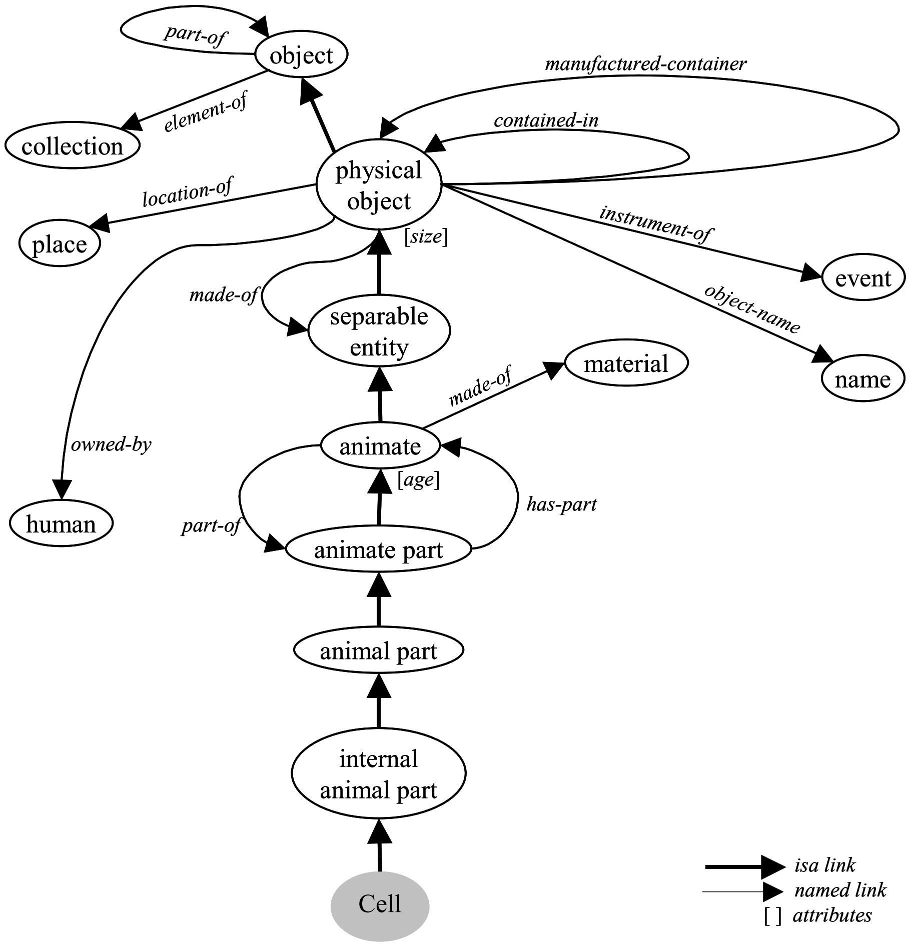


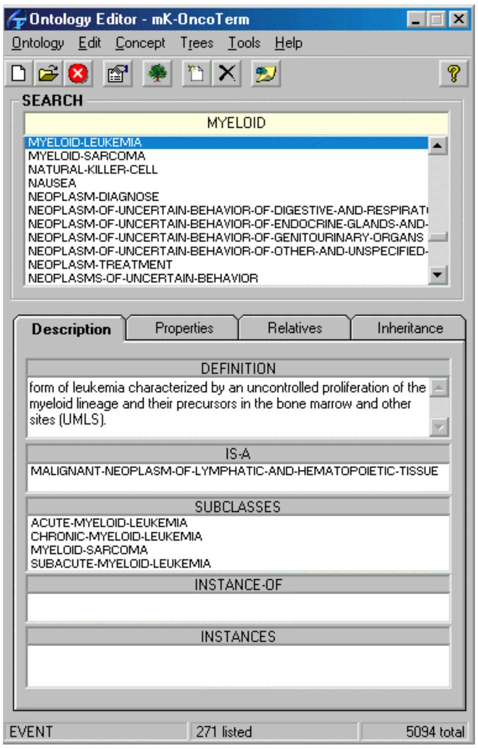
Figure 10. Representation for “cell” in μKosmos.

3.3.2 OntoTerm: an Ontology-based Terminology Management System

OntoTerm is a terminology management system (TMS) built by Antonio Moreno to overcome some of the problems of the existent terminology database management systems in two major senses. On the one hand, it is based on a *Conceptual modelling*, that is, the object domain or subject field must be conceptually structured prior to entering language-specific terms. The construction of an ontology becomes, in this approach, the first step in the construction of a term base. On the other hand, it is oriented to a terminology information exchange because it implements the ISO standards for terminology exchange: Martif (ISO 1220) and all the data categories from the CLS Framework (ISO 1620). OntoTerm does not allow to enter terms in a termbase unless you have previously entered and defined a concept explicitly in the ontology. We can then say that it is an ontology-based TMS including four modules:

- The Ontology Editor
- The TermBase Editor
- The Ontology Navigator
- The HTML Report Generator

The Ontology Editor is the module which allows the user to create a new ontology[15] or to open an existing one, and it becomes also the first screen that appears when the user runs the program. Once the program is running, a given ontology is visualised as follows:


Ontology Editor - mK-OncoTerm
Ontology Edit Concept Trees Tools Help
SEARCH
Description
Properties
Relatives
Inheritance

MYELOID

MYELOID-LEUKEMIA
MYELOID-SARCOMA
NATURAL-KILLER-CELL
NAUSEA
NEOPLASM-DIAGNOSE
NEOPLASM-OF-UNCERTAIN-BEHAVIOR-OF-DIGESTIVE-AND-RESPIRAT
NEOPLASM-OF-UNCERTAIN-BEHAVIOR-OF-ENDOCRINE-GLANDS-AND-
NEOPLASM-OF-UNCERTAIN-BEHAVIOR-OF-GENITOURINARY-ORGANS
NEOPLASM-OF-UNCERTAIN-BEHAVIOR-OF-OTHER-AND-UNSPECIFIED-
NEOPLASM-TREATMENT
NEOPLASMS-OF-UNCERTAIN-BEHAVIOR

DEFINITION
form of leukemia characterized by an uncontrolled proliferation of the myeloid lineage and their precursors in the bone marrow and other sites (UMLS).

IS-A
MALIGNANT-NEOPLASM-OF-LYMPHATIC-AND-HEMATOPOIETIC-TISSUE

SUBCLASSES
ACUTE-MYELOID-LEUKEMIA
CHRONIC-MYELOID-LEUKEMIA
MYELOID-SARCOMA
SUBACUTE-MYELOID-LEUKEMIA

INSTANCE-OF

INSTANCES

EVENT | 271 listed | 5094 total

Figure 11. OntoTerm main screen

The Ontology Editor permits to create or modify ontologies; to add, modify or delete selected concepts from a particular ontology; to view the whole or a partial piece of the tree ontology and to include some annotations concerning the concept.

For each concept, it is possible to include a description (concept definition, hyperonym and hyponym relations, instances of the concept), its properties, its relatives and the inheritance. The system organises information by means of concepts, attributes and relations. Attributes and relations can be locally assigned or inherited. As for inheritance, it is worth mentioning that the system presents exclusive inheritance (all relations and attributes which correspond to the concept itself and those inherited from its parent) or cumulative inheritance (exclusive inheritance plus all relations and attributes inherited from its hyperonymy path).

The second main module of OntoTerm is the Ontology Navigator. It allows to view —in a readable format and by navigation through hyperlinks on the pages

[15] When creating a new ontology, the system provides a subset of µKosmos ontology as a kind of upper-level ontology.

generated— all information extracted from the ontology, that is, concepts organised in a hierarchical tree, their definition and all relations among concepts.

Moreover, from a linguistic and terminological point of view, it is worth noting the TermBase Editor provided by OntoTerm. It is the module for creating, editing and browsing terminological units corresponding to concepts. It uses a standard set of data categories in order to assign linguistic information to each concept of the ontology. The TermBase Editor is visualised as follows:

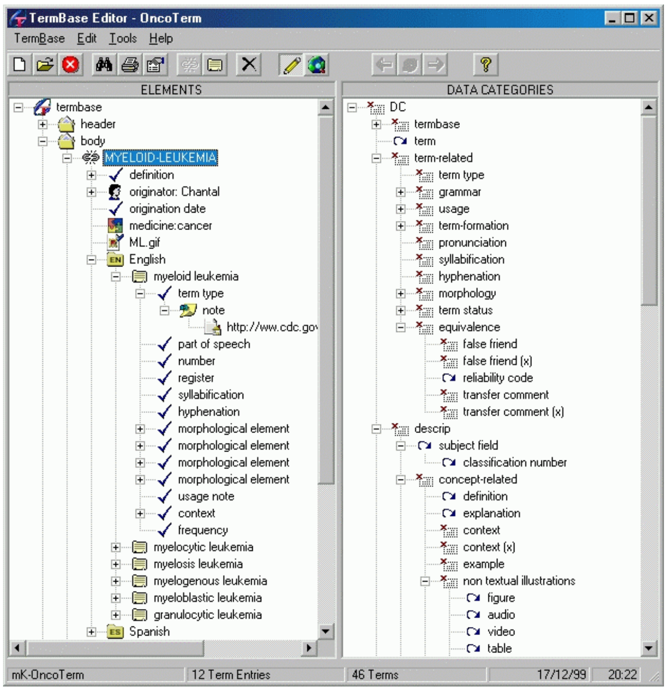


Figure 12. OntoTerm TermBase Editor.

Finally, the HTML Report Generator is a tool that allows the user to select those concepts indicated to generate web pages for. The HTML Report Generator is the output format for all information included in the previously defined modules of the OntoTerm tool[16].

3.3.3 Analysis

In spite of the claim that this is a general knowledge ontology, it is not clear which is its actual coverage for domains different from those of interest for the KBMT system. Some of their concept organisation for domains like medicine are not completely clear and convincing.

One of its main advantages is the existence of a support tool: OntoTerm. The facilities of this tool ease all management operations. It should be noted that this tool gives many facilities to the user, as for example, it is possible to add any kind of

[16] For more information about the HTML Report Generator result see http://www.ontoterm.com.

relations. However, a rigorous and systematic use of the system is necessary to avoid inconsistencies, duplicates, etc.

3.4 SIMPLE

The SIMPLE (*Semantic Information for Multifunctional Plurilingual Lexica*) project, which can be considered as a follow-up to the European PAROLE project, aims at providing a core set of language resources for the EU languages. Being an European project, it covers 12 different languages. Researchers involved believe that the production of a reusable core lexicon with morphological, syntactic and semantic information will allow many applications to be developed with a shorter time-scale and overall cost. The main objective is to add a semantic layer to the existing morphological and syntactic levels of the PAROLE resources. The semantic lexicons are general language corpus-based and the final objective is to cover about 10,000 word meanings (7000 for nouns, 2000 for verbs and 1000 for adjectives). The exchange format for the lexicons is SGML and, as for the morphological and syntactic layers, all semantic lexicons share the same DTD.

Until now, research work has been oriented to define a general architecture for encoding lexical meaning. For this reason, researchers have focused on the development of a top ontology of semantic types and a set of formal tools for the analysis of lexical items. At present, most efforts are devoted to ensure consistency and to facilitate the encoding of the lexicon. In the project framework, the ontology becomes a semantic tool used to represent main word senses related to each lexical entry by means of a template.

So, as it is mentioned in the SIMPLE Linguistic Specifications Paper, «each word sense corresponds to a given semantic type. Each semantic type is actually a cluster of structured information. Semantic types differ in terms of how much information they convey. In other words, word senses differ in their degree of complexity, which is explicitly part of their semantic type.»

In order to uniform information given for all 12 languages, SIMPLE has an information pattern containing the following:

- Semantic Type
- Template
- SemU

We have to bear in mind that SIMPLE work is not oriented to build an ontology. In contrast, it represents an attempt to encode lexical semantic information for an important number of languages. The formal specification for the representation and encoding information follows this guideline:

- Semantic type
- Domain information
- Glossa
- Argument structure
- Selectional restriction on the arguments
- Event type for verbs
- Link of the arguments to the syntactic subcategorisation frames

- Type hierarchy information
- Qualia information
- Information about regular polisemous alternation
- Information concerning cross-part of speech relations
- Eventual collocations retrieved from corpus
- Synonymy relations

All these items follow recommendations of the EAGLES Lexicon/Semantics Working Group and their organisation is based on extensions of Generative Lexicon Theory (Pustejovsky, 1995).

The SIMPLE ontology is divided into a *core ontology* (formed by those types which have been identified as the central and common ones for the construction of the different lexicons) and the *recommended ontology* (formed by more specific types that constitute lower nodes in the hierarchy). Thus, the ontology is used to provide the Conceptual Core shared by all the lexicons. It is the *Template* which provides the interface between the ontology and every lexicon. An schematic representation of a template, which includes the SemU position in the ontology, is as follows:

SemU:	Identifier of a SemU
SynU:	Identifier of the SynU to which the SemU is linked
BC Number:	Number of the corresponding Base Concept in EuroWordNet
Template_Type:	Semantic type of the SemU
Template_Supertype:	Semantic type which dominates the Template_Type of the SemU in the type-hierarchy
Unification_path:	Unification history of a template (for unified top-types)
Domain:	Domain information from LexiQuest domain list
Semantic Class:	One of the classes provided by LexiQuest
Gloss:	Lexicographic definition
Event Type:	Event sort (for event SemUs only)
Predicative Representation:	Predicate associated with the SemU, and its argument structure
Selectional Restr.:	Selectional restrictions on the arguments
Derivation:	Derivational relations between SemUs
Formal:	Formal relation between SemUs
Agentive:	Agentive relations between SemUs
Constitutive:	Constitutive relations between SemUs Constitutive semantic features
Telic:	Telic relations between SemUs
Synonymy:	Synonyms of the SemU
Collocates:	Collocate information
Complex:	Polysemous class of the SemU

Having reviewed all linguistic specifications concerning SIMPLE, we have to highlight all efforts carried on in order to encode semantic information even though the ontology is not developed enough in order to be used in NLP applications.

3.4.1 Sample

SIMPLE, like EWN, provides three entries for "cell", each one following its own template, that are shown bellow:

```
<SemU id="celula1_Organicobject"
example="e.g., célula (Unidad fundamental de los seres vivos, con
cierta autonomía)"
naming="célula"
weightvalsemfeaturel="WVSFTemplateOrganicobjectPROT
WVSFTemplateSuperTypeConcreteentityPROT
TSVP_OBJECT_TS_classificateur_de_nom_C TSVP_PLUS_TS_PART_T">
<RWeightValSemU semr="SRIsapartof" target="cuerpo_Organicobject"
weight="ESSENTIAL">

<SemU id="celula1_Instrument"
example="e.g., se ha disparado la célula fotoeléctrica (dispositivo
que transforma las variaciones de intensidad luminosa en variaciones
de de intensidad de una corriete)"
naming="célula"
weightvalsemfeaturel="WVSFTemplateInstrumentPROT
WVSFUnificationPathConcreteentity-ArtifactAgentive-TelicPROT
TSVP_APPARATUS_TS_classificateur_de_nom_C">
<RWeightValSemU semr="SRCreatedby" target="hacer_X"
weight="PROTOTYPICAL">
<RWeightValSemU semr="SRUsedfor" target="hacer_X"
weight="PROTOTYPICAL">

<SemU id="celula1_HumanGroup"
example="e.g., la célula del partido (Unidad o grupo separado de una
organización)"
naming="célula"
weightvalsemfeaturel="WVSFTemplateHumanGroupPROT
TSVP_PLUS_TS_HUMAN_T TSVP_PLUS_TS_COLLECTIVE_T
WVSFTemplateSuperTypeGroupPROT
TSVP_GROUP_NAMES_TS_classificateur_de_nom_C">
<RWeightValSemU semr="SRHasasmember" target="persona_Human"
weight="PROTOTYPICAL">
```

3.4.2 Analysis

The previous sample shows how information is structured in a SIMPLE lexical entry. As we can see, linguistic information is structured in terms of SemUs which contain the identification of the lexical entry. Each lexical entry includes an usage example and entries are linked, when corresponding, by the weight value semantic feature. In fact, the ontology is only used as a way to control information included in the template but it is not explicitly used for the lexicon organisation.

Most concretely, SemUs include attributes and relations (apart from the predicate). The attributes may contain information about the ontology and any other relevant information (i. e., ‘Connotation’, ‘Plus-Edible’, ‘Plus-Fictive’, ‘Plus-Human’, ‘Plus-Gradable’, etc.). Relations aim to describe the SemU in terms of the qualia structure (‘tellic’, ‘agentive’, etc.). Finally, semantic codification includes encyclopaedic and linguistic specification to explicitate the linguistic use. So, in the case of verbs, and for all predicates, it is intended to reflect the argument structure and the restrictions of selection that can be applied. All this information describes the lexical semantics for each unit and not its position on a given ontology.

3.5 UMLS

Among the domains that own specific resources, medicine is the main area that needs to be mentioned. This is mainly due to the largest request of normalisation[17] in this domain. Most resources have been developed for English. The only resource that includes some information for Spanish is the UMLS (*Unified Medical Language System*) project. It is a long term research and development effort to facilitate the retrieval and integration from multiple machine-readable biomedical information sources (UMLS, 2001). This long term project was initiated in 1986 and supported by the NLM (*National Library of Medicine*)[18].

The UMLS resources are currently used in a number of applications[19] (mainly related to information retrieval and integration from different resources) or research activities as the term extraction system described in Maynard (1999).

Currently, UMLS resources partially represent the information related to the Human Genome Project. However, several researchers (Yu *et al.*, 1999) propose different kinds of modifications and/or enlargement for adapting this resource to this project.

The UMLS is available free of charge to U.S. and international users. Anyway, the use of the Metathesaurus may require additional agreements with producers of the individual vocabularies it contains.

UMLS is a set of three knowledge sources: a) UMLS Metathesaurus, b) UMLS Semantic Network, and c) SPECIALIST Lexicon. The following sections will briefly describe the above mentioned sources.

UMLS Metathesaurus

The Metathesaurus contains information about biomedical concepts and terms obtained from a set of controlled vocabularies and classification systems. It preserves the names, meanings, hierarchical contexts, attributes, and inter-term relationships present in its source vocabularies; it adds certain basic information to each concept; and it establishes new relationships between terms from different source vocabularies. Its scope is determined by the combined scope of its source vocabularies. The Metathesaurus is produced by automated processing of machine-readable versions of its source vocabularies, followed by human review and editing by domain experts. Its structure allows to integrate resources from languages other than English like most of the European languages.

The 2001 UMLS Metathesaurus includes:

- 800,000 concepts
- 1,400,000 “terms” (Eye, Eyes, eye = 1)

[17] See, for instance, recent existing codification systems, ICD-9-CM, ICD-10, SNOMED and Read Code among others.

[18] More information about this project is available in [NLM, 1998] or through the following url: *http://www.nlm.nih.gov/research/umls/umlsmain.html*

[19] At the web page *http://www.nlm.nih.gov/research/umls/umlsapps.html* it is possible to consult a list of applications and services using this resource.

- 1,900,000 “strings”/concept names (Eye, Eyes, eye = 3)
- more than 50 source vocabularies

For managing this data, the UMLS distribution includes MetamorphoSys. It is a tool that allows the creation of a customized version of the Metathesaurus[20]; in this way, users may: exclude unnecessary vocabularies, alter “preferred name” precedence, exclude vocabularies as required by the license agreement, etc.

UMLS Semantic Network

The Semantic Network provides a consistent categorisation of all the concepts represented in the UMLS Metathesaurus. In the 2001 version there are 134 semantic types and 54 relations between them. The semantic types are the nodes in the Network, and the relationships between them are the links. Each node has a simple denominative tag. There are major groupings of semantic types for organisms, anatomical structures, biologic function, chemicals, events, physical objects, and concepts or ideas. All semantic types are divided in two major groups tagged as: ‘entity’ and ‘event’. Each concept of the Metathesaurus is associated with one or more semantic types.

The primary link is the “*is-a*” link. This establishes the hierarchy of types within the Network and it is used for deciding on the most specific semantic type available for assignment to a Metathesaurus concept (hyperonymic relation). In addition, a set of non-hierarchical relations between the semantic types are identified. These are grouped into five major categories, which are themselves relations: ‘physically related to’, ‘spatially_related_to’, ‘temporally_related_to’, ‘functionally_related_to’ and ‘conceptually_related_to.'

The relations are stated between semantic types and do not necessarily apply to all instances of concepts that have been assigned to those semantic types. That is, the relation may or may not hold between any particular pair of concepts. For example, although the relation ‘evaluation_of’ holds between the semantic types ‘Sign’ and ‘Organism Attribute’, a particular sign or a particular attribute may not be linked by this relation. Thus, signs such as “overweight” and “fever” are evaluations of the organism attributes “body weight” and “body temperature”, respectively. However, “overweight” is not an evaluation of “body temperature”, and “fever” is not an evaluation of “body weight”.

The organisation of the net allows a semantic type to receive information from its ancestors using an inheritance mechanism. In some cases, there will be a conflict between the placement of the types and the link to be inherited. In such cases the inheritance may be blocked.

[20] It has to be taken into account that MetamorphoSys has heavy computer requirements: a minimum of 256 MB of physical memory, as well as 8 GB recommended free disk space. It runs on operating systems like Unix, Linux and Windows.

SPECIALIST Lexicon

The SPECIALIST lexicon is an English language lexicon with many biomedical terms. It has been designed to be used in a NLP system. The UMLS 2000 version includes about 108,000 lexical records.

The lexicon entry for each word or term records syntactic, morphological, and orthographic information. Lexical entries may be single or multi-word terms. Entries which share their base form and spelling variants, if any, are collected into a single lexical record.[21]

Lexical information includes syntactic category, inflectional variation (e.g., singular and plural for nouns, the conjugations of verbs, the positive, comparative, and superlative degrees for adjectives and adverbs), and allowable complementation patterns (i.e., the objects and other arguments that verbs, nouns, and adjectives can take). The lexicon recognizes eleven parts of speech: verbs, nouns, adjectives, adverbs, auxiliaries, modals, pronouns, prepositions, conjunctions, complementizers, and determiners.

The basic sentence patterns of a language are determined by the number and nature of the complements taken by verbs. The UMLS lexicon recognizes five broad complementation patterns: intransitive, transitive, ditransitive, linking and complex-transitive. Verb entries also encode each of the inflected forms (principal parts of the verb). Verbs are inflectionally classified as regular, Greco-Latin regular or irregular units. Noun entries describe the inflection of the nouns and spelling variations. Complementation patterns for nouns and nominalisation information are also included when relevant. In addition to inflection and complement codes, adjectives in the lexicon have position codes to indicate the syntactic positions in which they may occur. An adjective may be a qualitative, classifying, or colour adjective. Adverbs in the lexicon are coded to indicate their modification properties. The lexicon recognizes sentence, verb phrase and intensifier type adverbs, and classifies sentence and verb phrase adverbs into manner, temporal and locative types.

The distribution includes a set of lexical programs, indexes, and databases that may be useful for developers who work with the UMLS knowledge sources. The set of tools allows operations like lowercasing, uninversion, sorting words in a multi-word term, stopword removal, possessive marker removal, punctuation removal, and generation of inflectional and derivational variants. The databases allow to know derivational variants (alternations such as “aphasic/aphasia"), closely related terms that mean the same thing but may have a different syntactic category (e.g., “hepatocellular/liver cells”), spelling alternations (e.g, “foetal/fetal”) and neoclassical combining forms with their meanings (e.g., “heart/cardi(o)”).

As mentioned above, all linguistic information is relevant to the English language, no information is included for other languages.

[21] The base form is the uninflected form of the lexical item; the singular form in the case of a noun, the infinitive form in the case of a verb, and the positive form in the case of an adjective or adverb.

3.5.1 Sample

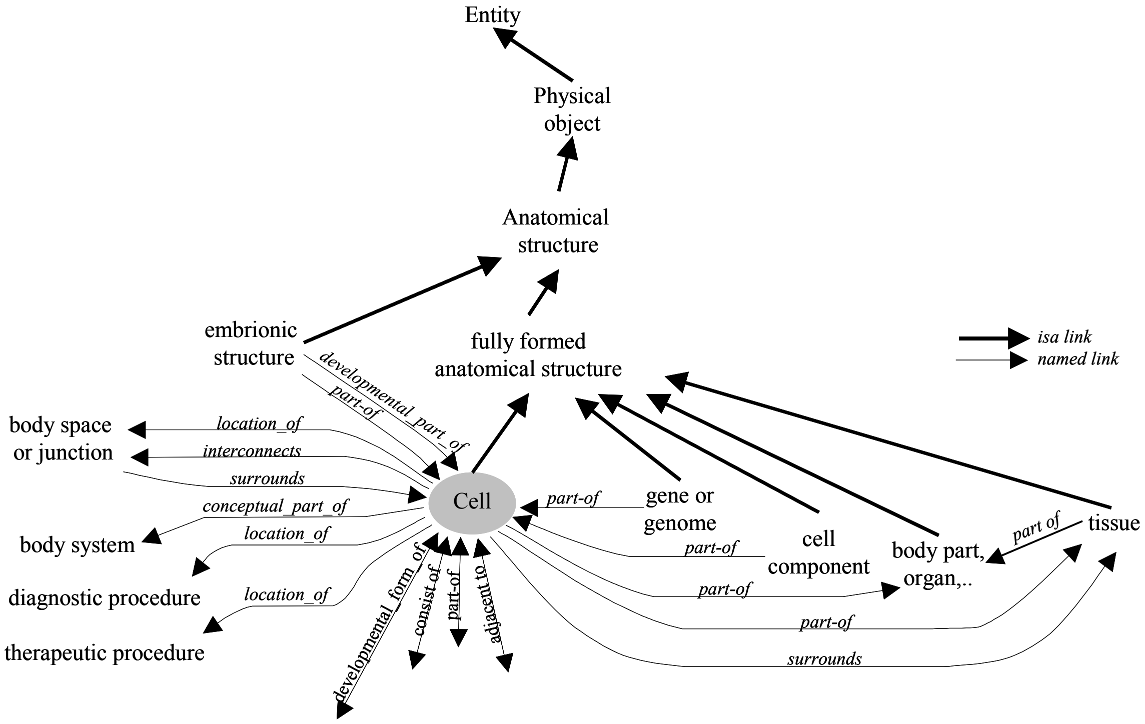


Figure 9. Representation for 'cell' in the UMLS Metathesaurus.

3.5.2 Analysis

On the one hand, the UMLS Semantic Network provides a very detailed organisation for medical concepts that has proved to be expandable to other medical subdomains recently emerged.[22]

On the other hand, although the Metathesaurus structure allows to include languages different from English, some limitations arise in using this kind of resources (at least for Spanish in the 1998 Edition):

- The coverage for such languages is relatively limited[23];
- Spanish entries do not have special characters like á, É, ñ... (they are reduced to a, E, n);
- Words taking part of every concept are not lemmatised[24].

[22] See for example the proposal of Yu *et al.* (1999). It should be noted that it seems to be a mere theoretical exercise because the authors do not provide any actual results in using their proposal.

[23] In 1998 Edition, the only source for Spanish in the Metathesaurus included almost 24 K concepts (the total number of concepts, at that time, was about 478 K in more than 40 resources). Later versions increase these figures but the basic problem is still remaining.

[24] See for example *autoantígenos* instead of *autoantígeno* or *complicaciones postoperatorias* instead of *complicación postoperatoria*. This problem is not present in English vocabularies due to the existence of a "lemmatised index".

A corrective action for solving some of these problems is possible but it may be time-consuming. Yet, it has the advantage to have a more complete and checked domain ontology specialised in the medical domain.

Regarding the concepts included, some of them are doubtful and some other are, at least, curious. See for example the following words that are related to UMLS concepts: ‘deportes’, ‘maltrato conyugal’, “distensiones y esguinces’, and so on.

3.6 Other Ontologies

In this section, we will briefly introduce other ontologies (in alphabetical order) which can be consulted mostly via the web. These ontologies have to be mentioned in order to provide a general framework about what it is available at present in this field, but they will not be analysed in depth because their distribution is generally restricted and most of them are specialised domain oriented.

3.6.1 $(KA)^2$

The *Knowledge Annotation Initiative of the Knowledge Acquisition Community* $(KA)^2$ was an initiative launched to develop an ontology that models the knowledge acquisition community (its researchers, topics, products, etc.). This ontology is decribed in Benjamins *et al.* (1999) and it will form the basis to annotate WWW documents of the knowledge acquisition community in order to enable intelligent access to these documents. $(KA)^2$ is an open joint initiative where the participants are actively involved in:

(i) a distributive ontological engineering process to model the knowledge acquisition community (a domain ontology), and

(ii) annotating webpages relevant for the KA community (the instances of the domain ontology).

$(KA)^2$ aims at “intelligent” knowledge retrieval from the Web and automatic derivation of “new” knowledge". In other words, it aims at knowledge-based reasoning on the Web, as opposed to the more usual information retrieval. Another objective of the initiative concerns a distributive ontological engineering process.

3.6.2 EDR

EDR (*Electronic Dictionary Research*) is a voluminous and rather exhaustive dictionary, described by Yokoi (1995) and developed in Japan by a private and public enterprise consortium[25]. The dictionary can be considered the biggest lexical database available. It contains five modules: a set of two monolingual dictionaries, one for English (190,000 lexical entries) and the other for Japanese (260,000 lexical entries), an English-Japanese bilingual dictionary (230,000 lexical entries), a dictionary of concepts (400,000 entries), two dictionaries of collocations for English (460,000 entries) and for Japanese (900,000 entries) and a dictionary of technical terms. This

[25] For more information, see *http://www.iijnet.or.jp/edr/*, including detailed information about this project.

application includes a lot of monolingual and bilingual information (i.e. complete data about syntactic and semantic categorisation patterns) with a concept hierarchy organised similarly to EWN. The main difference between EDR and EWN is that the nodes without a lexicalisation in the hierarchy are explicitely indicated.

3.6.3 EngMath

It is an ontology for mathematical modelling in engineering (Gruber *et al.*, 1994). The ontology includes conceptual foundations for scalar, vector, and tensor quantities, physical dimensions, units of measure, functions of quantities, and dimensionless quantities. The conceptualisation builds on abstract algebra and measurement theory explicitly designed for knowledge sharing purposes. The ontology is being used as a communication language among cooperating engineering agents, and as a foundation for other engineering ontologies.

3.6.4 Enterprise Ontology

The Enterprise Ontology is a collection of terms and definitions relevant for business enterprises. The ontology was developed in the Enterprise Project by the Artificial Intelligence Applications Institute at the University of Edinburgh together with a consortium of private companies. The project was supported by the UK's Department of Trade and Industry under the Intelligent Systems Integration Programme.[26]

Conceptually, the Enterprise Ontology is divided into a number of main sections such as activities and processes, organisation, strategy, marketing, and time. The number of concepts defined in the ontology is 130.

3.6.5 Generalized Upper Model

The *Generalized Upper Model* (GUM) is a general task and domain-independent linguistically motivated ontology that supports sophisticated natural language processing while significantly simplifying the interface between domain-specific knowledge and general linguistic resources. We also expect the proposed ontology to provide a solid basis for domain modelling in general, not only where natural language is concerned.

The GUM is, in the terms of Bateman (1994), an interface ontology. It occupies a level of abstraction midway between surface linguistic realisations and 'conceptual' or 'contextual' representations. It enables abstraction beyond the concrete details of syntactic and lexical representations, while still maintaining sufficient close contact with linguistic realisations to be solidly founded on objective criteria. That is: if there is no specifiable lexicogrammatical consequences for a 'concept', then it does not belong in the Generalized Upper Model. Finally, the Generalized Upper Model is not theory specific; it does not aim to be a lexical semantics; it is not language specific and, in any case, it can be considered an interlingua.

[26] See more information about the project at: *http://www.aiai.ed.ac.uk/project/enterprise/*.

3.6.6 Pennman Upper Model

The Pennman Upper Model (Bateman *et al.*, 1990) was built from work done in natural language generation at ISI in the 1980's. It emerged as a general and reusable resource, supporting semantic classification at an abstract level that was task and domain independent. One of its key features was the methodology underlying its construction. It was written in LOOM[27]. The original Pennman Upper Model was merged with the KOMET German Upper Model to create a single unified upper model and they have been the basis for the General Upper Model described above.

3.6.7 PhysSys

PhysSys is an engineering ontology oriented to modelise, simulate and conceive physical systems. It includes three ontologies on engineering concerning the system presentation, the physical processes and the descriptive mathematical relations involved. According to Borst *et al.* (1996), this ontology was built to attain knowledge sharing and to reuse it in complex industrial applications. Moreover, the practical use of this kind of ontologies in large-scale applications is not restricted to knowledge-based systems, for the domain of engineering system modelling, simulation and design. PhysSys ontology provides the foundation for the conceptual database schema of a library of reusable engineering model components, covering a variety of disciplines such as mechatronics and thermodynamics.

From the application scenario, it is possible to identify various viewpoints that are seen as natural within a large domain: broad and stable conceptual distinctions that give rise to a categorisation of concepts and properties. This provides a first mechanism to break up ontologies into smaller pieces with strong internal coherence but relatively loose coupling, thus reducing ontological commitments. In the criteria design, it is assumed that general and abstract ontological 'super' theories, for example mereology, topology, graph theory and system theory, can be used and reused as generic building blocks in ontology construction, which becomes an important element in knowledge sharing across domains. Ontology projections can occur in simple forms such as include-and-extend and include-and-specialise, but they are in their richest form very knowledge-intensive, being in fact themselves full-blown ontological theories.

3.7 Ontolingua[28]

Ontolingua is the generic name for the Ontolingua Server, a tool for collaborative ontology construction. Ontolingua researchers have developed a set of tools and services to support not only the development of ontologies by individuals, but also the process of achieving consensus on common ontologies by distributed groups.

[27] LOOM is a knowledge representation language developed at ISI. For more information, see *http://www.isi.edu/isd/LOOM/LOOM-HOME.html#OVERVIEW.*

[28] For a complete description of the system, see *http://ontolingua.stanford.edu* or go directly to *http://www-ksl-svc.stanford.edu:5915/.*

These tools use the web to enable wide access and provide users to publish, browse, create[29] and edit ontologies stored on the ontology server.

Figure 10 shows a schematic view of the system. The leftmost box depicts the general-purpose Ontolingua editor and server. The server provides access to a library of ontologies, it allows new ontologies to be created, and existing ontologies to be modified. There are three primary modes of interaction with the Ontolingua Server:

a) remote collaboration: using the web, remote users can create new ontologies and browse the already existing ones stored at the server;
b) remote applications: it is possible for remote applications to query ontologies stored at the server.
c) stand-alone applications: the system allows to translate an ontology into a format to be used by a specific application at the user host.

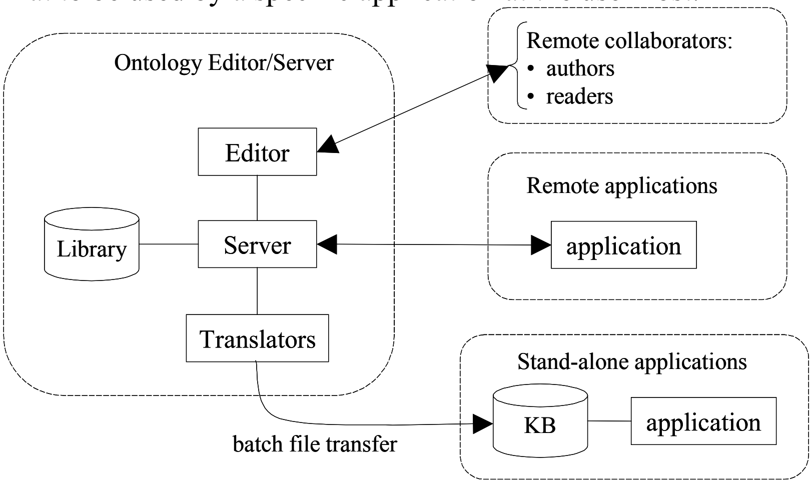


Figure 10. Architecture of the Ontolingua Server

The design of the web-based interface and the underlying structure is detailed in Farquhar *et al.* (1996). These tools and service providing facilities for the use of ontologies and knowledge representation are:

- A semi-formal representation language that supports the description of terms both informally in natural language and formally in a rigorous computer interpretable knowledge representation language. It is used an extended version of the

[29] A guided tour for developing ontologies can be followed at *http://www-ksl-svc.stanford.edu:5915/doc/frame-editor/guided-tour/index.html.*

Ontolingua language (Gruber, 1992) which provides a frame-like syntax and full first order logic as specified in the Knowledge Interchange Format (KIF)[30].

- Browsing and retrieval of ontologies from repositories. The presentation of the ontologies in the web is separated from their internal representation.
- Customisation and extension of ontologies from repositories. Users can assemble a new ontology from a library of modules, as well as extend or restrict definitions from the library.
- Facilities for translating ontologies from repositories into typical application environments such as Prolog, CLIPS, LOOM, KIF, etc.
- Facilities for programmatic access to ontologies so that remote applications have reliable access to up-to-date term definitions. It has been defined a network protocol and an application program interface (API) to enable remote applications to use an Ontolingua Server to learn about the vocabulary of an ontology or about relations between terms.
- Support for distributed, collaborative development of consensus ontologies by means of a development environment with a rich set of features to support concurrent ontology development such as locking mechanisms and analysis of alternative definitions from various authors.

There is a considerable number of ontologies available in the Ontolingua server. Some of them are private whereas some others are publicly available. Among the latter, it is worth mentioning the Frame Ontology which is the conceptual basis for the Ontolingua translators. Translators of ontologies written in KIF using the frame ontology allow to work from a common source format and continue to use existing representation systems.

4 Comparative Analysis

In this section, and having reviewed the main features of the five former selected ontologies, we will analyse some of the key parameters which 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 hard task. The samples included in

[30] KIF is a language designed to be used in the knowledge interchange process among different computer systems. Typically, when a computer system reads a knowledge base in KIF, it converts the data into its own internal form. All computation is done using these internal forms. When the computer system needs to communicate with another computer system, it maps its internal data structures into KIF.

Main features of this language are the following:

- It has a declarative semantics. It is possible to understand the meaning of expressions in the language without appealing to an interpreter for manipulating those expressions.
- It is logically comprehensive (it provides the expression of arbitrary logical sentences in the first-order predicate calculus).
- Although it is not intended to be used within programs as a representation or communication language, it can be used for that purpose if so desired.
- Although it is not primarily intended as a language for interaction with humans, its readability facilitates the use in describing representation language semantics.
- There is a draft proposing KIF to the American National Standard (NCITS.T2/98-004). The full text is available at the following url: *http://logic.stanford.edu/kif/dpans.html*.

the previous section show many differences on the design and purpose of each ontology.

However, some characteristics are in fact comparable. In this sense, the elements reviewed in the comparative analysis are the following: availability, 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[31], which has allowed us to have access in depth to the ontology organisation.

4.1 Availability

By availability we mean access facilities in order to obtain the ontology from its creator via web or via formal agreements for using in our lab. In the case of Cyc, it belongs to a private enterprise. Their creators allow the access to a reduced part of information which is not really transparent nor easily reusable. In the case of EWN and SIMPLE (and specifically PAROLE), the access is less restrictive because the user can obtain information related to this resources via ELRA (*European Language Resources Association*)[32]. Finally, as for μKosmos and UMLS is concerned, IULA has been accessing to both of them. UMLS (see § 3.5) has been freely obtained and the μKosmos ontology design has been directly consulted from literature on the web and, later on, it has been browsed by means of an ontology management system which includes this ontology. Table 4-1 summarises the information given above.

Resource	Availability
Cyc	Publicly available subset
EuroWordNet	License (ELRA)
μKosmos	Through OntoTerm
SIMPLE	License (ELRA)
UMLS	License (institutional agreement)

Table 4-1. Ontologies availability indications.

4.2 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:

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

[31] We would like to thank Antonio Moreno (Universidad de Málaga) for all his indications and interest in order to facilitate us the access to OntoTerm.

[32] More information may be obtained through the following url: *http://www.icp.inpg.fr/ELRA/home.html*.

b) *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.[33]

c) *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.[34]

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

e) *UMLS*. The only tool included in the UMLS distribution is MetamorphoSys, a system that allows to customize and create subsets of the UMLS Metathesaurus in order to better meet the user's needs.[35]

4.3 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. A brief comment about these characteristics is indicated below:

a) *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, skolemisation, and some second-order features (e.g., quantification over predicates is allowed in some circumstances).

b) *EuroWordNet*: It describes concepts (called synsets) as a set of variants. There is 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[36] principles. Semantic information for each concept is inherited from its ancestors except for the cases where some parts of this information are redefined.

c) *μKosmos*: Concepts are described by their position in the ontology and by the indication of their properties and values.[37] 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.

d) *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).

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

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

[35] 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.

[36] Pustejovsky (1995).

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

e) *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.

One can set two different groups 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.

4.4 Application Field

As mentioned before, 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 latter consideration, it has to be mentioned that the general ontologies do not have all the domains equally developed. Probably, μKosmos has considerably developed those branches of the ontology concerned with the joint-venture domain. Also, as mentioned in §3.2.2, EWN has asymmetrically developed the different domains.

4.5 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.

4.6 Size, Granularity and Completeness

The size of all the resources analyzed is very different. Table 4-2 shows the global sizes for each resource in the different languages considered.

Resources	Ontology Size	Size for each language		
		English	Spanish	Catalan
Cyc	3,000	14,000	0	0
EWN		90,000	50,000	20,000
μKosmos[38]	4,800	0	0	0
SIMPLE		?	3,000	3,000
UMLS (2001 Edition)	134	800,000	30,000	0

Table 4-2. Analyzed resources: size comparison.

Together with the information included in Table 4-2, it should be taken into account that not all ontologies have the same granularity level in all domains[39]. Figure 11 shows the data included in μKosmos and EWN ontologies for the concept "body part". Both ontologies include the concept but the number of hyponyms is quite different: 1,639 concepts for EWN against 42 for μKosmos. This unbalance is probably due to the fact that EWN is a lexical-oriented general ontology that has been enlarged in the medical field while μKosmos is also a general domain ontology, but designed for supporting a KBMT system oriented to the economical domain. Besides, the enlargement criteria is quite different between EWN and μKosmos. In the first case, it is intended to provide maximum detail in the ontology and, for this reason, the number of concepts is enlarged as necessary. As for μKosmos, the enlargement criteria seems to be restricted to add new concepts only if necessary. UMLS must be dealt aside because it covers the medicine domain with a controlled number of concepts non-characterized in much detail.[40]

It is a well-known phenomenon that in WN, and therefore EWN, not all domains have the same granularity level. Some of them, such as medicine or botany, have been more deeply developed than others.

[38] There are several 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.

[39] Notice that in some cases this information is even not indicated.

[40] Taking the lexical unit *asthma* as an example, we see that EWN defines the following hyperonym chain: *asthma→respiratory disease→disease*. In contrast, UMLS relates this lexical unit to the semantic type *Disease or Syndrome*. This lexical unit is not included in μKosmos.

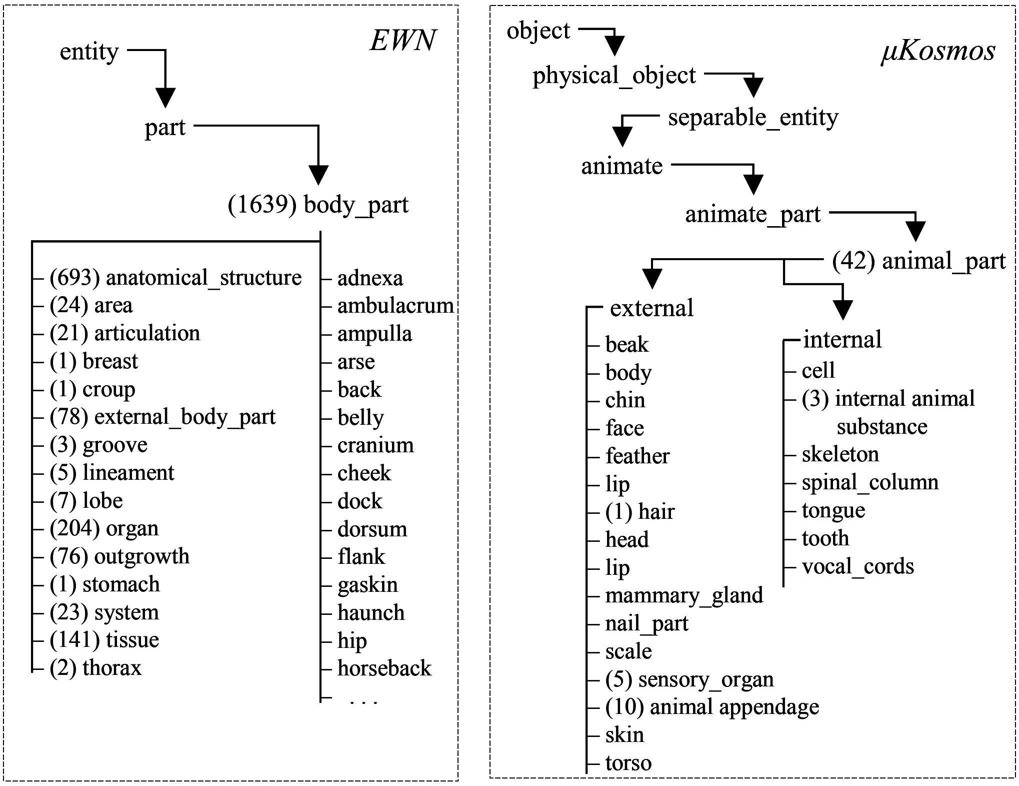


Figure 11. Completeness comparison example for an specific medical concept: μKosmos versus EWN[41].

5 New Trends in Ontologies

Internet has often been described as a vast electronic library which is expanding at a high rate. Most of that information is currently being represented using the Hypertext Markup Language (HTML), which is mainly designed to allow web developers to display information in a way that it becomes accessible to humans. Although HTML allows to visualize the information on a web browser, it provides limited capability to describe the information associated to a web page[42]. Consequently, most of the Internet resources are "machine readable" but not "machine understandable". A component of a web page that fully identifies it is the URI (Universal Resource Identifier), but again this identifier does not supply any universally understandable information about the resource itself. The key point is that «*missing is the part of the Web which contains information about information*» (W3C, 2000).

The World Wide Web Consortium (W3C) has developed the Extensible Markup Language (XML) which allows information to be more accurately described using tags. As an example, the word *Java* on a web site might represent a computer

[41] The "*is-a*" is the only relation represented in this figure. The number indicated next to some concepts shows the quantity of hyponyms.

[42] The HTML language provides a tag ('<meta>') to include information about the content of the web page, but just a small number of pages take profit of this tag.

language, an island or a coffee variety. The use of XML to provide metadata markup, for example regarding the unit *Java*, makes the meaning of the word unambiguous. XML has limited capacity to describe the relationships (schemas or ontologies) with respect to objects. This capability has been further expanded with the development of the Resource Description Framework (RDF).

RDF is part of the W3C Metadata Activity where it is defined as "a declarative language that provides a standard way for using XML to represent metadata in the form of statements about properties and relationships of items on the web". In other words, RDF uses XML to exchange machine-understandable descriptions of Web resources; such resources can be of any type, including XML and non-XML resources. RDF can be used in a variety of application areas, such as:

- Resource discovery: to provide better search engine capabilities.
- Cataloging: for describing the content and content relationships available at a particular Web site, page, or digital library.
- Intelligent software agents: to facilitate knowledge sharing and exchange.
- Content rating systems.
- Page/site description: describing collections of pages that represent a single logical "document".
- Describing intellectual property rights of Web pages.
- To express the privacy preferences of a user as well as the privacy policies of a Web site.
- Together with digital signatures, RDF will be essential for e-commerce, collaboration, and other applications.

The use of ontologies provides a very powerful way to describe objects and relationships among them. It is becoming commonly used in the web to describe taxonomies, ranging from large ones for representing huge web sites to smaller ones to categorize products for sale. In any case, the use of ontologies is increasingly widespread and its production is moving from AI laboratories to web publisher's desks. Due to this evolving situation, specific standards have emerged and several tools have been designed.

The Ontology Interchange Language (OIL) is a proposal for integrating ontologies into web standards. It is a web-based representation and inference layer for ontologies, which combines the widely used modeling primitives from frame-based languages with the formal semantics and reasoning services provided by description logics.

The DARPA Agent Markup Language (DAML) is being developed as an extension to XML and the Resource Description Framework (RDF). The latest release of the language (DAML+OIL) provides a rich set of constructs with which it is possible to create ontologies and to markup information so that it becomes machine readable and understandable.[43]

[43] More information about these emerging standards may be found at the following URL: *http://www.ontoknowledge.org/oil/* (OIL), *http://DAML.SemanticWeb.org/* (DAML), http://dublincore.org/ (DC). The full set of standards related to XML/RDF may be found at: *http://www.w3.org/TR/.*

The Dublin Core Metadata Initiative (DC) is another cross-disciplinary international effort to develop mechanisms for the discovery-oriented description of diverse resources in an interdisciplinary environment. DC uses XML/RDF to provide the structure in order to express information without ambiguity.

The Platform for Internet Content Selection (PICS) is a system for associating metadata (PICS “labels") with Internet content. PICS provides a mechanism whereby independent groups can develop metadata vocabularies without naming conflict.

At the same time that new standards and/or proposals are published, there is also an emerging large quantity of related resources. They may take the form of complete tools for creating ontologies or any sort of software modules for processing XML/RDF documents (RDF Parsers and Compilers, RDF database server, visualisation systems, RDF Schema editor, etc.). In Duineld *et al.* (2000) there is an evaluation of different tools for supporting the construction of ontologies. For showing how rapidly this framework is evolving, it should be noted that some of the tools examined in such paper have already been replaced for newer and much more powerful versions. Having reviewed the three different tools that we consider the most relevant ones: Protégé-2000, OilEd and Ontoprise, we will now describe main utilities concerning the latest version of Protégé-2000.

5.1 Protégé-2000

Protégé-2000 is a knowledge-base design and a knowledge-acquisition tool developed by the Knowledge Modelling Group at Stanford University. It takes profit of the large experience of this group in building tools for knowledge-base construction. This tool is available as free software under the open-source Mozilla Public License[44]. It provides an integrated knowledge-base editing environment and an extensible architecture. It executes on any computer running the Java programming language.

This tool allows the user to construct a domain ontology; to customize knowledge acquisition forms; and to enter domain knowledge. Moreover, it is conceived as a platform which can be extended with graphical widgets for tables, diagrams, animation components to access other knowledge-based system embedded applications. And it also becomes a library which other applications can use to access and display knowledge bases.

One of the main objectives in the design of Protégé-2000 was the capacity to reuse already existent knowledge bases. This goal was achieved making its knowledge model compatible with the Open Knowledge Base Connectivity (OKBC) protocol[45]. As a result, users may import ontologies from other OKBC knowledge

[44] Protègè-2000 and related documentation may be obtained at the following url: *http://protege.stanford.edu/*.

[45] This protocol provides a set of operations for a generic interface to a underlying knowledge representation systems.

servers. Protégé-2000 restrict the use of this protocol but at the same time it extends some of its features[46].

A main feature in Protégé-2000 is the use of forms to acquire instance data. Protégé-2000 generates these forms automatically based on the class definitions and the users can then custom-tailor the forms, if necessary. Protégé-2000 uses metaclasses widely, it implements the internal structure of its own knowledge model in a metaclass architecture. A metaclass is a template that is used to define new classes in an ontology. Therefore, it is possible to customize the forms for specifying classes and slots in the same way that it customizes forms for acquiring instances. The metaclass architecture in Protégé-2000 along with its component based approach also enables developers to use this tool as an editor for knowledge representation systems using other knowledge models. This flexibility allows to define RDF as a new metaclass architecture in Protégé-2000. The knowledge model underlying RDF (RDF Schema) is different from the Protégé-2000 knowledge model. However, it is possible to define the main elements of the RDF knowledge model by defining the metaclasses that will add RDF-specific features to the templates used to create new classes and slots. This definition enables the use of this tool for the creation and editing of RDF documents.

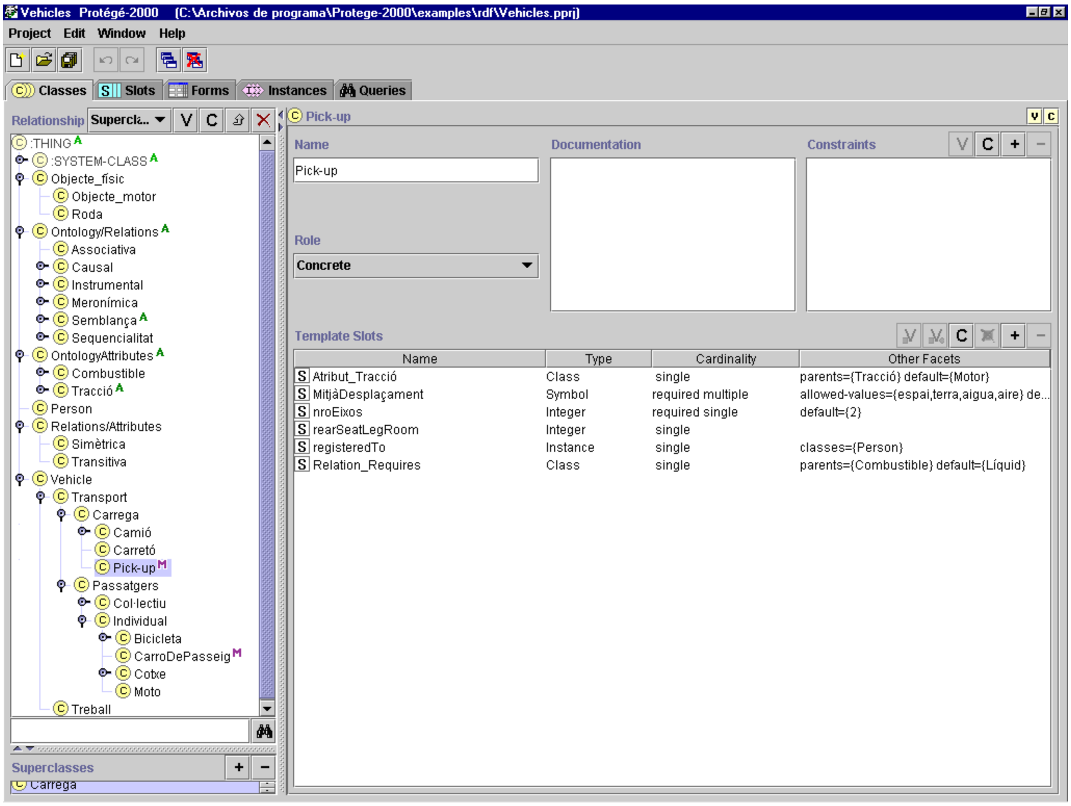
The knowledge model of Protégé-2000 is frame-based, that is, frames are the main building block of a knowledge base. An ontology consists of classes, slots, facets, and axioms. Classes are concepts in the domain of discourse. Slots are frames that describe properties or attributes of classes. Facets describe properties of slots (cardinality, restrictions on the value type, etc.). Axioms specify additional constraints. A Protégé-2000 knowledge base includes both the ontology and individual instances of classes with specific values for slots. Classes constitute a taxonomic hierarchy. Then, if a class A is a subclass of a class B, every instance of A is also an instance of B.

Protégé-2000 accesses all the components of the ontology through a uniform graphical user interface whose top-level consists of overlapping tabs for compact presentation of the parts and for convenient co-editing between them. This “tabbed” top-level design permits an integration of (1) the modeling of an ontology of classes describing a particular subject; (2) the creation of a knowledge-acquisition tool for collecting knowledge; (3) the entering of specific instances of data and creation of a knowledge base, and (4) the execution of applications. The ontology defines the set of concepts and their relationships. The knowledge-acquisition tool is designed to be domain-specific, allowing domain experts to easily and naturally enter their knowledge of the area.

This tool assumes that building and maintaining knowledge bases are expensive processes. For this reason, it is designed to allow developers to reuse already existent domain ontologies and problem-solving methods, thereby shortening the time needed for development and program maintenance. Several applications can use the same

[46] See Noy (2000) for the details of the knowledge model used in Protégé-2000.

domain ontology to solve different problems, and the same problem-solving method can be used with different ontologies.



Name: Pick-up

Role: Concrete

Name	Type	Cardinality	Other Facets
Atribut_Tracció	Class	single	parents={Tracció} default={Motor}
MitjàDesplaçament	Symbol	required multiple	allowed-values={espai,terra,aigua,aire} de...
nroEixos	Integer	required single	default={2}
rearSeatLegRoom	Integer	single	
registeredTo	Instance	single	classes={Person}
Relation_Requires	Class	single	parents={Combustible} default={Líquid}

Figure 12. The representation of an ontology in Protégé-2000.

Protégé-2000 provides a nice graphical user interface which should be easy to handle by any one (see Figure 16). The ontology structure is made up just like a hierarchical directory structure giving a good overview of the ontology. Classes and subclasses can be selected, added and edited using the mouse buttons and/or clearly identified buttons (on the left side of the window). The slots corresponding to the selected class are fully described on the right side of the window. Again, using the mouse or the buttons the user can easily edit already defined slots as well as to add new ones.

6 Final Remarks

Having reviewed the five major ontologies and taking into account that in IULA's ongoing project we will need an ontology in the near future, we will point out the most important parameters we have to bear in mind in order to determine if any of the ontologies described could be applied to our purposes. For this reason, we will concentrate in the following two sections on the review of the main requirements about each ontology and its management tools.

6.1 Ontology

Aiming at a general ontology that allows enlargement, we have 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 to 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 semantic information to a dictionary and it cannot be considered as an ontology itself. However, UMLS could be an important source to enlarge the selected ontology for our project and SIMPLE would be useful in order to complete and refine linguistic information for NLP.

Both remaining ontologies, μKosmos and EWN, are general domain ontologies that satisfy the basic requirements of IULA's ongoing project framework. In spite of this, there are important differences between such resources. Table 6-1 indicates the most salient parameters of both resources.

Parameter	μKosmos (Ontoterm)	EWN
Type	Conceptual	Lexical
Completeness	Medium	High
Medicine coverage	Low	High
Implementation OS	Windows	Unix

Table 6-1. Main characteristics of μKosmos and EWN.

6.2 Management Tool

Before describing main utilities of the management tools available for μKosmos and EWN, we will highlight some requirements to be fulfilled by a management tool adequate to our research purposes.

In our opinion, an ontology management tool should include some basic management facilities:

a) To enlarge the hierarchy.
b) To assign predefined attributes to nodes.
c) To add pre-established and fully-organised relations among concepts in the hierarchy.
d) To block repetition of concepts and/or relations and to avoid circularities inside the system.
e) To care about the mechanism of inheritance (monotonic / non-monotonic inheritance, multiple inheritance control, among others).
f) To have a mechanism to connect the ontology with the possible existing dictionaries.
g) To export the ontology contents to different formats.
h) To be a user friendly application.

Thus, it is absolutely necessary to maintain the coherence of the whole system. It means that the addition of new attributes and relations should be protected from free and uncontrolled decisions. All attributes must have a particular space, where corresponding values must be pre-defined (for discrete attributes) or range limited (for numeric attributes). As for relations, they must be completely defined and only assigned to concepts after having passed a validation test.

After introducing these general desiderata on the requirements of a management tool, we briefly summarise main similarities and differences of the selected management tools according to the above listed parameters. As it has been mentioned in section 6.1, we consider µKosmos and EWN as the most appropriate ontologies for our project. In this final considerations we analyse, in accordance with the parameters mentioned above, their corresponding management tools and Protégé-2000.

Parameters	OntoTerm	Protégé-2000	EWN
Enlargement	✓	✓	✓
Attribute assignation	Free	Restricted	—
Conceptual relations	Free	Restricted	Closed
Concept repetition	✕	✕	✕
Inheritance	✓	✓	✓
Ontology-Lexicon	✓	✕	—
Export	HTML	XML, JDBC...	✕
Friendliness	✓	✓	No

Table 6-2. Management tools: comparative analysis.

As already seen, µKosmos is a concept-oriented ontology and, for this reason, in its ontology manager it is clearly seen a unidirectional path from the ontology specification towards the lexicon. Conversely, as far as EWN is concerned, there is a tight relation between the lexicon and the ontology and, for each language, it is possible to develop a particular and different ontology. Protégé-2000, as a facility to build new ontologies, does not foresee any associated lexicon.

Talking about the possibility to add new concepts and relations, it has to be said that OntoTerm is not very restrictive and its design criteria can derive into coherence problems. Regarding EWN, the way to introduce new information is exhaustively controlled. Thus, EWN enlargement tool is much more restrictive than OntoTerm as for synsets and relations additions are concerned.

As for information about relations, it is worth mentioning that both ontology managers only show the "*is-a*" relation. That is, OntoTerm provides a conceptual tree indicating this type of relation among all concepts contained in the ontology up to the top level. In the case of EWN, it is more difficult to see all the ontology because the system provides the visualisation of the hyperonimy path of a particular lexical unit. Moreover, OntoTerm appears as a much more user friendly tool than EWN.

Having summarised the main differences between these two management tools, we have to conclude that both tools need further development in order to fulfill our research project requirements and to be used in the Genome Knowledge Base construction.

6.3 The Genome Project: Towards an Ontology and its Management Tool Selection

In the Genome project framework, we aim to create an ontology which will be further developed in some areas. The ontology will be unavoidably tied to its management tool. So, the final selection must take into consideration both aspects: facility and adequacy of the ontology and an appropriate management environment for its development.

Following the μKosmos design on the basis of two separated modules, the ontology and the lexicon, we believe that this approach will fulfill our project requirement. However, it will be an ontology built upon the just mentioned criteria and expanded with information drawn from some other ontologies, such as UMLS, EWN. UMLS will be reused in order to enlarge the new ontology in the medical branch. EWN will be considered as a linguistic information source, such as SIMPLE. EWN will 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.

OntoTerm, as the management tool available and associated to μKosmos, may be considered as an appropriate tool. However, as seen in Table 6-2, OntoTerm does not foresee some of the requirements specified for the desired management tool. For this reason, we propose as an ideal management tool to integrate in OntoTerm some of the facilities offered by Protégé. Particularly, we would aim a resulting tool treating attributes and relations in a similar way to Protégé and some additional management features, i. e., only a restricted number of users should be licensed to define new relations, attributes and its corresponding values.

Finally, we are aware that much work must be done until the factual creation of the Genome Project ontology and the adaptation of a completely satisfying management tool. However, we believe that the objectives of the paper have been achieved by setting the foundations for this project to become a reality.

7 Bibliography

Bateman J.; Magnini B. and J. Rinaldi (1994) "The Generalized Upper Model". *Actas de ECAI.*

Bateman J.; Kasper, R. T.; Moore, J. and R. Whitney (1990) "A General Organization of Knowledge for Natural Language Processing: the Penman Upper Model". Technical Report, USC/ISI, Marina del Rey, California.

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

Benjamins, R.; Fensel, D.; Deckes and A. Gómez-Pérez (1999) "(KA)^2 Building ontologies for the Internet: a mid term report". International Journal of Human Computer Studies.

Borst, P.; Akkermans, H. and J. Top (1996) "Engineering ontologies". Report INF-96-09. University of Twente.

Duineveld, A. J.; R. Stoter; M.R. Weiden; B. Kenepa and V. R. Benjamins (2000) "Wonder tools? A comparative study of ontological engineering tools". International Journal of Human-Computer Studies. Academic Press, 52, pag. 1111-1133.

EAGLES (1999) "Preliminary Recommendations on Lexical Semantic Encoding. Final Report". The EAGLES Lexicon Interest Group. EAGLES Document: LE3-4244: *http://www.ilc.pi.cnr.it/EAGLES96/browse.html#wg2.*

Farquhar, A.; Fikes, R. and J. Rice (1996) "The Ontolingua Server: a Tool for Collaborative Ontology Construction". Available at: *http://www-ksl-svc.stanford.edu:5915/doc/project-papers.html.*

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

Gruber, T. R. and G. Olsen (1994) "An ontology for engineering mathematics". Fourth International Conference on Principles of Knowledge Representation. Morgan Kaufmann.

Gruber, T. R. (1992) "Ontolingua: A mechanism to Support Portable Ontologies". Report KSL 91-66, Stanford University.

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

LE2-4003a (1998) Deliverable D027: "EuroWordNet Subset2 for Dutch, Spanish and Italian". Reachable at *http://www.ley.una.ln/~ewn/docs/.*

LE2-4003b (1998) Deliverable D017: "The EuroWordNet Base Concepts and Top Ontology". Reachable at *http://www.ley.una.ln/~ewn/docs/.*

LE2-4003c (1997) Deliverable D005: "Definition of the links and subsets for nouns of the EuroWordNet project". Reachable at *http://www.ley.una.ln/~ewn/docs/.*

LE3-4244 (1999) "Preliminary Recommendations on Lexical Semantic Encoding. Final Report". The EAGLES Lexicon Interest Group. Reachable at *http://www.ilc.pi.cnr.it/EAGLES96/browse.html#wg2*.

Maynard, D. (1999) "Term recognition using combined knowledge sources". PhD Thesis. Manchester Metropolitan University. Faculty of Science and Engineering.

Miller G. A.; Beckwith R.; Fellbaum, C.; Gross, D. and K. Miller (1993) «Introduction to WordNet: An on-line lexical database».

Moreno A. and C. Pérez (2000) "Reusing the MikroKosmos Ontology for Concept-Based Multilingual Terminology Databases". Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC-2000). Athens, May 31th-June 2nd, pag. 1061-1067.

Moreno A. (1999) "An introduction to OntoTerm". IULA's Workshop, Junio 1999.

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

Noy, N. F.; Fergerson, R. W. and M. A. Musen (2000) "Knowledge-Acquisition Interfaces for Domain Experts: An Empirical Evaluation of Protege-2000". Proceedings of the Twelfth International Conference on Software Engineering and Knowledge Engineering (SEKE2000), Chicago, IL.

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

Yokoi T. (1995) "The Impact of the EDR Electronic Dictionary on Very Large Knowledge Bases", De Mars N. (ed.) Towards Very Large Knowledge Bases. IOS Press.

Computers and the Humanities (1998), Special Issue on EuroWordNet, Vol. 32, Kluwer Academic Publishers, Dordrecht.

Proc. COLING/ACL'98 Workshop on Usage of WordNet for Natural Language Processing.

PhysSys [http://www.cs.utwente.nl/memoranda/r/memo.inf.96/memoranda.html].

Soler, C. (1997) *Desajustes léxicos nominales y su representación en una base de conocimientos léxicos. Valores semánticos del adjetivo*. PhD dissertation. Universitat Politècnica de Catalunya. Barcelona.

Vossen, P. (ed.) (1999) "EuroWordNet General Document". University of Amsterdam. Documento obtenido en: *http://www.hum.uva.nl/~ewn*.

Yu, H.; Friedman, C; Rhzetsky, A. and P. Kra (1999) "Representing Genomic Knowledge in the UMLS Semantic Network". AMIA Conference.