An Environment for Ontology Design and Enrichment from Texts

Michel Simonet[1], Delphine Bernhard[1], Gayo Diallo[1],
Séverine Gedzelman[1], Radja Messai[1] and Rémi Patriarche[1]

[1] TIMC-IMAG,
Institut de l'Ingénierie et de l'Information de Santé,
Faculté de Médecine,
38700 La Tronche, France
{Michel.Simonet, Delphine.Bernhard, Gayo.Diallo,
Séverine.Gedzelman, Radja.Messai, Remi.Patriarche}@imag.fr
http://www-timc.imag.fr

Abstract. Ontology design is a difficult task for which there is no agreed upon methodology. Texts of the domain can provide the words and terms of the domain, and support the abstraction of concepts and relationships which constitute the skeleton of the ontology. An environment to help the ontology designer has been built. Its main components are a term extractor, a concordancer and an ontology editor specially designed for multilingual treatment.

1 Introduction

Ontology design has become an important issue in knowledge management. However, here is not yet an agreed upon methodology to build an ontology [6][7][8][9]. It is most often the result of an expert's work or that of a group of experts, and it is recognized a difficult and time-consuming task.

The identification of the concepts of the domain is at the heart of ontology design. A concept is defined by its name, its definition and its relationships with other concepts. According to the intended usage of the ontology, it may also be necessary to develop the terminology associated with a concept. For example, if the ontology is designed for Information Retrieval or Semantic Annotation, it is important that it provides all the terms which refer to the concepts in texts or in users' queries.

Even if the intended usage of the ontology is not directly related to texts, the texts of the domain can be a valuable source of information and data to help building the ontology. It is a long time since the French TIA community (Terminologie et Intelligence Artificielle) has noticed the similarities between ontologies and terminologies and promotes an approach based on text analysis to build an ontology [1]. We have taken the same approach for two project described in Section 2. In order to build or enrich ontologies using text corpora we have designed a specific environment, named GNOMIC. The components of this environment are detailed in Section 3.

1 Context

During the now completed European project INFACE we worked on the enrichment of an ontology of breast cancer for health professionals using text corpora [12]. This work is currently being complemented by the building of an ontology of breast cancer for patients, using the same methodology. For each of these ontologies of breast cancer we have built a corpus of around 500 texts. This corpus mainly contains scientific texts for the former and web pages for the latter. For these two projects we have been using tools to assist the ontology engineer in his/her design task: a term extractor to isolate terms (repeated groups of consecutive words) and identify terms which are not in the ontology, a concordancer to visualise terms in their context and an ontology editor which supports editing multilingual vocabulary.

In the context of the European project NOESIS (www.noesis-eu.org) we have again been confronted to the task of building an ontology with a rich vocabulary as its purpose is to support concept-based information retrieval and semantic annotation of texts. As the domain is that of cardio-vascular diseases, we decided to start from an existing medical classification, the MeSH (Medical Subject Headings) thesaurus, which is currently used to index biomedical literature (http://www.nlm.nih.gov/mesh/). We have used a bilingual version of the MeSH which was provided by the French institute INSERM.

After putting the 690 concepts of the cardio-vascular subset into an OWL format we have proceeded to a first enrichment phase with terms from the UMLS (Unified Medical Language System) metathesaurus (http://umlsks.nlm.nih.gov) for these concepts.

```
<owl:Class rdf:ID='M0001551'>
  <skos:prefLabel xml:lang='en'>Aortic Rupture</skos:prefLabel>
  <skos:altLabel xml:lang='en'>Aortic Ruptures</skos:altLabel>
  <skos:altLabel xml:lang='en'>Rupture Aortic</skos:altLabel>
  <skos:altLabel xml:lang='en'>Ruptures Aortic</skos:altLabel>
                             …
  <skos:prefLabel xml:lang='fr'>Rupture de l'aorte</skos:prefLabel>
  <skos:altLabel xml:lang='fr'>Rupture aortique</skos:altLabel>
  <skos:altLabel xml:lang='fr'>Rupture aorte</skos:altLabel>
  <rdfs:comment>do not coord with rupture spontaneous unless    par-
ticularly discussed and then only nim </rdfs:comment>
  <rdfs:isDefinedBy>Tearing of aortic tissue. It may be rupture of an
aneurysm or it may be due to trauma.  </rdfs:isDefinedBy>
  <rdfs:subClassOf rdf:resource='#M0026605'/>
  <rdfs:subClassOf rdf:resource='#M0001545'/>
</owl:Class>
```

Figure 1. Example concept extracted from the MeSH and the UMLS.

The initial vocabulary provided by the MeSH for the cardio-vascular subset of 690 concepts was 2070 English terms and 966 French terms. The first enrichment phase through UMLS increased the English vocabulary by 10000 new terms (of which 2500 are true synonyms while the others are lexical variants) and 13000 terms for other languages (French, Italian, Spanish and German); the Greek vocabulary, absent from

UMLS, was provided by a Greek version of the MeSH and was limited to 690 terms (one term per concept).

In the next phase of the enrichment process we will proceed from texts. For this purpose a corpus of around 500 scientific texts of the CV domain has been constituted from www.infobiomed.org. As this enrichment process has to be performed by doctors which are neither IT nor ontology specialists, we have built an environment to help them select terms and add them to the ontology.

2 The GNOMIC Environment

GNOMIC (Grenoble Noesis Ontology Management Integrating Corpora) has been designed to help ontology engineering from texts in different languages. Its input data are a corpus of texts of the domain and an ontology (which may be empty in the case of initial design). GNOMIC aims at providing a friendly environment to the user who is neither an IT nor an ontology specialist in order to help him/her build or enrich an ontology. The methodology which has guided its design consists in extracting the terms from the texts in the corpus, identifying those absent in the ontology, enabling the designer to visualize them in their context (concordancer) and editing the ontology. Ontology edition means adding, removing or modifying concepts, relationships, vocabulary and definitions. We now present the main components of the GNOMIC environment.

2.1 Term Extractor

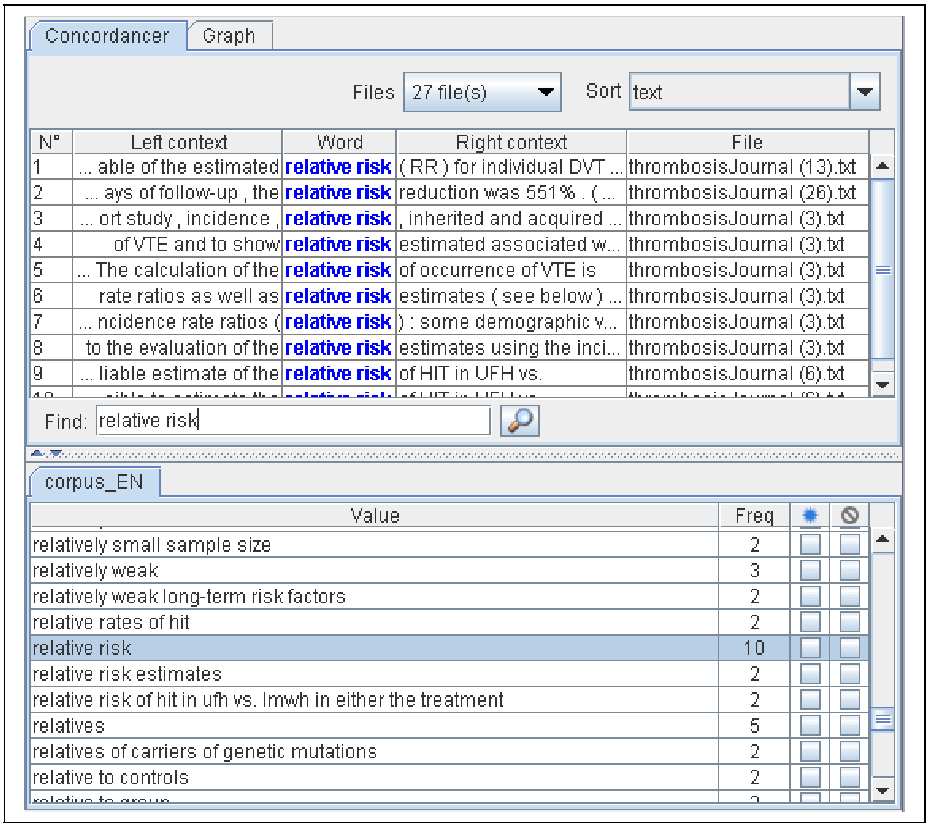
Candidate terms are extracted using the method described by [3] based on repeated segments. When terms have been extracted, they are displayed as a list (see Fig. 2, bottom) so that the user may validate the terms he wants to keep for further association with a concept, and reject those terms he considers not relevant for the domain under consideration.

Terms are either simple (unique words) or complex (sequence of words). Relevant word sequences are extracted using a list of word sequence delimiters (like determiners and prepositions). These delimiters cannot be found either at the beginning or at the end of a term. Moreover, terms cannot contain conjunctions like *and*. Further filtering is done to eliminate terms included in a larger term which occurs in the corpus with the same frequency or nearly the same frequency. Despite automatic filtering, the resulting list is still rather noisy. It is therefore necessary to manually select relevant terms. Some terms may be eliminated straightforwardly, others need to be checked. In order to help users perform this task, GNOMIC integrates a concordancer.

2.2 Concordancer

Concordancers have proved useful for lexicography and lexical semantic studies. Both share the same aim with knowledge resources building: unravelling meaning. Indeed, they reveal uses of language by displaying words in the midst of their linguistic context and this applies even to rare and seldom used words. Contextual collocates provide observable evidence of word meaning.

The concordancer lists all the occurrences in the corpus of the term under study (Fig. 2 shows occurrences of the term *relative risk*). These occurrences may be sorted using words found in the near context of the search term as a sort criterion. Concordances are especially useful to check the meaning of a term.



Files: 27 file(s)
Sort: text

N°	Left context	Word	Right context	File
1	... able of the estimated	**relative risk**	(RR) for individual DVT ...	thrombosisJournal (13).txt
2	... ays of follow-up , the	**relative risk**	reduction was 551 % . (...	thrombosisJournal (26).txt
3	... ort study , incidence ,	**relative risk**	, inherited and acquired ...	thrombosisJournal (3).txt
4	of VTE and to show	**relative risk**	estimated associated w...	thrombosisJournal (3).txt
5	... The calculation of the	**relative risk**	of occurrence of VTE is	thrombosisJournal (3).txt
6	rate ratios as well as	**relative risk**	estimates (see below) ...	thrombosisJournal (3).txt
7	... ncidence rate ratios (	**relative risk**	) : some demographic v...	thrombosisJournal (3).txt
8	to the evaluation of the	**relative risk**	estimates using the inci...	thrombosisJournal (3).txt
9	... liable estimate of the	**relative risk**	of HIT in UFH vs.	thrombosisJournal (6).txt

Find: relative risk

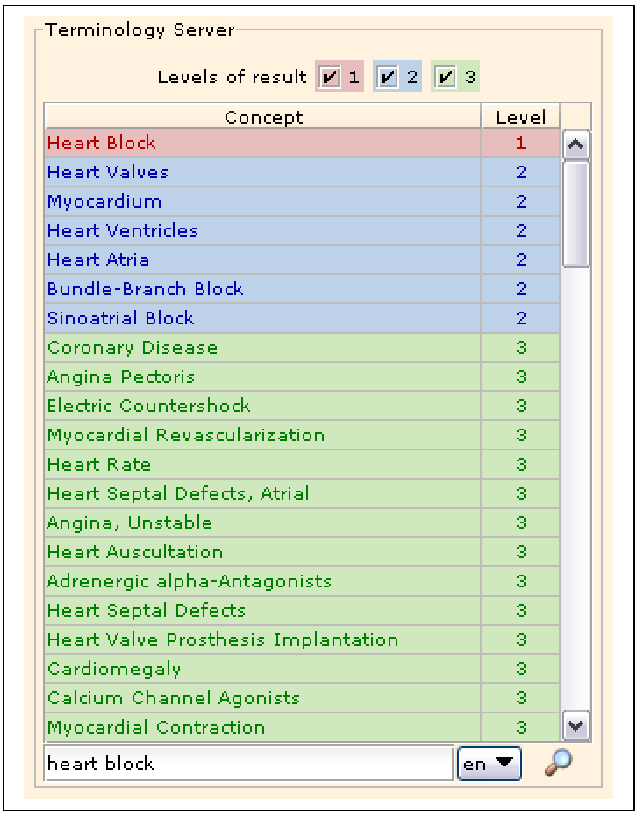
Value	Freq	✸	⊘
relatively small sample size	2	□	□
relatively weak	3	□	□
relatively weak long-term risk factors	2	□	□
relative rates of hit	2	□	□
relative risk	10	□	□
relative risk estimates	2	□	□
relative risk of hit in ufh vs. lmwh in either the treatment	2	□	□
relatives	5	□	□
relatives of carriers of genetic mutations	2	□	□
relative to controls	2	□	□

Figure 2: List of extracted terms and concordancer User Interface

During the work on the ontology of Breast cancer for professionals, we were confronted to the term « cancer of Canada » which had been extracted from the texts under consideration. In order to decide whether this was a term relevant to the domain of cancer, we used the concordancer, which displayed the various uses of this term. It was immediately shown that this term appeared only in the expression « Institute of cancer of Canada ». Therefore the term « cancer of Canada » was part of the noise produced by the term extractor and was straightforwardly eliminated.

2.3 Terminology Server

When a term has been validated, it has to be associated with the concept it represents. The terminology server automatically looks for links between a given term and the concepts of the ontology. It calculates a similarity function, using the Levenhstein Distance, between the stemmed term and those associated with the concepts. The result is a ranked list of concepts, as shown in Fig. 3 for the term *heart block*. The result levels depend on the search strategy used to retrieve the concept represented by a term. These strategies are detailed in Table 1.



Terminology Server

Levels of result ☑ 1 ☑ 2 ☑ 3

Concept	Level
Heart Block	1
Heart Valves	2
Myocardium	2
Heart Ventricles	2
Heart Atria	2
Bundle-Branch Block	2
Sinoatrial Block	2
Coronary Disease	3
Angina Pectoris	3
Electric Countershock	3
Myocardial Revascularization	3
Heart Rate	3
Heart Septal Defects, Atrial	3
Angina, Unstable	3
Heart Auscultation	3
Adrenergic alpha-Antagonists	3
Heart Septal Defects	3
Heart Valve Prosthesis Implantation	3
Cardiomegaly	3
Calcium Channel Agonists	3
Myocardial Contraction	3

heart block

en

Figure 3: Results of Terminology Server for the query “heart block”

Examples in Tab. 1 are given for the query term "heart block". The level 1 result is the concept whose preferred term is "Heart Block", which matches perfectly the query term. However, if no perfect match is found among the alternative terms representing a concept, similar terms, whose Levenshtein Distance with the query term is lower than a given threshold, are considered as accurate results. By using the Levenshtein Distance it is possible to match lexical variations of the query term like for instance "heart blocks".

Table 1. Search strategies used by the terminological server to relate a term to a concept.

	Query Term	**Result**	**Search strategy**
Level 1	heart block	Heart Block	Identity or strong similarity
Level 2	heart block	Heart Valves Myocardium Heart Ventricles Heart Atria Bundle-Branch Block Sinoatrial Block	Inclusion of all the words of the queried term in one of the alternative terms of the concept
Level 3	heart block	Coronary Disease Angina Pectoris Electric Countershock Myocardial Revascularization Heart Rate Heart Septal Defects, Atrial Angina, Unstable Heart Auscultation Adrenergic alpha-Antagonists Heart Septal Defects Heart Valve Prosthesis Implantation Cardiomegaly Calcium Channel Agonists Myocardial Contraction **...**	Inclusion of at least one word of queried term in one of the alternative terms of the concept

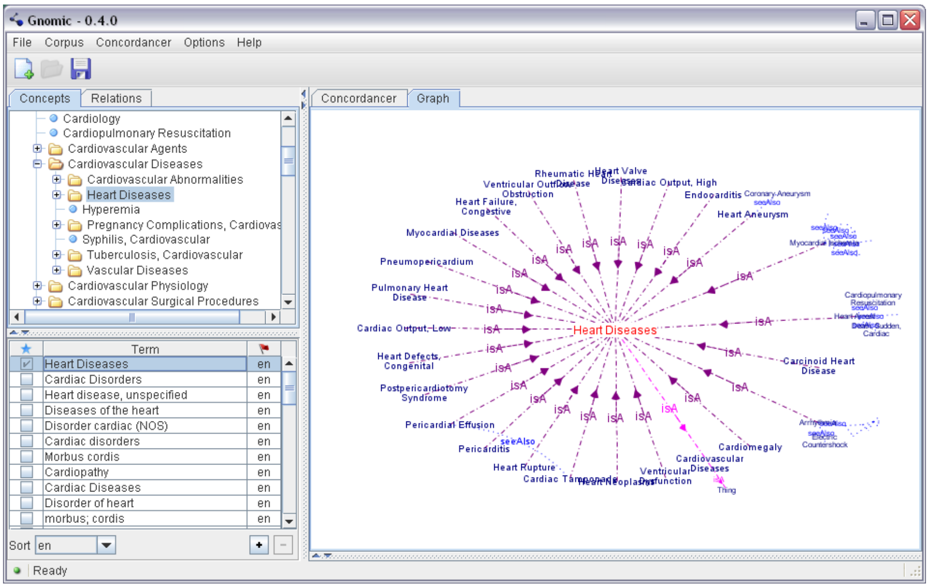
Level 2 results list the preferred terms for concepts which have at least one representing term encompassing all of the words in the query term. For instance, one of the representing terms for the concept "Bundle-Branch Block" is "Heart block bundle branch". "Heart block bundle branch" contains both the words "heart" and "block" which form the query term "heart block". It is therefore listed in the match proposals for level 2.

Level 3 results list the preferred terms for concepts which have at least one representing term containing one of the words in the query term. For instance, the terms "heart block" and "Heart Rate" share the word "heart". "Heart Rate" is therefore listed as level 3 result.

2.4 Ontology editor

We have developed our own ontology editor because the available ones, including the most popular, Protégé, did not provide some features we considered important:

- Ontology model: relationships between concepts should be treated symmetrically, while in standard editors a relation is considered as an attribute of a class representing the concept, thus losing the natural symmetry between the two concepts of the relation. We considered this feature essential as we use an ontological model for database conceptual modeling and we have found that using an "object" representation of an ontology during this phase of the modeling process (as is done in UML) induced untimely representation choices which might later prove to be erroneous.
- Multilingualism: as an ontology is meant to represent a consensus of a community on some domain and is to be used as a communication means, the vocabulary associated with the concepts plays an essential role, which is not always considered. When different languages are considered and a variety of synonyms for a concept, we need a preferred term in each language, a feature which is not supported by Protégé. To do so we used the SKOS library, which can be embedded in OWL [13].
- Ergonomy: as this environment is to be used by non-IT specialists, the user interface has to be strictly adapted to their needs. Again, Protégé proved to be too complex for this category of users. A hyperbolic presentation completes the classical tree-like presentation and makes possible the visualization of concepts and relations (one can select the relations to represent) to a depth which can be parametrized.



- Cardiology
- Cardiopulmonary Resuscitation
- ⊞ Cardiovascular Agents
- ⊟ Cardiovascular Diseases
  - ⊞ Cardiovascular Abnormalities
  - ⊞ Heart Diseases
  - Hyperemia
  - ⊞ Pregnancy Complications, Cardiovas
  - Syphilis, Cardiovascular
  - ⊞ Tuberculosis, Cardiovascular
  - ⊞ Vascular Diseases
- ⊞ Cardiovascular Physiology
- ⊞ Cardiovascular Surgical Procedures

★	Term	
☑	Heart Diseases	en
□	Cardiac Disorders	en
□	Heart disease, unspecified	en
□	Diseases of the heart	en
□	Disorder cardiac (NOS)	en
□	Cardiac disorders	en
□	Morbus cordis	en
□	Cardiopathy	en
□	Cardiac Diseases	en
□	Disorder of heart	en
□	morbus; cordis	en

Sort en

Ready

Figure 4: Ontology graphical visualizations with representative terms for the concept Heart Diseases

Fig. 4 shows the ontology of the Cardio-Vascular domain in tree form (upper left), the vocabulary associated with the *Heart Diseases* concept (bottom left) and the hyperbolic presentation of the concept with its immediate environment (right). The list of candidate terms and the terminology server (see Fig 4) are hidden.

2.5 Technical characteristics

The GNOMIC environment has been developed in Java. The ontology is represented in OWL and managed through the JENA API (http://jena.sourceforge.net). Jena [2] is a set of tools (API) developed within the framework of the HP Labs Semantic Web Program project, making possible to manipulate ontologies and to apply inference mechanisms. Jena provides integrated implementations of the W3C Semantic Web Recommendations, centred on RDF graph. It includes support for RDFS and OWL, including advanced OWL Full support.

We have used the HyperGraph Java source code (http://hypergraph.sourceforge.net) to visualise the ontology as a hyperbolic tree. We had to slightly modify the sources in order to link the hyperbolic graphical objects with the Jena model objects. The advantage of the hyperbolic view is that users can understand the ontology structure by seeing the relational cross-links between the vertices ; however this needs to be supplemented by a tree view explorer when there are many children. We exploit the advantages of both views in different circumstances by allowing them to complement one another. Each view allows the user to change the current selection by clicking on vertices or tree nodes.

3 Conclusions and perspectives

The project is still under development and will be used at the end of 2005 by NOESIS medical users to proceed to the enrichment of the ontology from a corpus of English texts in the cardio-vascular domain. The set of English terms will then be translated into the five other languages of the project (French, Italian, Spanish, German and Greek). A corpus of French texts on cardio-vascular diseases will also be constituted and the French terms tested against these texts so as to validate the translation process.

We do not take part in the debate about whether the vocabulary associated with a concept is part of the ontology or not. In certain situations, such as concept-based information retrieval, concepts need to be associated with terms in natural language, all the more so when the indexing and the search must be language-independent.

Other ongoing work related to ontology building and enrichment concerns the discovery of semantic relationships based on morphological and distributional similarities.

Acknowledgements

This work has been partly supported by the European Commission (NOESIS project, IST-2002-507960).

We thank Claudie Hasenfuss from INSERM for her support in providing the bilingual French version of the MeSH thesaurus.

References

1. N. Aussenac-Gilles, B. Biebow, S. Szulman, “Revisiting ontology design: a methodology based on corpus analysis”, in *European Workshop on Knowledge Acquisition, Modeling and Management*, pp 172-188, Oct. 02-06, 2000.
2. J. J. Carroll, I. Dickinson, C. Dollin, D. Reynolds, A. Seaborne, K. Wilkinson, “Jena: implementing the semantic web recommendations”, in Proceedings of the 13th International World Wide Web Conference, pp. 74 - 83, New York, 2004.
3. P. Frath, R. Oueslati, F. Rousselot. “Identification de relations sémantiques par repérage et analyse de cooccurrences de signes linguistiques”, in *Ingénierie des Connaissances. Evolutions récentes et nouveaux défis*, Eds J. Charlet, M. Zacklad, G. Kassel, D. Bourigault, Paris, Eyrolles, 2000.
4. S. Gedzelman, M. Simonet, D. Bernhard, G. Diallo. “Building an ontology of Cardio-Vascular diseases for Concept-Based Information Retrieval”, in *Computers in Cardiology*, Lyon, France, Sept. 2005.
5. T. R. Gruber. A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2):199-220, 1993.
6. M. Grüninger, M. S. Fox, “Methodology for the design and evaluation of ontologies”, in *IJCAI workshop on Basic Ontological Issues in Knowledge Sharing*, Montreal, Quebec, Canada, 1995.
7. N. Guarino, “Some ontological principles for designing upper level lexical resources”, in *Procs of the First International Conference on Lexical Resources and Evaluation*, Granada, Spain, 28-30 May 1998.
8. N Guarino,. (ed.), Formal Ontology in Information Systems, IOS Press, 1998.
9. R. Patriarche, S. Gedzelman, G. Diallo, D. Bernhard, C-G. Bassolet, S. Ferriol, A.Girard, M. Mouries, P. Palmer, A. Simonet, M. Simonet, “ A Tool for Textual and Conceptual Annotations of Documents ” in *E-challenges Conference*, Ljubljana, 19-21 Oct.2005.
10. R. Patriarche, S. Gedzelman, G. Diallo, D. Bernhard, C-G. Bassolet, S. Ferriol, A.Girard, M. Mouries, P. Palmer, A. Simonet, M. Simonet, “Noesis Annotation Tool. un outil pour l’annotation textuelle et conceptuelle de documents”, in *16èmes journées francophones d’ingénierie des connaissances (IC’05)*, Nice, 30 mai – 3 juin 2005
11. M. Simonet, G. Diallo, P. Palmer, A. Simonet, “Ontologies pour l’organisation des connaissances – Vers une ontologie anatomo-fonctionnelle du cerveau”, in *Création d'une base de connaissances anatomo-fonctionnelle : application au cerveau et au cœur, Santé et Systémique*,. Vol.7, N. 3-4, Hermès, 2003, pp 47-75.
12. M. Simonet , D. Bernhard, G. Diallo, P. Palmer, S. Ferriol, F. Baldesare, M. Casella Dos Santos, H. Cools, C. Dhaen, “Enrichissement d’une ontologie multilingue à partir de textes pour le cancer du sein”, in *WSM 04- Web Sémantique Médical*, Rouen, 2004.
13. SKOS : Schema for Knowledge Organisation Systems http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20050510/
14. J. F. Sowa, *Knowledge Representation: Logical, Philosophical, and Computational Foundations*, Brooks Cole Publishing Co., Pacific Grove, CA, 2000.
15. M. Uschold, M. King, “Towards a methodology for building ontologies” in *Procs IJCAI-95, Workshop on Basic Ontological Issues in Knowledge Sharing*, Canada, 1995.
16. M. Uschold, “Building Ontologies: towards a unified methodology”, in 16th Annual Conf. of the British Computer Society Specialist Group on Expert Systems, 1996.