Building an Ontology of Cardio-Vascular Diseases for Concept-Based Information Retrieval

S Gedzelman, M Simonet, D Bernhard, G Diallo, P Palmer

Université Joseph Fourier, TIMC Laboratory, Grenoble, France

Abstract

Word-based Information Retrieval (IR) suffers from several drawbacks which the Semantic Web initiative aims at overcoming through the use of ontologies. The European project Noesis (www.noesis-eu.org) is currently developing a concept-based approach to IR with Cardio-Vascular diseases as an application domain. For this purpose, an ontology of CV diseases had to be built. We present the process of its construction and its possible usages.

1. Introduction

The NOESIS project is an Integrated Project of the European 6th Framework Program (2004 – 2006). Its objective is to build a platform for wide scale integration and visual integration of medical intelligence, with particular emphasis on Knowledge Management and Decision aid. The application domain is that of Cardio-Vascular diseases. A kernel ontology has been designed for automatic and manual indexing purposes, bringing a direct and indirect support to users, respectively within the Noesis Annotation Tool and the Information Retrieval system. Working at a conceptual level in texts minimizes the ambiguïty of natural language and enhances retrieval accuracy.

Building a specialized biomedical ontology from scratch, with experts' knowledge or from text analysis (NLP algorithms [2]) requires a huge effort of conceptualization [1] and a long editing time (with ontology editors such as Protégé [10]) which has already been done in classical thesaurii and terminologies. Therefore, our choice was to rely on existing resources such as the UMLS (Unified Medical Language System [15]) metathesaurus and the MeSH (Medical Subject Headings [7]) thesaurus, often used in document indexing. For instance, the Medline system [8] indexes scientific biomedical articles with MeSH concepts and CISMeF [5] uses a French version of the thesaurus to classify different kind of documents (French translation provided by INSERM).

2. Knowledge resources

To highlight the difference between terminologies and ontologies, Bodenreider explains in [3] that the first are used to identify "the concepts in the text whereas the second help identify the relationships among concepts suggested by syntactic and discourse structures".

Indeed, usually the inner structure of terminologies does not provide more than vocabulary and subsomption relationships. But to enhance information retrieval, both associative and hierarchical relationships are exploited, notably to build expanded queries. The MeSH and UMLS are in fact more than simple terminologies, since they can provide such information [20].

About the MeSH Thesaurus

The MeSH thesaurus has a particular semantic clustering. *Concepts* are not the highest level of conceptualization, but there are *descriptors*, an entity uniting several closely related concepts. For example, the descriptor "***Anti-Arrhythmia Agents***" shown in fig.1 gathers three concepts including itself, "*Cardiac Depressants*", and "*Antifibrillatory Agents*". The descriptor's label is usually taken from the broadest concept within the set or the most common concept used in indexing.

This entity (descriptor) increases the possible number of terms to be found in texts and moreover allows language interoperability. For instance, the concept "*Heart Block*" does not have its counterpart in every language (no possible translation in Italian while it is translated into "*Bloc cardiaque*" in French, "*Herzblock*" in German, "*Bloqueo cardiaco*" in Spanish). However the descriptor "***Heart Block***" includes two additionnal concepts: "*Auriculo Ventricular Dissociation*" and "*Atrioventricular Block*", the latter having an Italian corresponding term "*Blocco atrioventricolare*".

Relationships are defined in the MeSH at the descriptor level as well as at the concept level (Fig. 1). "Is-A" or "Is-narrower" are both describing subsomption between descriptors and between concepts.

This general relationship "See-Also" is present

between concepts (Related) and between descriptors (SeeRelatedDescriptor). Descriptors can also be linked to semantic types (from the UMLS semantic network) and to qualifiers, another MeSH entity characterizing contextual use. In our example, the descriptor "***Anti-Arrhythmia Agents***" is a semantic type of "*Pharmacologic substance*, and can be seen in the context of "*Analysis*", "*Adverse Effects*" or "*Administration & Dosage*" and so on.

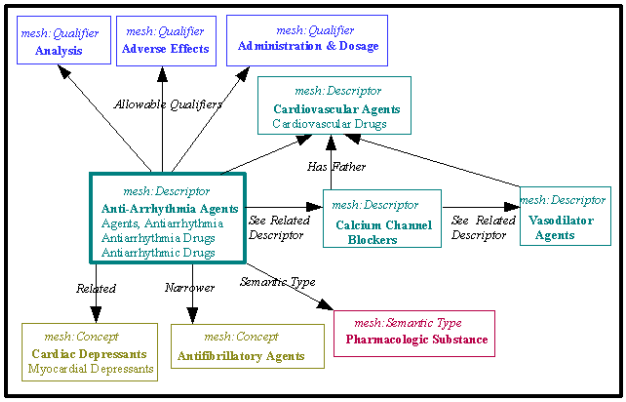

Fig. 1 : Schema presenting the different MeSH elements related to the descriptor "Anti-Arrhythmia Agents".

Since the MeSH thesaurus provides rich information about concepts and relations, efforts have been made to simply migrate this thesaurus to OWL [13], a formal web ontology language relying on description logics. Contrary to [16], the CISMeF project [14] altered to some extent the inner structure of the thesaurus, correcting for example some IsA relationships into PartOf ones. However, this work was done only on the French version of the MeSH. To represent the link between documents and MeSH concepts, their documents are instances of classes which can either be simple classes (like MeSH concepts) or classes aggregated by the union or intersection operators, by which the document would be indexed. For example, the resource 112 is declared as an instance of the R_112 class which is the intersection of "viral vaccine" and "hepatitis diagnostic" (see declaration below).

```
<owl:class rdf:ID= "R_112">
 <owl:intersectionOf
       rdf:parseType="Collection">
 <owl:class rdf:about= "#viral vaccine">
 <owl:Restriction>
  <owl:onPropertyrdf:resource= "#qu_diagnosis">
  <owl:someValuesFromrdf:resource= "#hepatitis">
 </owl:Restriction>
</owl:class>
```

As the Noesis ontology is designed for Information Retrieval and Semantic Annotation in the domain of Cardio-Vascular diseases, it has to contain as much vocabulary from the field as possible. The core ontology was taken from the MeSH, by selecting the concepts in 10 hierarchies dealing with the CV domain:

– Cardiovascular Diseases (C14)
– Cardiovascular Agents (D18)
– Cardiovascular Physiology (G09-330)
– Cardiovascular Surgical Procedures (E04.100)
– Cardiovascular System (A07)
– Syphillis Cardiovascular (C01.252.400.840.744.657)
– Cardiovascular Abnormalities (C16.131.240)
– Cardiovascular Diagnostic and Techniques (E01.370.370)
– Cardiology (G02.403.776.409.163)
– Cardiovascular Models (H01.770.461.395.161)

Some pieces of information taken from the MeSH descriptors like definitions, comments, terms, relations ("is-a" and "see-also"), concept references (id) have also been integrated into the core ontology. From French and Greek versions of the MeSH we could also get French and Greek terms. However, the MeSH thesaurus mainly provides the concept names and only few natural language synonyms.

The enrichment of the vocabulary has been designed in two steps. First the UMLS terms corresponding to the core concepts have been added in English, French, Italian, Spanish and German (this is the situation to-date). In a second phase we will consider English texts of the CV domain and add the vocabulary found in these texts which is absent in the ontology. The new terms will then be considered for translation into the five other languages.

About the UMLS metathesaurus

UMLS unifies most of the known medical classifications and thesauri. It integrates more than 2 million names for some 900 000 concepts, and is still growing, for example with the Consumer Health Vocabulary project [17] which will include patients' words and phrases about health.

Although UMLS gathers hierarchies from various sources (MeSH, SNOMED, and so forth), these remain quite independent from one another. Common concepts are referred to by a same identification code, which allows coreference of their terms. Although there are many inconsistencies (e.g., cycles) in the UMLS metathesaurus [2][18], it is a very rich source of Knowledge and it provides a variety of terms in several languages.

In UMLS, a concept identifier (CUI) has been assigned to every concept in the MeSH, contrary to us and [14] who have chosen to consider only the descriptor level (the concept names inside a descriptor are considered as synonyms) because of our objective, information retrieval, while UMLS aims at uniting terminologies.

3. Ontology building

Method

The extraction of MeSH and UMLS information has been automatized thanks to a GUI tool (MOB API [6]). This tool can also be used to build ontologies for other domains with the same OWL structure as described in the Results section. Users may select any level of the MeSH hierarchies they wish to extract. They can provide their own metadata for the ontology (author, title and description of the ontology) and choose target languages for UMLS vocabulary enrichment. The structure of the OWL file automatically produced by the tool will soon be customizable.

Results

The ontology extracted from the MeSH thesaurus for the CV domain contains 690 OWL classes, each corresponding to a MeSH descriptor, 2070 English terms and 966 French terms. The first enrichment step increased the English vocabulary by 10000 new terms, 2500 being true synonyms and the other lexical variants. The concept is represented by its MeSH *id* and by one term in each language, called the *preferred term* (there is one preferred term per language). As neither OWL nor RDF offers a support to distinguish preferred terms and multiple languages, we have used SKOS [12], which provides both *prefLabel* for preferred terms and *altLabel* for other terms.

A simplified example of an OWL class is given below for the concept ANTI-ARRHYTHMIA AGENTS, whose preferred term in English is "Anti-Arrhythmia Agent" and examples of alternative terms are "Anti Arrhythmia Drugs", "Cardiac Depressants", "Myocardial Depressants" and "Antifibrillatory Agents".

```
<owl:Class rdf:ID="M0001326">
  <skos:prefLabel xml:lang="en">Anti-Arrhythmia
Agents</skos:prefLabel>
  <skos:altLabel xml:lang="en">Anti Arrhythmia
Drugs</skos:altLabel>
  <skos:altLabel xml:lang="en">Cardiac
Depressants </skos:altLabel>
  <skos:altLabel xml:lang="en">Myocardial
Depressants </skos:altLabel>
  <skos:altLabel xml:lang="en"> Antifibrillatory
Agents </skos:altLabel>
  <rdfs:subClassOf rdf:resource="#M0003472"/>
  <!--class = Cardiovascular Agents-->
  <rdfs:seeAlso rdf:resource="#M0000495"/>
  <!--class = Adrenergic beta-Antagonists-->
  <rdfs:seeAlso rdf:resource="#M0003165"/>
  <!--class = Calcium Channel Blockers-->
  <rdfs:comment>consider also arrhytmia
</rdfs:comment>
 <rdfs:isDefinedBy>Agents used for the treatment
or prevention of cardiac arrhythmias …
</rdfs:isDefinedBy>
</owl:Class>
```

4. Uses of the cardiovascular ontology within the Noesis project

The Schema in Fig. 2 shows the role of the ontology for Knowledge Management in the Noesis project.

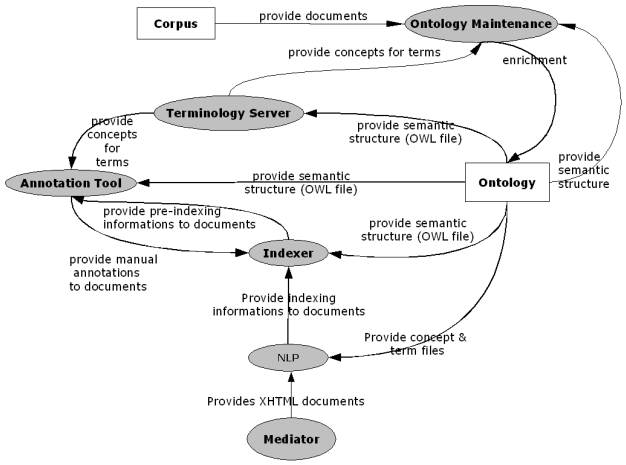

Fig. 2: Noesis modules interaction with the ontology

Several modules strongly interact with the ontology. The **Indexer** performs automatic indexing of the texts according to the ontology. A CF/ICF (*concept frequency / inversed concept frequency*) approach has been developed to adapt at the concept level the classical TF/IDF method which operates at the term level [11].

The **Annotation Tool** [9] aims at enabling the sharing of knowledge between users within the Noesis community (Fig.3). Users can add personal notes to scientific documents through textual annotations. They can also experience collaborative indexing, guided by the ontology via a graphical tree (Fig. 3, bottom left) or aided by a Terminology Server (bottom right).

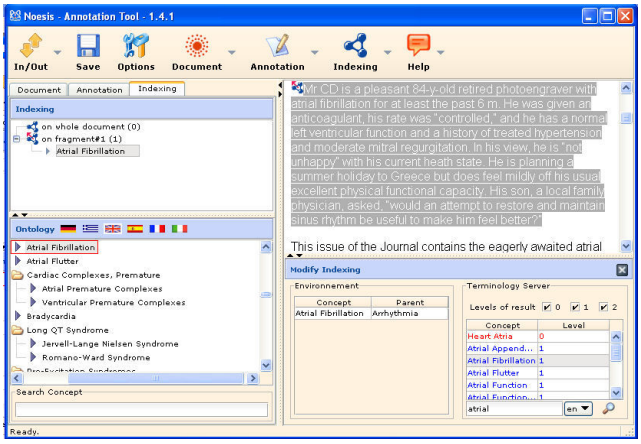

Fig. 3: The indexing window in the Annotation Tool.

The **Terminology Server** proposes concepts corresponding to user terms, thus helping her/him to find the concept(s) which best fit the part of the text she/he wants to semantically annotate. Fig. 3 (bottom right) shows the concepts proposed for the term "atrial".

Semantic annotation is done manually. It can complement or correct the results of automatic indexing.

5. Discussion and conclusions

At the present stage the ontology still has to be enriched by vocabulary taken from the scientific literature in the CV domain. For this purpose, a corpus of 500 English texts has been constituted (articles taken from www.biomedcentral.com). Designing and enriching an ontology is a difficult task, for which there is no agreed upon methodology. Some authors [4][19] have already investigated the richness of corpus for ontology enrichment.

From the experience gained in previous projects we have designed an environment for ontology design and enrichment based on texts. This environment comprises various tools such as a term extractor (to propose new terms extracted from the texts), a concordancer (to visualize terms in their context), a Terminology Server (to help the user find existing concepts from a given term) and an ontology editor which has been designed to manage multilingual ontologies – a feature which is notably absent from current ontology editors.

Further work will deal with enrichment of the SeeAlso relationship (currently 270 occurrences), which is the basis for the semantic expansion of users'queries.

Acknowledgements

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

References

[1] Bensilamane D, Arara A, Yetongnon K, Gargouri F, Ben-Abdallah H. Two approaches for ontologies building: from-scratch and from existing data sources. International Conference on Information Systems and Engineering, ISE 2003, July 20 - 25, 2003, Montreal, Canada.

[2] Bodenreider O, Burgun A. Aligning Knowledge Sources in the UMLS: Methods, Quantitative Results, and Applications. Medinfo. 2004 Sept.; 2004:327-331.

[3] Bodenreider O, Mitchell JA, McCray AT. Biomedical Ontologies. Pacific Symposium on Biocomputing: World Scientific. 2003;:562-564.

[4] Bourrigault D, Aussenac-Gilles N. Construction d'ontologies à partir de textes. Actes de la 10ème conférence annuelle sur le Traitement Automatique des Langues (TALN 2003), Batz-sur-Mer, 2003, T2 ;. 27-50.

[5] CISMeF: Catalog and Index of French-language Health Internet resources. http://www.chu-rouen.fr/cismef/cismefeng.html A quality-controlled subject gateway, Rouen, Avril 2002.

[6] Gedzelman S. MOB: Mesh Ontology Builder, access and use, soon available on TIMC website. http://www-timc.imag.fr/osiris

[7] Medical Subject Heading Browser. http://www.nlm.nih.gov/mesh/MBrowser.html

[8] Medline. http://medline.cos.com/

[9] Patriarche R, Gedzelman S, Diallo G, Bernhard D, Bassolet CG, Ferriol S, Girard A, Mouries M, Palmer P, Simonet A, Simonet M. A Tool for Conceptual and Textual Annotation of Documents. e-challenges conference, Ljubljana, 19-21 Oct. 2005.

[10] Protégé (2000). http://protege.stanford.edu/index.html

[11] Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Information Processing and Management, Vol. 24, Issue 5, 1998; 513-523.

[12] SKOS : Schema for Knowledge Organisation Systems. http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20050510/

[13] Smith MK, Welty C, McGuiness D. Owl web ontology guide. http://www.w3c.org/TR/owl-guide.

[14] Soualmia LN, Golbreich C, Darmoni SJ. Representing the MeSH in OWL: Towards a semi-automatic Migration. First International Workshop on Formal Biomedical Knowledge Representation, collocated with KR 2004; 1-12. Whistler, Canada.

[15] The Unified Medical Language System. http://umlsks.nlm.nih.gov

[16] Van Assem M, Menken M, Schreiber G, Wielemaker J, Wielinga B. A Method for Converting Thesauri to RDF/OWL. ISWC'04, Hiroshima, LNCS, volume 3298; 17-31, 2004.

[17] Zeng QT, Tse T, Crowell J, Divita G, Roth L, Browne AC. Identifying Consumer-Friendly Display (CFD) Names for Health Concepts. Proc AMIA Symp 2005: In press. http://www.consumerhealthvocab.org/

[18] Zweigenbaum P. L'UMLS entre langue et ontologie: une approche pragmatique dans le domaine médical. Revue d'Intelligence Artificielle, 18:111-137, 2004.

[19] Zweigenbaum P, Grabar N. Corpus-based associations provide additional morphological variants to medical terminologies. In Mark Musen, editor, Actes AMIA Annual Fall Symposium 2003, Washington, DC, novembre 2003. AMIA.

[20] Zweigenbaum P. Encoder l'information médicale: des terminologies aux systèmes de représentation des connaissances. Innovation Stratégique en Information de Santé, (2-3):27-47, 1999.

Address for correspondence
Michel SIMONET
TIMC, Faculté de Médecine
Institut de l'Ingénierie et de l'Information de Santé.
38700 LA TRONCHE
FRANCE
Email : Michel.Simonet@imag.fr