Representing the MeSH in OWL: Towards a Semi-Automatic Migration

LF. Soualmia[1,2], C. Golbreich[3], SJ. Darmoni[1,2]

[1] CISMeF & L@STICS, Rouen University Hospital, 76031 Rouen, France
{Lina.Soualmia, Stefan.Darmoni}@chu-rouen.fr
[2]PSI Laboratory - FRE CNRS 2645, INSA-Rouen, 76131 Mont-Saint Aignan, France
[3] Laboratoire d'Informatique Médicale, University Rennes 1, 35033 Rennes, France
Christine.Golbreich@univ-rennes1.fr

Abstract

Due to the numerous health documents available on the Web, information retrieval remains problematic with existing tools. This paper is positioned within the context of the CISMeF project (acronym of Catalogue and Index of French-speaking Medical Sites) and of a future Semantic Web. In CISMeF the resources are described using a set of metadata based on a structured terminology which "encapsulates" the MeSH thesaurus in its French version. Now, the objective is to migrate the CISMeF terminology, and thus the MeSH thesaurus, to a formal ontology, so as to get a more powerful search tool. The paper presents the very first stage and results of this ongoing project, aiming at migrating the MeSH to OWL. It reports on the first steps, which have presently been done, concerning the automatic transformation of the terminology into OWL-DL. First, the CISMeF terminology has been "formalized" in OWL. Then, the resulting OWL "ontology" has been imported under the Protégé editor which makes possible to check its consistency and its classification in using Racer. Finally, the paper concludes on the current results and encountered difficulties, and gives future work perspectives.

INTRODUCTION

The amount of health information available on Internet is considerable. Information retrieval remains problematic: users are now experiencing huge difficulties in finding precisely what they are looking for, among the tons of documents available online. Generic search engines (e.g. Google) or generic catalogues (e.g. Yahoo) cannot solve this problem efficiently and offer a selection of documents that turns out to be either too large or ill-suited to the query. Free text word-based (or phrase-based) search engines typically return innumerable completely irrelevant hits requiring much manual weeding by the user and might miss important information resources. Free text search is not always efficient and effective: the sought page might be using a different term (synonym) that points to the same concept; spelling mistakes and variants are considered as different terms; search engines cannot process HTML *intelligently*, the most the most widespread language on the Web.

This paper is positioned within the context of the CISMeF[1] project (acronym of Catalogue and Index of French-speaking Medical Sites) and of a future Semantic Web[2]. The CISMeF catalogue was developed since 1995 to assist the health professionals, the students and the general public in their search for health information on the Web. CISMeF is a quality-controlled health gateway, cataloguing the most important and quality-controlled sources of institutional health information in French in order to allow end-users to search them quickly and precisely.

In CISMeF the resources are described using a set of metadata elements based on a structured terminology which "encapsulates" the MeSH[3] (Medical Subject headings) thesaurus in its French version. The present work follows that done in[4] and aims at migrating the CISMeF terminology, and thus the MeSH thesaurus, to a formal ontology, so as to get a more powerful search tool[5]. Every year the MeSH thesaurus is modified and new concepts are added. As the rapid evolution of medical knowledge and the very dynamic nature of web information require frequent updates, a formal knowledge representation also contributes in maintaining a consistent terminology, by detecting the inconsistencies that might result from updates or modifications. We chose the OWL DL sublanguage[6] to represent the CISMeF terminology, as being the W3C standard and also as it provides powerful reasoning services based on Description Logics.

The paper presents the very first stage and results of an ongoing project aiming at "formalizing" the MeSH in OWL. The long term goal is to migrate the existing terminology to a formal representation in OWL and to enhance it. The paper main contribution concerns the modeling choices underlying the automatic migration process used for migrating MeSH to OWL. Section 2 introduces the CISMeF catalogue in which these experimentations are carried out. Modeling choices underlying the automatic transformation towards OWL are detailed in section 3. Section 4 presents the results of the consistency checking and classification of the

OWL “ontology”, after its import under the Protégé editor[7], using Racer[8]. Section 5 draws conclusion from the results and gives future work perspectives.

THE CISMeF TERMINOLOGY AND USE FOR RESOURCES INDEXING

The CISMeF catalogue describes and indexes a large number of health information resources (*n=13,198*) and has three main topics: guidelines for health professionals, teaching material for students in medicine, and consumer health information. A resource is any support that may contain health information : it can be a Web site, Web pages, documents, reports and teaching material. Metadata based on a terminology "ontology-oriented" are used to describe the resources.

CISMeF Terminology. The catalogue resources are indexed according to the CISMeF terminology, which is based on the French version of the MeSH concepts provided by the INSERM (National Institute of Health and Medical Research). The MeSH thesaurus in its 2003 version includes approximately 22,000 keywords (e.g. *abdomen, hepatitis*) and 84 qualifiers (e.g. *diagnosis, complications*). These concepts are organized into hierarchies from the most general to the most specific concept. For example, the keyword *hepatitis* is more general than the keyword *hepatitis viral A*. The qualifiers are used to specify which particular point of view of a keyword is addressed. For example the association of the keyword *hepatitis* and the qualifier *diagnosis* (noted *hepatitis/diagnosis*) restricts *hepatitis* to its *diagnosis* aspect. Qualifiers are also organized into hierarchies.

The heterogeneity of Internet health resources and the great specificity of MeSH keywords, which makes it difficult to refer broadly to a medical specialty, led the CISMeF group to introduce two new concepts, namely *metaterms* and *resource types*. Metaterms (*n*=66) concern medical specialties. The *resource types* (*n*=127) describe the nature of the resource e.g. *teaching material*, *clinical guidelines*. The keywords and qualifiers in CISMeF are thus clustered according to *metaterms.* Each one is related to one or several metaterms. The metaterms and resource types enhance information retrieval into the catalogue. In fact, meta-terms have been created to optimize information retrieval in CISMeF and to overcome the relatively restrictive nature of MeSH keywords. For instance, the queries 'guidelines in cardiology' and *'databases in psychiatry'* where *cardiology* and *psychiatry* are only MeSH keywords get few or no answers. Introducing *cardiology* and *psychiatry* as metaterms is an efficient strategy to get more results because instead of exploding one single MeSH tree (e.g. *psychiatry* as a MeSH keyword), using metaterms results in an automatic expansion of the queries by exploding other related MeSH or CISMeF trees as well as the current tree (e.g. *psychiatric hospital* as a MeSH keyword or *mental health dispensary* as a resource type will be exploded in the case of the *psychiatry* query).

The CISMeF terminology and the catalogue resources are stored in a relational database (Oracle 8.i). The CISMeF terminology has the same structure as a terminological ontology[9]:

- The vocabulary, that describes major terms of the medical domain, is well known by the librarians and the health professional.

- Each concept has:

  - a preferred term (Descriptor) to express it in natural language.

  a set of properties.

  - a natural language definition that allows to differentiate it from the concepts it subsumes and those that it subsumes.

  - a set of synonyms.

  - a set of constraints to apply on the qualifiers. For example the qualifier *‘Complications’* could only be used for the *‘Diseases’* arborescence.

  - a set of equivalences. For example the association ‘*Hepatitis/chemically induced’* is equivalent to the keyword *‘Hepatitis, toxic’*.

Many ways of navigation and information retrieval are possible into the catalogue. *Simple search* which is based on the subsumption relationships is the most often used. If the query, a given word or expression, can be matched with an existing term, then the result of the query is the union of the resources indexed by the term, and by the terms it subsumes, directly or indirectly, in all the hierarchies it belongs to. For example a query on *Hepatitis* will return as answer all the resources related to *Hepatitis* and also those related to *Hepatitis A, Hepatitis B…*etc. If the query cannot be matched, then the search is done over the other fields of the metadata. If it fails, a full-text search is carried out.

But although quite powerful, this kind of search requires a good knowledge of the medical domain, and exhibits some limitations.

Indeed, the consistency of this terminology has not yet been studied and some defaults may arise. For example, in the *‘Anatomy’* tree, some keywords are hierarchically organized according to a 'specialization' relationship, while in fact they are related by the *'part of'* relationship. As a consequence, a query on *‘headache*’ also returns documents on *‘mouth pain’*, *‘eye pain’* and *‘ear pain’* among others.

Another problem in query processing concerns the associations between keyword/qualifier. A query on "*hepatitis/diagnosis*" is processed in CISMeF as a conjunction of two queries one on "*hepatitis"* and one on "*diagnosis*". Thus, when exploded, this query returns also resources on "*lumbago/diagnosis*" and resources on "*lumbago/radiography*" since "*radiography*" is subsumed by "*diagnosis*".

The descriptions are incomplete. For example, the keyword "*abdominal neoplasm*" is defined as a "*neoplasm*" and not as an "*abdominal disease*" whereas "*stomach neoplasm*" is defined as "*neoplasm*" and a "*stomach disease*". The term "*abdominal disease*" does not exist in the MeSH. Therefore some improvements are now investigated. Because of its size, automatic tools are needed. A formal representation may be promising, in particular to verify the terminology consistency and the overall classification.

Metadata. The notion of metadata appeared before Internet but its interest has growth with the number of electronic publications and digital libraries. « The Semantic Web dream is of a Web where resources are machine understandable and where both automated agents and humans can exchange and process information.[1]». The solution proposed by the W3C is to use metadata to describe the data contained on the Web and to add semantic markup to Web resources that describes their content and functionalities, from the vocabulary defined in ontologies. Metadata are data about data or in the Web context, data describing Web resources. When properly implemented, metadata shall unambiguously describe resources, so enhancing information retrieval.

In CISMeF we use several sets of metadata. Among them there is the Dublin Core[10] (DC) metadata set, which is a 15-element set, intended to aid discovery of electronic resources. The resources indexed in CISMeF are described by eleven of the elements of Dublin Core: *author, date, description, format, identifier, language, editor, type of resource, rights, subject* and *title.* DC is not a complete solution; it cannot be used to describe the quality or location of a resource. To fill these gaps, CISMeF uses its own elements to extend the DC standard. Eight elements are specific to CISMeF: *institution, city, province, country, target public, access type, sponsorships* and *cost.* The user type is also taken into account. CISMeF defined two additional fields for the resources intended for the health professionals: indication of the *evidence-based medicine* and the *method* used to determine it. In the teaching resources eleven elements of the IEEE 1484 LOM (Learning Object Metadata) "Educational" category are added.

The metadata format was the HTML language in 1995. In 2000, in order to allow interoperability with other platforms the XML language became the metadata format. Since December 2002, the format used is RDF a basic Semantic Web language, within the EU-project MedCIRCLE framework[11] in which CISMeF is a partner. This project was initiated to qualify the quality of health information and to guide consumers to trustworthy health information. The vocabulary of the HIDDEL (High Information Description Disclosure Evaluation Language) metadata is contained in an ontology (represented in RDF Schema) and the resources are described in RDF according the concepts of the HIDDEL ontology.

AUTOMATIC MIGRATION TO OWL

There are several works concerning the UMLS® [24] and its Semantic Network representation with a formal language[12-15], but as far as we know, the MeSH formalization (a component of the UMLS metathesaurus), has not yet been studied. MeSH suffers from its size, its numerous inconsistencies and ambiguities concerning the medical concepts. For example, *'diagnosis'* is defined as a medical specialty and also a qualifier. In previous works MeSH has partly been enhanced by introducing new concepts in CISMeF[1] but it now appears not sufficient. An advantage of using description logics is to benefit of advanced inference services (satisfiability, subsumption, classification, consistency checking, instanciation, realization and retrieval), which can contribute to maintain a consistent terminological system and to improve results of queries thanks to inferences.

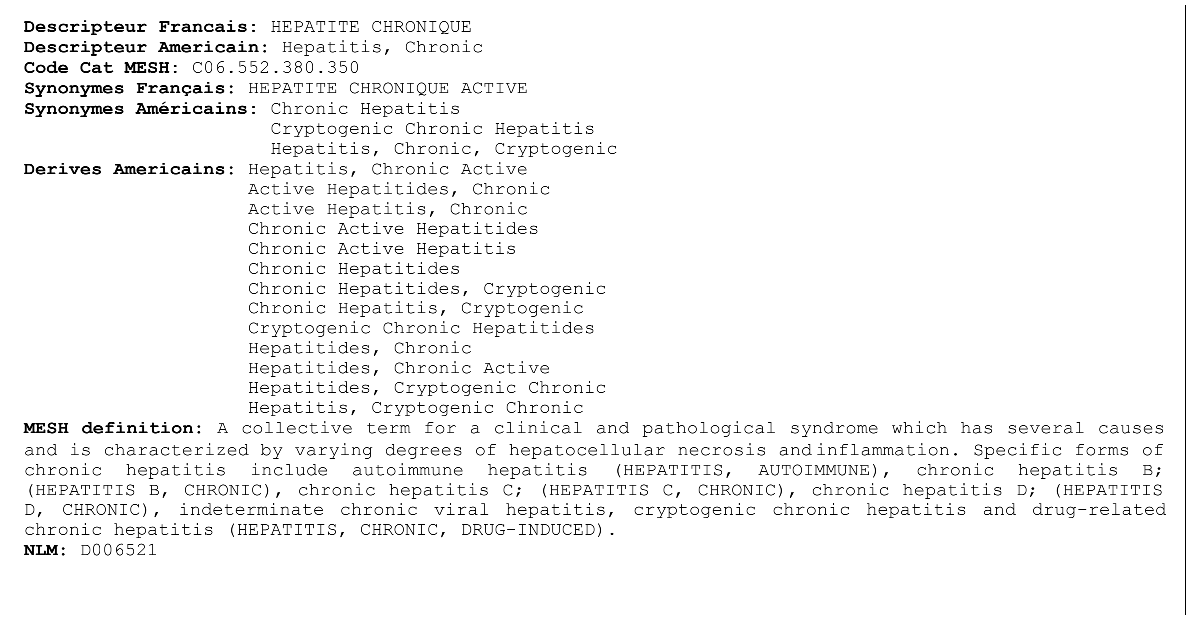
This section reports on the first stage of a general process aiming at the migration and enhancement of the MeSH.

Modeling principles. A first modeling principle was to "clean" the MeSH taxonomy, in distinguishing between the *'part-of'* and the *'is-a'* relationships (the *Anatomy, Biological Sciences* and *Geographic Locations* hierarchies are processed separately).

The second one was to clearly distinguish between the different notions: specialty, keyword, and qualifier. For example the specialty "*diagnosis*" is distinguished from the qualifier "*diagnosis*" because they denote different notions (resp. "*virology*").

The third one concerns the elicitation of the qualifiers domain. Qualifiers cannot be associated to all the keywords. It is a MeSH restriction. For example, the qualifier "*diagnosis*" can be associated to the keyword "*diseases*" (and thus to

[1] Ian Horrocks, IEEE Intelligent systems March / April 2002



```
Descripteur Francais: HEPATITE CHRONIQUE
Descripteur Americain: Hepatitis, Chronic
Code Cat MESH: C06.552.380.350
Synonymes Français: HEPATITE CHRONIQUE ACTIVE
Synonymes Américains: Chronic Hepatitis
                      Cryptogenic Chronic Hepatitis
                      Hepatitis, Chronic, Cryptogenic
Derives Americains: Hepatitis, Chronic Active
                    Active Hepatitides, Chronic
                    Active Hepatitis, Chronic
                    Chronic Active Hepatitides
                    Chronic Active Hepatitis
                    Chronic Hepatitides
                    Chronic Hepatitides, Cryptogenic
                    Chronic Hepatitis, Cryptogenic
                    Cryptogenic Chronic Hepatitides
                    Hepatitides, Chronic
                    Hepatitides, Chronic Active
                    Hepatitides, Cryptogenic Chronic
                    Hepatitis, Cryptogenic Chronic
MESH definition: A collective term for a clinical and pathological syndrome which has several causes
and is characterized by varying degrees of hepatocellular necrosis andinflammation. Specific forms of
chronic hepatitis include autoimmune hepatitis (HEPATITIS, AUTOIMMUNE), chronic hepatitis B;
(HEPATITIS B, CHRONIC), chronic hepatitis C; (HEPATITIS C, CHRONIC), chronic hepatitis D; (HEPATITIS
D, CHRONIC), indeterminate chronic viral hepatitis, cryptogenic chronic hepatitis and drug-related
chronic hepatitis (HEPATITIS, CHRONIC, DRUG-INDUCED).
NLM: D006521
```

Figure 1. Concept definition in the MeSH text file provided by the INSERM

all its descendants), but not to the "*geographic locations*". These restrictions on the qualifiers were formalized to check whether a qualifier is not wrongly associated to a keyword, and can be viewed as defining the domains of the qualifiers.

The fourth one concerns multiple hierarchies. A keyword in the MeSH may belong to several trees. In this case, for the moment, the keyword is associated to the intersection of its direct super-concepts.

Finally, since the objective is to remain as much as possible compatible with the original MeSH indexing, for each resource, the related MeSH concepts used for its indexing, serve to define a new concept of the ontology used for the resource new formal indexing. This new concept is defined from the conjunction of the original ones and will be used to define the individuals.

From Text Files to a Database. Each year the MeSH text files (Fig1.) are first processed using a awk script on a Unix platform to inform the table TB_MeSH in the CISMeF database which contains the following items: *Descripteur Français, Code Cat MeSH* and *NLM*. The other fields are not yet taken into account (e.g. MeSH definition) because they are in English. Nevertheless, the definitions are under translation into French in the context of the VuMeF project[16].

The *Code Cat MeSH* indicates in which hierarchy the descriptor is located and refers to a level position. A descriptor may belong to many hierarchies. This information is very useful to represent the hierarchies. For example one can deduce that *Hepatitis, Chronic* (C06.552.380.350) is subsumed by *Hepatitis* (C06.552.380) with a difference of level of 1. In practice, a join is done on the tables TB_MeSH and TB_MC, which contains all the descriptors used in the catalogue (n= 9,765), to update the terminology and also to compute all the existing links between descriptors and the levels in the hierarchies.

From the Database to a Terminological Knowledge Base. OWL-DL is a Description Logics (DL) language[17]. DL structures the domain knowledge at two levels: a terminological level (TBox or ontology), containing the classes of domain objects (concepts), with their properties (roles) and an assertional level (ABox), containing individuals (instances). In our case the ontology contains the specialties, keywords, qualifiers and resource types OWL-DL classes. Instances represent the indexed resources (to be soon included in the ABox under construction).

A DL system not only stores terminologies in a formal logic-based language, but also provides reasoning services. Main reasoning tasks concern satisfiability (existence of a model of the ontology), subsumption (supporting the classification of a concept in the hierarchy), and instance recognition (enabling to identify for a particular individual the most specific concepts it is an instance of).

The CISMeF terminology, is automatically transformed from the previous relational database into an OWL ontology, in using Java and SQL queries. The construction is a Top-Down construction, going from the Top concept to the specialties, and then to the keywords and resource types. The qualifiers hierarchy is modeled separately. The objective is to automate as much as possible all the process. As in [18] the illegal characters (- : , &) and spaces of the original descriptor names were replaced by underscores. All accented characters (*e.g.*, *"éèêë"*) were replaced by non-accented (*"e"*) ones. Names that began with numbers were prefixed with underscores. For example, "*11-hydroxycorticostéroïdes*" is renamed by "*_11_hydroxycorticosteroides*".

Representing the terminology in OWL

- ***OWL classes***

The keywords, metaterms and resource types, are represented as OWL classes. When two concepts have the same label but correspond to distinct notions, they are prefixed by *mt_* when it is a speciality, *tr_* when it is a resource type, *qu_* when it is a qualifier.

The specialties are, for the moment, represented as primitive concepts, without any OWL definition. Each specialty from the CISMEF specialty table, is automatically transformed into such a concept, for example, the specialty ‘*cardiology’* is represented by the OWL class:

```
<owl:Class rdf:ID="mt_cardiology">
```

- ***OWL hierarchies structuration***

The ‘is-a’ relations from the “cleaned” MeSH terminology are represented thanks OWL subsumption axioms. First, the keywords and resource types who are direct sons of the specialities are described. Then their descendants are progressively added level by level. For example ‘*accident domestique’* is a sub-concept of ‘*accidents’*:

```
<owl:Class rdf:ID="accident_domestique">
  <rdfs:subClassOf>
    <owl:Class rdf:about="#accidents" />
  </rdfs:subClassOf>
</owl:Class>
```

If a concept has more than one super-concept, it is represented as a subclass of the intersection of its super-concepts, for example ‘*accident_radiation’* is defined using the intersection of ‘*accidents*’ and ‘*accident_travail’*(occupational accident):

```
<owl:Class rdf:ID="accident_radiation">
  <rdfs:subClassOf>
    <owl:Class>
  <owl:intersectionOf rdf:parseType="Collection">
      <owl:Class rdf:about="#accident_travail" />
      <owl:Class rdf:about="#accidents" />
     </owl:intersectionOf>
    </owl:Class>
  </rdfs:subClassOf>
</owl:Class>
```

- ***OWL properties***

The qualifiers are represented as OWL properties, hierarchically organized. Each qualifier from the CISMeF qualifiers table, issued from the MeSH text files, is automatically transformed into a corresponding OWL property with a defined domain *“domain_qu_*”, but without any range. For example, the CISMEF qualifier ‘*contre-indications*’ is transformed into:

```
<owl:ObjectProperty
rdf:ID="qu_contre_indications">
<rdfs:domain
rdf:resource="#domain_qu_contre_indications" />
<rdfs:subPropertyOf>
<intersectionOf rdf:parseType="Collection">
<owl:ObjectProperty  rdf:about="#qu_pharmacologie"
/>
<owl:ObjectProperty
rdf:about="#qu_usage_therapeutique" />
</intersectionOf>
</rdfs:subPropertyOf>
</owl:ObjectProperty>
```

- ***The “part-of” property***

The keywords that belong to the trees *Anatomy, Biological Sciences* and *Geographic Locations* are organized hierarchically according to the *part-of* relationship. They are processed separately. The OWL property *partOf* is defined as:

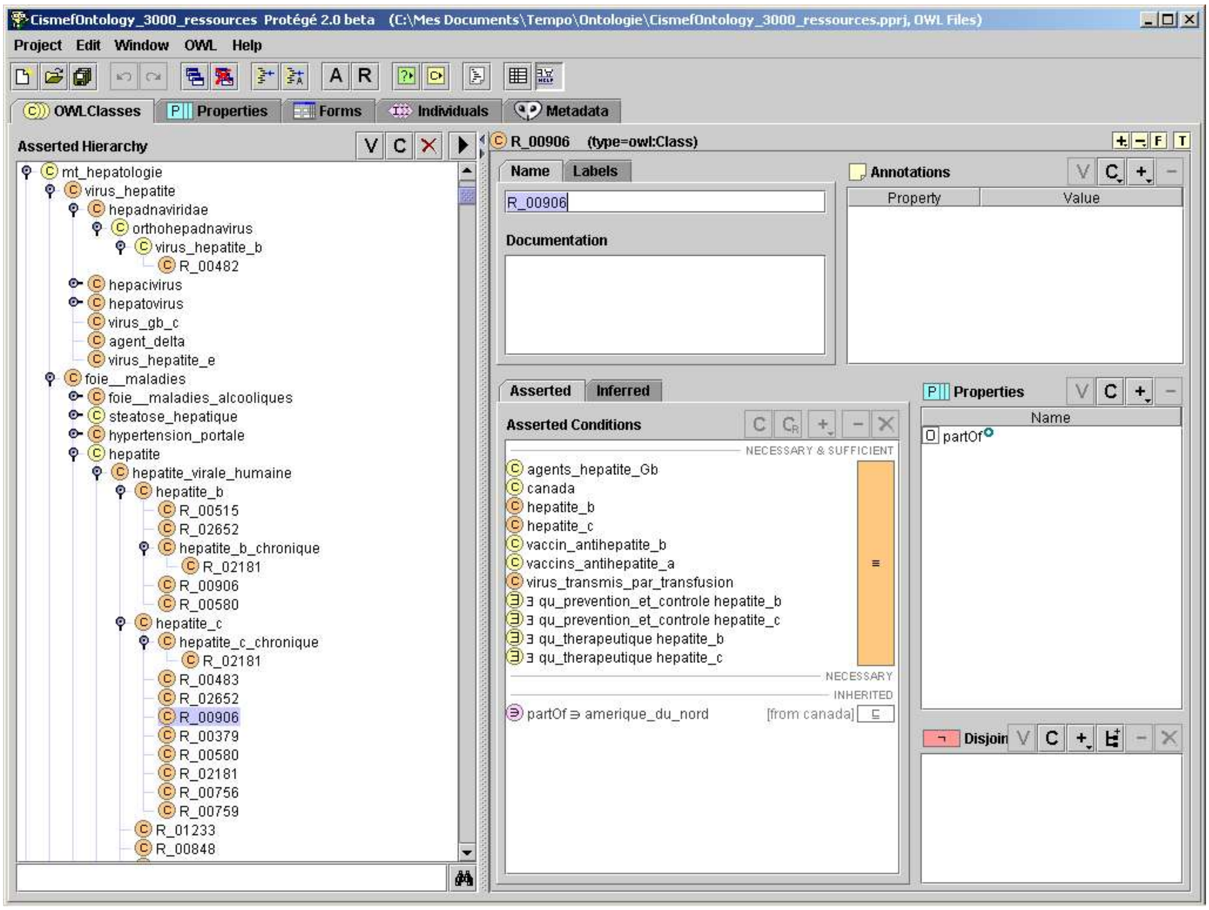
```
<owl:ObjectProperty rdf:ID="partOf">
</owl:ObjectProperty>
```

In the CISMeF (MeSH) terminology, the keyword *"abdomen"* is placed under the keyword "*region corps*" (body region) in the *Anatomy* tree. As this hierarchical relation corresponds in fact to a “*partOf”* relationship the concept *"abdomen"* is defined as:

```
<owl:Class rdf:ID="abdomen">
  <rdfs:subClassOf>
    <owl:Restriction>
        <owl:onProperty rdf:resource="#partOf" />
         <owl:someValuesFrom
         rdf:resource="#region_corps" />
    </owl:Restriction>
  </rdfs:subClassOf>
</owl:Class>
```

- ***Domains of properties***

Since a qualifier can be applied to several hierarchies of keywords, the domain of a property associated to a qualifier is represented by the union of the related qualified concepts. In CISMeF, this information is stored as a string in the item *Restriction* and the hierarchies roots are delimited by a comma, and was inserted manually by the medical librarian into the database. For example, “C01-C05, D, G” indicates that the considered qualifier can be applied to the keywords C01 to C05, D01 to D27, G01 to G14. For each restriction (84) such strings have been automatically processed so as to determine all the related keywords. For example the domain of the property *"qu_contre_indications"* has been defined as:



Name: R_00906

agents_hepatite_Gb
canada
hepatite_b
hepatite_c
vaccin_antihepatite_b
vaccins_antihepatite_a
virus_transmis_par_transfusion
∃ qu_prevention_et_controle hepatite_b
∃ qu_prevention_et_controle hepatite_c
∃ qu_therapeutique hepatite_b
∃ qu_therapeutique hepatite_c

partOf ∋ amerique_du_nord [from canada] ⊑

Figure 2. OWL ontology import into Protégé.

```
<owl:Class rdf:ID="domain_qu_complications">
<owl:unionOf rdf:parseType="Collection">
<owl:Class rdf:about="#anesthesie_analgesie" />
<owl:Class  rdf:about="#intervention_chirurgicale"
/>
<owl:Class
rdf:about="#produits_chimiques_inorganiques" />
        …
</owl:unionOf>
</owl:Class>
```

Representing the resources descriptions in OWL

The concepts related to the resources (n=13,198) have also been defined. For each resource, a new concept of the ontology has been created. For example the resource number 112, which is concerned by a diagnosis of some hepatitis and a viral vaccine, is indexed by *'hepatite/diagnostic'* (hepatitis/diagnosis) and *'vaccin antiviral'* (antiviral vaccines), therefore its description field of the metadata is represented as an instance of the defined concept R_112 = ∃ diagnostic.hepatite ∩ vaccin_antiviral.

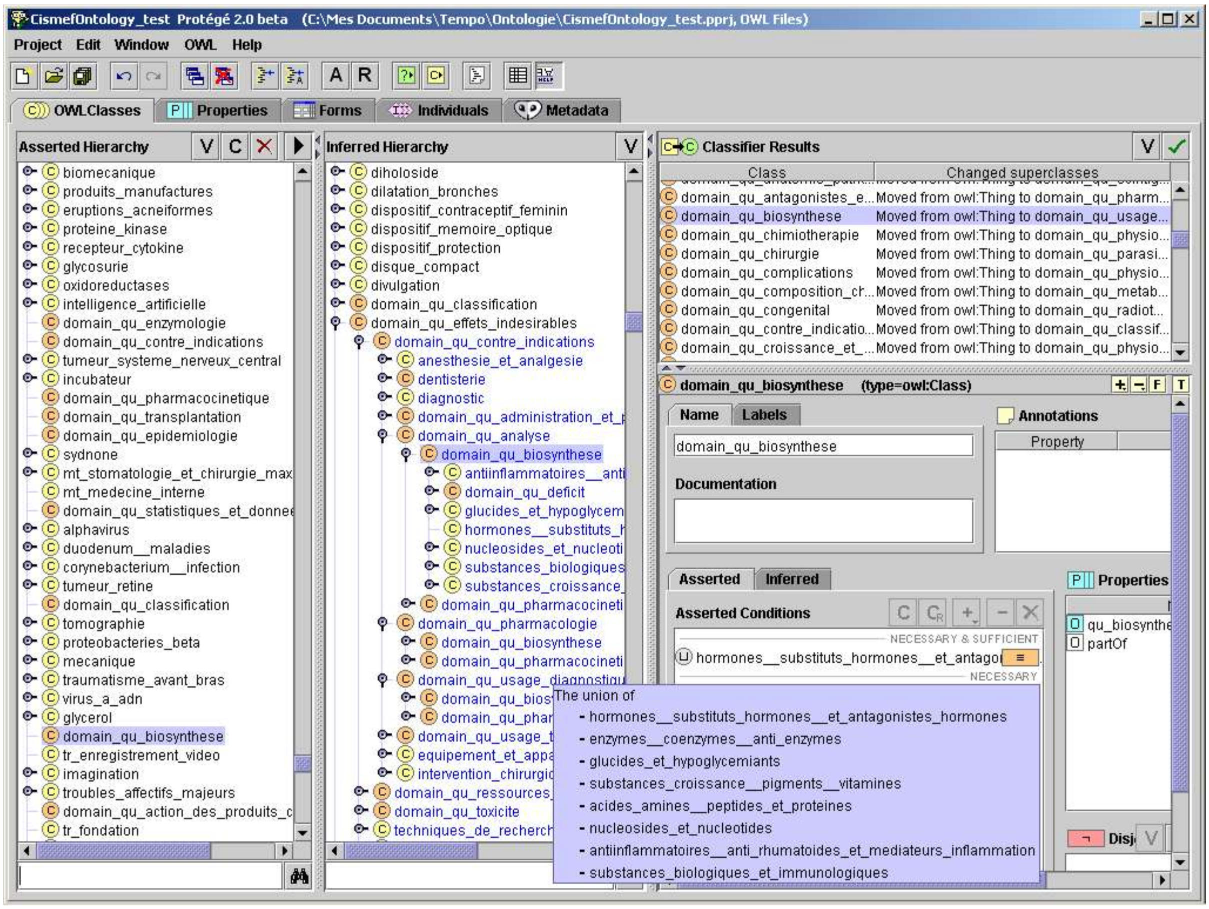
```
<owl:Class rdf:ID="R_112">
<owl:intersectionOf rdf:parseType="Collection">
<owl:Class rdf:about="#vaccin_antiviral" />
<owl:Restriction>
<owl:onProperty rdf:resource="#qu_diagnostic" />
<owl:someValuesFrom rdf:resource="#hepatite" />
</owl:Restriction>
</owl:intersectionOf>
</owl:Class>
```

CHECKING AND CLASSIFYING THE IMPORTED OWL 'ONTOLOGY'

Protégé OWL import. The size of the TBox is very large: 23,239 concepts (9,765 keywords; 65 specialties; 127 resource types; 84 domains; 13,198 concepts related to the resources) and 85 relations (84 qualifiers plus the relation *partOf*). It was not possible to import the resulting OWL file into the Protégé 2000 editor[7] thanks to its OWL (plug-in build 119) because the virtual Java machine had no sufficient memory, due to the file size (20.75 MB). Thus it was necessary to reduce the number of the concepts related to the resources to 3,000. The file loading has then been successfully processed in ~ 30 min (Fig.3). The ontology sub-language has been checked to be OWL-DL.

Figure 3 shows the concept R_00906, which represents a resource indexed with the concepts *agents hepatite Gb, Canada, hepatite b, hepatite c, vaccin anti-hepatite b, vaccins anti-hepatite a, virus transmis par transfusion, hepatite b/prevenrion et controle, hepatite c/ prevention et controle, hepatite b/therapeutique, hepatite c/therapeutique.* It which inherits the property partOf from its definition, as the concept *canada* is part of the concept *amerique_du_nord* (America, North).

Consistency checking. The consistency checking of all the terminology, augmented by the subset of 3,000 concepts describing resources, has approximately taken three hours (with Protégé 2.0 beta and the OWL plug-in build 119) using Racer[8]. No inconsistent class has been found. A little surprising, this may be explained by several reasons:



Class	Changed superclasses
domain_qu_antagonistes_e...	Moved from owl:Thing to domain_qu_pharm...
domain_qu_biosynthese	Moved from owl:Thing to domain_qu_usage...
domain_qu_chimiotherapie	Moved from owl:Thing to domain_qu_physio...
domain_qu_chirurgie	Moved from owl:Thing to domain_qu_parasi...
domain_qu_complications	Moved from owl:Thing to domain_qu_physio...
domain_qu_composition_ch...	Moved from owl:Thing to domain_qu_metab...
domain_qu_congenital	Moved from owl:Thing to domain_qu_radiot...
domain_qu_contre_indicatio...	Moved from owl:Thing to domain_qu_classif...
domain_qu_croissance_et_...	Moved from owl:Thing to domain_qu_physio...

domain_qu_biosynthese (type=owl:Class)

domain_qu_biosynthese

⊔ hormones__substituts_hormones__et_antago

The union of
- hormones__substituts_hormones__et_antagonistes_hormones
- enzymes__coenzymes__anti_enzymes
- glucides_et_hypoglycemiants
- substances_croissance__pigments__vitamines
- acides_amines__peptides_et_proteines
- nucleosides_et_nucleotides
- antiinflammatoires__anti_rhumatoides_et_mediateurs_inflammation
- substances_biologiques_et_immunologiques

Figure 3. New concepts classification (domains and resources' concepts).

- the pre-processing import of the MeSH into a structured database
- the distinction between the different notions (specialties, keywords, qualifiers and resource types)
- the use of the intersection operator for a class (object property) having several super-classes (super-properties)
- the classes, except those for resources and domains, have no description
- classes that describe the resources are OWL defined concepts, based on the CISMeF manual indexation, checked by the librarian team.

Classification. The classification was also very long (checking first whether the ontology is consistent). A new hierarchy has been inferred and all the domains and the concepts used for the resources have been classified according to their description. Fig 2 and Fig 3 show how the domains, initially defined as subclasses of Thing, and also many resources concepts, have been moved from Thing to another place. For example, the class '*domain_qu_biosynthese'* representing the domain of the property *qu_biosynthese* is subsumed by the class *domain_qu_analyse* describing the domain of the property *qu_analyse* (Fig.3) :

Because of it definition, the domain *domain_qu_biosynthese* has been moved from `owl:Thing` to *domain_qu_analyse:*

```
<owl:Class rdf:ID="domain_qu_biosynthese">
<owl:unionOf rdf:parseType="Collection">
<owl:Class
rdf:about="#hormones_substituts_hormones" />
<owl:Class
rdf:about="#enzymes__coenzymes__anti_enzymes" />
<owl:Class
rdf:about="#glucides_et_hypoglycemiants" />
<owl:Class
rdf:about="#acides_amines__peptides_et_proteines"
/>
<owl:Class
rdf:about="#nucleosides_et_nucleotides" />
<owl:Class
rdf:about="#substances_biologiques_immunologiques
" />
</owl:unionOf>
</owl:Class>

<owl:Class rdf:ID="domain_qu_analyse">
<owl:unionOf rdf:parseType="Collection">
<owl:Class
rdf:about="#produits_chimiques_inorganiques" />
<owl:Class
rdf:about="#composes_chimiques_organiques" />
<owl:Class
rdf:about="#composes_heterocycliques" />
<owl:Class
rdf:about="#hydrocarbures_polycycliques" />
<owl:Class
rdf:about="#hormones__substituts_hormones" />
<owl:Class
rdf:about="#agents_regulateurs_reproduction" />
<owl:Class
rdf:about="#enzymes__coenzymes__anti_enzymes" />
<owl:Class
rdf:about="#glucides_et_hypoglycemiants" />
<owl:Class rdf:about="#lipides_et_hypolipemiants"
/>
<owl:Class
rdf:about="#acides_amines__peptides_et_proteines"
/>
<owl:Class
rdf:about="#nucleosides_et_nucleotides" />
<owl:Class
rdf:about="#agents_systeme_nerveux_central" />
<owl:Class
rdf:about="#agents_systeme_nerveux_peripherique"
/>
<owl:Class
rdf:about="#agents_cardiovasculaires" />
<owl:Class rdf:about="#antiinfectieux" />
<owl:Class
rdf:about="#antineoplasiques_et_immunodepresseurs
" />
<owl:Class
rdf:about="#produits_dermatologiques" />
<owl:Class
rdf:about="#substances_biologiques_immunologiques
" />
<owl:Class
rdf:about="#materiaux_biomedicaux_et_dentaires" /
>
<owl:Class
rdf:about="#drogues_et_agents_divers" />
<owl:Class
rdf:about="#actions_chimiques_et_utilisations" />
</owl:unionOf>
</owl:Class>
```

CONCLUSION AND FUTURE WORK

Like the Gene Ontology migration[23], the MeSH formalization is a several steps process. This paper has presented the first steps achieved to transform the MeSH thesaurus into OWL-DL. The main contributions are its modeling principles, such as the distinction between *is-a* and *part-of* hierarchies, between concepts denoting different notions, the elicitation of properties domains etc., which support the automatic process. These first steps aiming at being automatic, are mainly based on syntactic transformations, achieved from the existing MeSH hierarchical organization. For the moment, this one has only been partly enhanced, but we are aware that a more "semantic" step, based on a careful investigation, is still needed and further improvements are planed. For example, particular links in the *Anatomy* hierarchy should be fixed, and defined as "is-a" relations instead of "part of": the MeSH sub-trees A11 (cells), A12 (fluids and secretions) and A15 (hemic and immune systems) are *is-a* hierarchies, and "*blood cell"* [A11.118] *is-a* "*cell"* [A11]. Other problems come from the MeSH 'is-a' hierarchies, that are not really well principled. For example *diagnosis_error* is defined in the MeSH, thus in consequence also in our OWL ontology, as a specialization of *diagnosis* and *medical_ error*, although an error *is not* a diagnosis. Instead, the concept *diagnosis_error* should be defined as a *medical_error "*about" a *diagnosis*, thus represented in OWL by *medical_error* $\cap$ $\forall$*about.diagnosis,* instead of their conjunction. A possibility to improve it and obtain such descriptions, is to use the UMLS Semantic Network relations, for instance like *is_complicated_by*, *is_treated_by* etc. for the diseases hierarchy. In addition, other properties, such as classical metadata (title, authors, format etc…), may be added to the concepts that describe the resources. The next steps of this project will be to enhance the OWL representation, to define all the individuals (resources), to use the retrieval reasoning service for query processing. Such a formal ontology issued from the MeSH, is promising and may be exploited in many applications, based on the MeSH thesaurus, mainly bibliographic databases such as Medline, and health gateways [19-22].

References

[1] Darmoni, SJ., Thirion, B., Leroy, JP. et al. (2001). A Search Tool based on 'Encapsulated' MeSH Thesaurus to Retrieve Quality Health Resources on the Internet. *Medical Informatics & the Internet in Medicine*, 26(3):165-178.

[2] Berners-Lee, T., Hendler, J., Lassila, O. (2001). The Semantic Web. *Scientific American*, 284(5):34-43.

[3] Nelson, SJ., Johnson, WD., Humphreys, BL. (2001) Relationships in Medical Subject Headings. In Bean and Green (Eds), 171-184.

[4] Soualmia, LF., Barry, C., Darmoni, SJ. (2003). Knowledge-Based Query Expansion over a Medical Terminology Oriented Ontology. Dojat, Keravnou, Barahona (Eds.), *LNAI # 2780*, Springer-Verlag, p.209-213.

[5] Golbreich, C. (2003) Towards a Sophisticated Multimedia Documents Search Engine. *Bulletin AFIA,* n°55, 40-54.

[6] Horrocks, I., Patel-Schneider, PF., van Harmelen, F. (2003) From SHIQ and RDF to OWL: The making of a web ontology language. *Journal of Web Semantics*. 1(1):7-26, 2003

[7] Noy, NF., Sintek, M., Decker, S., et al. (2001) Creating Semantic Web Contents with Protege-2000. *IEEE Intelligent Systems* 16(2):60-71.

[8] Haarslev, V., Möller, R. (2001) Description of the RACER System and its Applications. In *International Workshop in Description Logics 2001 (DL2001),* Stanford.

[9] Sowa, JF. (2000) Ontology, Metadata and Semiotics. B.Ganter, G.W.Mineau (Eds), Conceptual Structures: Logical, Linguistic, and Computational Issues, *LNAI #1867*, 55-81.

[10] Baker, T. (2000) A Grammar of Dublin Core. *Digital-Library Magazine,* vol 6 n°10.

[11] Mayer, MA., Darmoni, SJ., Fiene, M., et al. (2003). MedCIRCLE Modeling on the Semantic Web. Surjan, Engelbrecht, McNair (Eds) *Stud. Health Technol. Inf.* 95:667-672.

[12] Schulz, S. Hahn, U.(2001) Medical Knowledge Re-engineering – converting major portions of the UMLS into a terminological knowledge base. *IJMI,* 64(2-3):207-221.

[13] Horrocks, I., Rector, A. (1997) Experience Building a Large, Re-usable Medical Ontology using a Description Logic with Transitivity and Concept Inclusions. *Workshop on Ontological Engineering AAA Spring Symposium.*

[14] Cornet, R., Abu-Hanna, A. (2002) Usability of Expressive Description Logics – A Case Study in UMLS. *AMIA 2002,* 180-184.

[15] Kashyap, V., Borgida, A. (2003) Representing the UMLS Semantic Network using OWL. *ISWC 2003.*

[16] Darmoni, SJ., Jarousse, E., Zweigenbaum, P., et al. VuMeF: Extending the French Involvement in the UMLS Metathesaurus, *AMIA 2003,* 824.

[17] Baader, F, Calvanese, D., McGuinness,D., et al (2003) The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press.

[18] Golbeck, J., Fragoso, G., Hartel, F., et al. (2003) The National Cancer Institute's Thésaurus and Ontology. *Journal of Web Semantics*.

[19] Hersh, WR., Brown, KE., Donohoe, LC., et al. (1996) CliniWeb: managing clinical information on the World Wide Web. *JAMIA,* 3(4):273-80.

[20] Norman, F. (1998) Organising Medical Networks' information. *Med. Inf.* 23:43-51.

[21] Boyer, C., Baujard O., Baujard, V., et al. (1997) Health On the Net automated database of Health and medical information. *IJMI* 47(1-2):27-9.

[22] Deacon, P., Smith, JB., Tow, S. (2001) Using metadata to create navigation paths in the HealthInsite Internet gateway. *Health Info Libr J.* 18 (1) piii: 20-9.

[23] Wroe, C.J., Stevens R.., Goble C.A., Ashburner M.. A Methodology To Migrate The Gene Ontology To A Description Logic Environment Using DAML+OIL. *Proceedings of the 8th Pacific Symposium on Biocomputing (PSB)*, Hawaii. January 2003.

[24] Lindberg DAB, Humphreys BL, McCray AT, The Unified Medical Language System. Meth Inform Med, 1993, 32(4): 281-91