Une Terminologie du Domaine Médical: Structure et Exploitation

Lina F. Soualmia*,, Aurélie Névéol*,**, Magaly Douyère***

Benoît Thirion*, Alexandrina Rogozan, Stéfan J. Darmoni*,****

** Equipe CISMeF, L@STICS, CHU & Faculté de Médecine de Rouen*
1, rue de Germont , 76031 Rouen Cedex
{lina.soualmia, magaly.douyere, benoit.thirion, stefan.darmoni}@chu-rouen.fr

*** Laboratoire PSI CNRS FRE-2645, INSA & Université de Rouen*
Place Emile Blondel, BP-68 , 76131 Mont Saint Aignan
{aneveol, arogozan}@insa-rouen.fr

Résumé:

Nous détaillons ici la structure d' une terminologie du domaine médical qui est utilisée dans un catalogue de santé pour indexer et rechercher des documents. Nous montrons comment celle-ci a été modélisée à partir d'un thésaurus et comment elle est exploitée.

1. Introduction

Avec l'explosion du Web et la prolifération des connaissances biomédicales, les utilisateurs ont potentiellement accès à des informations de plus en plus nombreuses mais en réalité, ils sont obligés de naviguer dans un vrai labyrinthe de pages. C'est dans ce contexte que le catalogue CISMeF (Darmoni et al. 2001) a été développé en 1995 pour assister les professionnels de santé, les étudiants et le grand public dans leur recherche d'information de santé sur le Web. CISMeF et Doc'CISMeF, le moteur de recherche associé, prennent en compte la diversité des utilisateurs et leur permettent de trouver des documents de qualité qui répondent à un besoin précis. Un grand nombre de ressources (n=13,642) sont sélectionnées en fonction de critères stricts par une équipe de documentalistes et sont répertoriées selon une méthodologie de mise à jour du catalogue. Une ressource peut être un site Web, une page Web, un document, un rapport : tout support qui contient des informations relatives à la santé. La description de ces ressources se fait à l'aide de *notices* en se basant sur un ensemble de méta-données et une terminologie structurée semblable à une ontologie documentaire du domaine médical. Nous nous intéressons ici à la structure de cette terminologie. ((Soualmia et al., 2002) pour plus de détails concernant les métadonnées).

2. La Terminologie CISMeF

La terminologie CISMeF a été construite à partir des concepts du thésaurus MeSH[1] (développé depuis 1960) et de sa traduction en français fournie par l'INSERM[2]. Le MeSH dans sa version 2003 est composé d'environ 22,000 *mots clés* (exemples: *abdomen*, *hépatite*) et 84 *qualificatifs* (exemples: *diagnostic, complications, thérapeutique…*) regroupés sous la forme d'arborescences. Les mots clés correspondent à des concepts médicaux et sont organisés sous la forme de hiérarchie à 9 niveaux allant du terme le plus général en haut de la hiérarchie aux termes les plus spécifiques en bas de la hiérarchie. Par exemple le mot clé *aberration chromosomique* est plus général que le mot clé *trisomie*. Les qualificatifs, organisés également en hiérarchie, permettent de préciser le sens des mots clés en limitant leur étendue à certains aspects. Par exemple l'association du mot clé *lombalgie* et du qualificatif *diagnostic* (notée *lombalgie/diagnostic*) permet de restreindre la *lombalgie* au seul aspect *diagnostic*. Bien qu'il existe des ontologies médicales générales, comme GALEN (Rodrigues et al. 1998), ou spécifiques à un domaine comme MENELAS (Bouaud et al. 1995), c'est le MeSH qui a été choisi car il correspond aux attentes des documentalistes et il est connu des professionnels de santé. Les mots clés ont été regroupés dans CISMeF en fonction de spécialités médicales (n=66) intitulés *métatermes* (Cancérologie). Ce sont des super-concepts qui permettent une vision plus globale

[1] Medical Subject Headings. Le MeSH est produit par la US-National Library of Medicine pour la base documentaire Medline.
[2] Institut National de la Santé et de la Recherche Médicale http://dicdoc.kb.inserm.fr:2010/basimesh/mesh.html

concernant une spécialité. Les métatermes permettent de connaître l'ensemble des termes MeSH qui sont répartis dans plusieurs arborescences mais qui concernent une même spécialité. Une hiérarchie de *types de ressources* (*n*=127) a été modélisée et elle permet de décrire la nature de la ressource (*cours, information patient*). Les métatermes et les types de ressources permettent d'exprimer des requêtes complexes dans CISMeF comme des '*recommandations en cardiologie'* ou encore des *'cours en virologie'* ce qui n'est pas possible avec la structure actuelle du MeSH. A partir du MeSH fournit sous la forme de fichiers texte (Fig.1), seules les relations du type '*est-un*' et *'partie-de'* sont utilisées pour définir des liens père-fils dans la hiérarchie des mots clés CISMeF. Ces liens hiérarchiques sont exploités pour la recherche d'information et la navigation dans le catalogue (cf. Section 3). Par exemple le mot clé *Oreille* initialement défini comme étant *partie-de* du mot clé *Tête*, est défini dans la terminologie CISMeF par *Oreille* est fils de *Tête*. Les fichiers MeSH sont traités automatiquement pour renseigner la terminologie CISMeF afin qu'elle soit exploitable au niveau du site. La médecine étant un domaine qui évolue constamment (nouvelles maladies, nouveaux traitements…etc.), ces fichiers sont mis à jour chaque année (nouveaux mots clés, qualificatifs, nouvelles organisations dans les hiérarchies).

```
LT = POMPES IONIQUES
UF = POMPES A IONS
NT = ANTIPORTEURS
NT = PROTEINES DE TRANSPORT ANIONS
NT = PROTEINES DE TRANSPORT CATIONS
NT = SYMPORTEURS
RT = CANAL MEMBRANAIRE
RT = TRANSPORT BIOLOGIQUE ACTIF
RT = TRANSPORT IONIQUE
```

```
LT= Terme principal
UF= Synonyme
NT= terme spécifique
RT= Voir Aussi
```

Figure 1. Fichier texte fourni par l'INSERM

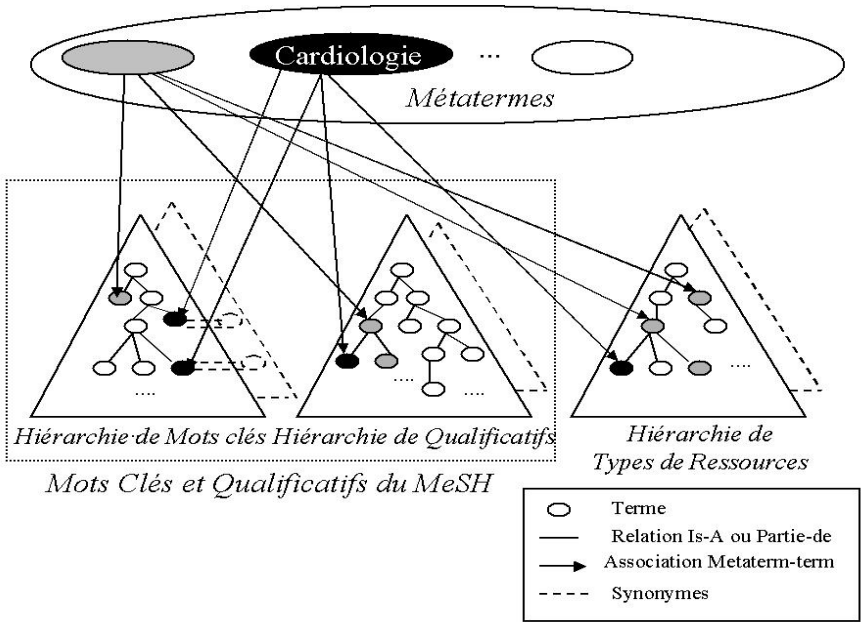

Figure 2. Structure de la terminologie CISMeF

La terminologie CISMeF (Fig.2) a une structure semblable une ontologie terminologique (Sowa, 2000) :

✓ Le vocabulaire est bien connu des documentalistes et des professionnels de la santé et il correspond à celui du domaine médical.

✓ Chaque concept (Fig.3) a :
- un terme préférentiel (Descripteur) pour l'exprimer en langue naturelle
- un ensemble de propriétés

- une définition en langage naturel pour quelquefois le différencier des concepts le subsumant et de ceux qu'il subsume (principe de (Bachimont, 2000))
- un ensemble de synonymes
- un ensemble de règles et de contraintes (Fig.4)

✓ Les concepts sont organisés selon une relation de subsomption allant du concept le plus général au plus spécifique.

```
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 définition: A collective term for a clinical and pathological syndrome
which has several causes and is characterized by varying degrees of
hepatocellular necrosis and inflammation. 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).
Numero NLM: D006521
```

Figure 3: Description du concept Hépatite Chronique

D'après l'exemple de définition de la Figure 3, le terme associé au concept ayant l'identifiant unique D006521 est *Hépatite Chronique*. Le code cat.MeSH indique à quel niveau ce concept est situé dans la hiérarchie: on peut déduire que *Hépatite Chronique (*C06.552.380.350) est subsumé par *Hépatite* (C06.552.380).

```
Hepatitis : C06.552.380+
Viral Hepatitis = Hepatitis, Viral Human and Hepatitis, Viral Animal
/chemically induced = Hepatitis, Toxic
/veterinary = Hepatitis, Animal or Hepatitis, Viral, Animal
hepatitis parenterally transmitted= Hepatitis C
hepatitis enterally transmitted = Hepatitis E
not specified as parenteral or enteral = probably Hepatitis, Viral, Human
Non-A, Non-B hepatitis = probably Hepatiis C
```

Figure 4: Exemple de règles et de contraintes

La Figure 4 est un exemple de contraintes sous la forme de règles à appliquer sur les concepts. Par exemple l'association *hépatite/induit chimiquement* est équivalente (=) au concept hépatite, toxique. Toutes ces informations ainsi que les notices descriptives des ressources sont stockées dans une base de données relationnelle qui est exploitée par le serveur du site Web de CISMeF. La structure de la terminologie est également exploitée pour l'indexation des ressources, la visualisation et la navigation dans les hiérarchies des termes du domaine et la recherche de ressources par le moteur Doc'CISMeF.

3. Exploitation

3.1. Indexation des ressources

A l'heure actuelle, l'indexation des ressources ajoutées dans le catalogue CISMeF (au rythme de 55 nouvelles ressources par semaine) est réalisée manuellement. Afin d'alléger la charge horaire importante liée à cette tache documentaire, nous travaillons actuellement au développement d'un

système d'indexation automatique qui devra correspondre au cahier des charges de l'indexation manuelle. Cela signifie que l'indexation doit consister en une liste de mots clés associés ou non à des qualificatifs. A chaque mot clé (ou couple (mot clé/ qualificatif)) est attribuée une pondération majeure ou mineure selon son importance dans le document. Par ailleurs, la notion de descripteur obligatoire doit être prise en compte. Les informations contenues dans la terminologie interviennent à plusieurs niveaux dans le processus d'indexation, aussi bien manuelle qu'automatique. Tout d'abord, les associations mot clé /qualificatif sont régies par la terminologie. En effet, chaque mot clé comporte une liste de qualificatifs pouvant lui être associés. Ainsi, le qualificatif *prévention et contrôle* pourra être associé au mot clef *hépatite*, mais pas au mot clef *oreille* alors que le qualificatif *virologie* pourra être associé à chacun de ces mots clés. L'indexation utilise également les relations de subsomption entre termes. Ainsi un document sur l'hépatite A devra être indexé au seul mot clé *hépatite A* qui est plus précis que *hépatite* ou *foie, maladies*. Certains aspects de la terminologie sont utilisées plus spécifiquement par l'indexation automatique (ainsi que par la recherche d'information). En effet, lors de l'analyse du texte pour en extraire les mot clefs MeSH, il est nécessaire de repérer dans un premier temps une série d'éléments textuels qui seront ensuite reliés aux mot clés correspondants. Ces éléments textuels comprennent les flexions des mots MeSH, mais également les synonymes de ces mots, contenus dans la terminologie. Ainsi, l'expression *femme enceinte* sera extraite, puis rapportée au mot clef *grossesse*.

3.2. Classification et Navigation

Classification de ressources: A partir d'un ensemble d'heuristiques et d'un algorithme de classification décrit dans (Névéol et al., 2004), les spécialités médicales (métatermes) auxquelles se rattachent les ressources sont déduites en utilisant les différents liens existants entre (métaterme-mot clé), (métaterme-qualificatif) et (métaterme-type de ressource) et classées en fonction de leur niveau d'importance.

Navigation: La navigation dans l'ontologie, grâce à un index thématique et alphabétique, permet à l'utilisateur d'appréhender les termes du domaine et leurs relations en affichant les différentes hiérarchies auxquelles il appartient. Chaque terme a sa propre page associée et des liens qui permettent de rechercher toutes les ressources qui y sont rattachées, ou encore restreindre la recherche en fonction de l'utilisateur (ressources destinées aux professionnels, aux étudiants ou aux patients et au grand public).

3.2. Recherche d'Information

Différents modes sont possibles :

- ✓ La recherche dite " simple " permet à l'utilisateur de saisir une requête en texte libre en français ou en anglais avec ou sans accent en majuscule ou en minuscule.
- ✓ La recherche dite " avancée " engage des recherches plus pointues à l'aide d'un formulaire contenant des listes déroulantes et permet de combiner plusieurs champs (mots clés, titre, année…etc.) avec des opérateurs booléens (ET, OU, SAUF).
- ✓ La recherche " logique " s'effectue à l'aide d'un langage de requêtes associé, des opérateurs booléens et des caractères spéciaux.

La recherche " simple " telle qu'en place aujourd'hui se base sur les relations de subsomption. Si le terme (un mot ou une expression) saisi par l'utilisateur est un terme existant dans la terminologie, le résultat de la requête est l'union de toutes les ressources instances du terme et des ressources instances des termes qu'il subsume, directement ou indirectement, et ce dans toutes les hiérarchies dans lesquelles il peut se trouver. Par exemple une requête sur le terme *tumeur* va renvoyer comme réponse l'ensemble des ressources rattachées à *tumeur* mais également celles rattachées à *tumeur colon* , *tumeur rectum*…etc. De même qu'une requête sur *tête* va renvoyer les ressources rattachées à *tête* mais également à *oreille, nez*…etc. et c'est d'ailleurs pour cette raison que nous considérons les liens *partie-de* du MeSH comme des liens de subsomption dans CISMeF. Si le terme saisi par l'utilisateur n'est pas un terme réservé, une recherche sur tous les autres champs de méta-données est effectuée, voire en plein texte sur tous les documents indexés. Ce type de recherche " simple " nécessite donc une bonne connaissance des termes de CISMeF, ce qui n'est pas évident pour un utilisateur novice.

4. Perspectives

Nos travaux actuels s'orientent vers l'amélioration du moteur de recherche pour permettre une *recherche d'information intelligente* et avons développé le système KnowQuE (Knowledge-based Query Expansion) (Soualmia et al., 2003) qui se fonde sur l'utilisation conjointe d'une base de connaissances morphologiques (Grabar et al., 2003) et d'une base de règles d'associations extraites par Data Mining. Le troisième composant (une ontologie formelle en OWL est en cours de réalisation).

En parallèle, le développement et le test du système d'indexation automatique sont en cours, avec une attention particulière pour les associations mot clé / qualificatif.

L'indexation et la recherche combinée texte et image, dans le catalogue de santé CISMeF, s'affirme comme une nécessité stratégique. L'enjeu de ce travail de recherche que nous avons déjà démarré dans le cadre d'une thèse de doctorat, est double : l'indexation d'images fixes, d'une part, et la définition des requêtes et des stratégies de recherche par similarité des images dans le catalogue, d'autre part. La recherche d'images dans les applications médicales (IRMA) est un processus complexe [Keysers et al, 2003]. Nous envisageons donc de passer par une étape intermédiaire : la catégorisation de la requête (texte et/ou image) sur un ensemble d'images prototypes. Cette étape nous permettrait d'augmenter l'expression sémantique de la requête d'images, par la sélection d'attributs d'image spécifiques pour une classe d'images, et l'optimisation du processus de recherche, grâce aux projections des termes et des relations de la terminologie CISMeF sur les images prototypes.

Références :

BACHIMONT, B. (2000) Engagement Sémantique et Engagement Ontologique : Conception et Réalisation d'Ontologies et Ingénierie des Connaissances. Charlet et al. (Eds). Ingénierie des Connaissances, Evolutions Récentes et Nouveaux Défis.

BOUAUD, J., BACHIMONT, B., CHARLET, J. and ZWEIGENBAUM, P. (1995) Methodological Principles for Structuring an « Ontology ». *Proceedings of IJCAI conference*.

DARMONI, SJ., THIRION, B., LEROY, JP., DOUYÈRE, M 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.

GRABAR, N., ZWEIGENBAUM, P., SOUALMIA, LF., and DARMONI SJ.(2003) Matching Controlled Vocabulary. Medical Informatics Europe, p.445-450.

KEYSÈRS D, DAHMEN J., NEY H., WEIN B., Lehmann T. : Statistical Framework for Model-based Image Retrieval in Medical Applications. Journal of Electronic Imaging (Special Section on Model-based Medical Image Processing and Analysis), Vol. 12, No. 1, pp. 59-68, January (2003).

NÉVÉOL, A ,SOUALMIA, LF ,DOUYÈRE M., ROGOZAN A., THIRION B., DARMONi, SJ.(2004) Using CISMeF MeSH encapsulated terminology and a rule-based algorithm for health resources categorization, to appear in International Journal of Medical Informatics.

RODRIGUES, JM., TROMBERT-PAVIOT, B., BAUD, R., WAGNER, J. and MEUSINET-CARRIOT, F.(1998) GALEN-In-Use : using Artificial Intelligence Terminology Tools to Improve the Linguistic Coherence of a National Coding System for Surgical Procedures. Cesnik et al. (eds). MedInfo'1998.

SOUALMIA, LF., BARRY-GRÉBOVAL, C., ABDULRAB, H., DARMONI, SJ. (2002) Modélisation et représentation des connaissances dans un catalogue de santé. Ingénierie des Connaissances 2002, p.139-149.

SOUALMIA, LF., BARRY C., DARMONI SJ. (2003) "Knowledge-Based Query Expansion over a Medical Terminology Oriented Ontology"; in M.Dojat, E.Keravnou and P.Barahona (Eds.): AIME 2003, Lecture Notes in AI#2780, Springer-Verlag, p.209-213.

SOWA, JF. (2000) Ontology, Metadata and Semiotics. Lecture Notes in AI #1867, Springer Verlag, p.55-81.