Recherche multi-terminologique de l'information de santé sur l'Internet

Saoussen Sakji*,**

**CISMeF, Centre Hospitalo-universitaire de Rouen*
Sakji.Saoussen @chu-rouen.fr
***GCSIS, LITIS EA 4108, Institut de recherche biomédicale, Université de Rouen.*

RÉSUMÉ. *La recherche d'informations et des connaissances médicales devient de plus en plus facile et accessible sur Internet pour le professionnel de santé, l'étudiant, mais aussi pour le patient et le cyber citoyen. CISMeF (Catalogue et Index des Sites Médicaux Francophones) est un outil visant à cataloguer et indexer les sources les plus importantes d'information de santé institutionnelles en France afin de les mettre à disposition du public. L'indexation des ressources Internet est mono-terminologique du fait qu'elle soit fondée exclusivement sur le thésaurus MeSH (traduit par l'US National Library of Medicine). En 2007, l'équipe CISMeF oriente ses objectifs vers un univers multi-terminologique qui s'appuie sur un extracteur automatique multi-terminologique et le développement préindustriel d'un serveur multi-terminologique médical. Le projet de recherche d'information multi-terminologique a débuté par l'intégration d'une terminologie complémentaire du MeSH concernant les substances chimiques et nous projetons d'intégrer les terminologies médicales françaises (CCAM) et celles traduites en français (CIM-10, SNOMED) afin d'améliorer la recherche d'information de CISMeF dans un contexte hétérogène.*

ABSTRACT. *The search for information and medical knowledge becomes increasingly easier and accessible on Internet not only for the health professionals and students, but also for patients and any Internet user. CISMeF (Catalogue and Index of Medical Sites in French) is a tool aiming at cataloguing and indexing the most significant sources of institutional health information in France and making them publicly available. The indexing of Internet resources is mono-terminological owing to the fact that is founded exclusively on the MeSH thesaurus (developed by the US National Library of Medicine). In 2007, the CISMeF team directed their objectives towards a multi-terminological universe based on a multi-terminological automatic extractor and the preindustrial development of a medical multi-terminological server. The multi-terminological information research project began by the integration of a terminology complementary to MeSH, concerning the chemical substances. and we project to integrate all the French medical terminologies (e.g. CCAM) and those translated into French (e.g. CIM-10, SNOMED)in order to improve the CISMeF information retrieval in a heterogeneous context.*

MOTSCLEFS : *recherche d'information, catalogue, Indexation, terminologies médicales.*
KEYWORDS : *information retrieval, catalogue, Indexation, medical terminologies.*

1. Introduction

La recherche d'information est un domaine historiquement lié aux sciences de l'information et à la bibliothéconomie qui ont toujours eu le souci d'établir des représentations des documents dans le but d'en récupérer des informations appropriées. Aujourd'hui, ce domaine est transdisciplinaire, ce qui devrait permettre de trouver des solutions pour améliorer son efficacité.

S'intéressant au domaine médical trouver l'information médicale la plus pertinente n'est pas une tâche aisée pour l'utilisateur ce qui nécessite donc de recourir à une classification spécialisée et hiérarchisée dans laquelle on peut naviguer afin d'avoir la réponse la plus adéquate, ainsi est né le CISMeF (Catalogue et Index des Sites Médicaux Francophones). Face à la diversité et la complexité de l'information pour le professionnel de santé et/ou les patients, CISMeF s'est orienté vers une indexation multi-terminologique qui constitue la base de notre travail pour la recherche multi-terminologique ce qui permet d'avoir une vision multithématique du résultat obtenu.

A travers cet article, nous présentons l'étendue de notre recherche, en commençant par l'état de l'art pour présenter le catalogue CISMeF, nous exposons par la suite, la problématique ainsi la contribution de notre travail. La quatrième section décrit une première partie réalisée de notre centre d'intérêt pour l'amélioration de la recherche d'information (en particulier le rappel). Pour conclure, nous présentons, dans la cinquième section, une discussion autour du thème et une conclusion et les perspectives, au niveau de la sixième section.

2. Etat de l'art

Créé en février 1995, dès la création du site Web du CHU de Rouen, le CISMeF (Darmoni et al, 2000) s'intéresse aux principaux sites et documents provenant des institutions francophones dans le domaine de la santé. En février 2008, il décrit et indexe environ 42 000 ressources

CISMeF inclut également, depuis l'année 2000 le système de recherche d'information Doc'CISMeF (Darmoni et al, 2001) qui jusqu'à 2006, était fondé sur un univers mono-terminologique où la recherche s'effectuait sur le seul thésaurus MeSH. Ce dernier est utilisé, notamment, pour indexer les articles scientifiques de la base de données bibliographiques MEDLINE (Clarke, 1997).

Une des particularités de CISMeF, outre son module pour les professionnels de santé, est de fournir des ressources pour les patients ainsi que pour les étudiants. Chaque ressource du catalogue est décrite et indexée par son contenant et son contenu en utilisant respectivement un ensemble de métadonnées et la terminologie CISMeF, qui englobe le thésaurus MeSH. Les métadonnées se réfèrent aux informations descriptives des ressources Web et décrites par dix éléments du Dublin

Core (DC) (Dekkers et al, 2003) dont les plus importantes sont : le titre, l'identifiant, la date, le contenu, le mot clef et le type de ressource. Pour plus de précision, huit métadonnées spécifiques (Darmoni et al, 2000) à CISMeF ont été ajoutées tels que : pays, institution…. La terminologie CISMeF encapsule la version française du thésaurus MeSH (Nelson et al, 2001) dans la mesure où, d'une part, elle représente une extension des concepts déjà existants et, d'autre part, elle emploie de nouveaux concepts. Dans le but de généraliser le MeSH aux ressources de santé sur Internet, des améliorations ont été réalisées depuis l'année 2000. En plus des descripteurs, des qualificatifs et des types de ressources MeSH (expansion de la liste déjà existante au niveau du MeSH), la notion de méta-termes (un concept innovant à la terminologie) a été ajoutée. La paire (descripteur/qualificatif) décrit le centre d'intérêt de la ressource. Les types de ressources sont une généralisation des types de publication de MEDLINE et sont utilisés afin de catégoriser la nature de la ressource. Les méta-termes sont, en général, des spécialités médicales ou des sciences biologiques, qui ont un lien sémantique avec un ou plusieurs termes MeSH, qualificatifs et types de ressources (Soualmia, 2006). L'ajout d'un tel concept permet d'avoir une vision globale sur le domaine médical et d'améliorer, par conséquent, la recherche d'information dans la mesure où les résultats retournés englobent plus d'informations et de ressources que si on cherche par mots clefs MeSH (Gehanno et al, 2007). Le processus de la recherche mono terminologique, basée sur un corpus indexé que par le thésaurus MeSH, commence, désormais, par vérifier la correspondance entre la requête de l'utilisateur et les termes présents dans le titre et les mots réservés (mot clef, qualificatif, type de ressource et méta terme), par la suite, en cas d'absence de réponses, on passe à une recherche sur toutes les métadonnées (i.e la description des ressources : auteur, mot réservé, date..) et finalement, en cas de zéro réponse à la deuxième étape, on effectue une recherche en texte intégral (le contenu des ressources). Un flag « interprétation de la requête » permet de mettre en relief le degré de pertinence des ressources retournées, selon les correspondances trouvées suite à l'application de l'algorithme « sac de mots » (Soualmia, 2006).

3. Problématique et contribution

Dès 2005, une décision stratégique de l'équipe CISMeF a permis le passage d'un univers mono-terminologique à un univers multi-terminologique par la mise au point d'un extracteur automatique multi-terminologique, le développement d'un Serveur Multi-Terminologique Médical (SMTM) qui rassemble sur un même portail les données des différentes terminologies : CISMeF, MeSH, SNOMED, CIM-10, CCAM, CISP2, DRC, CIF et la partie correspondante du réseau sémantique d'UMLS[1]. Cette union respecte la même structure d'origine basée sur la technologie du web sémantique notamment le langage OWL. Finalement, l'objectif de cette thèse est d'étudier et évaluer différents modèles, méthodes et outils de recherche d'information dans un univers multi-terminologique. Ce travail, en

collaboration avec le laboratoire LERTIM de Marseille, s'effectue au sein de l'équipe CISMeF au centre hospitalo-universitaire de Rouen.

L'indexation et la recherche des sources de données médicales étaient, dans un premier temps mono-terminologique du fait qu'elles soient basées que sur le thésaurus MeSH. Le but du passage vers un univers multi-terminologique, qui fait référence à plusieurs terminologies et ce dans le cadre de différents contextes médicaux, est d'améliorer le taux du rappel, sans pour autant perdre le niveau de la précision, ce qui permet d'étendre la finalité de notre outil de recherche d'information. Une telle approche, permet, outre la recherche documentaire en s'appuyant sur le thésaurus MeSH, la recherche des dossiers de santé électroniques via la nomenclature SNOMED, la recherche médicamenteuse par le biais des codes ATC, CIS, CIP.., l'analyse des données de mortalité et de morbidité recueillies dans différents pays grâce à la classification CIM-10.. En effet, le processus de recherche d'information peut être amélioré via ce contexte multithématique d'où notre contribution à ce stade est de modéliser toutes les terminologies suivant leurs structures d'origine afin de développer un nouveau moteur d'indexation et de recherche en santé fondé sur une structure multi-terminologique

4. Amélioration de la recherche d'information

4.1. *Univers mono-terminologique*

Pour la recherche d'information mono-terminologique, une méthode a été développée fondée sur l'expansion des requêtes des utilisateurs par leur enrichissement à l'aide d'éléments de connaissances. Cette approche générale combine l'exploitation des connaissances au système de recherche d'information et est fondée sur l'algorithme « sac de mots », qui permet de corriger, d'affiner et de préciser la requête de l'utilisateur pour que le système réponde au mieux à ses besoins informationnels (Soualmia, 2004).

Par ailleurs, pour être viable, le développement du catalogue CISMeF doit impérativement envisager une automatisation, au moins partielle, des tâches documentaires, qui peut être supervisée par les documentalistes aptes à contrôler la qualité de l'indexation. Cela implique donc de développer des outils informatiques capables de manipuler les documents. C'est suite à un tel constat que l'équipe CISMeF s'est orientée vers l'automatisation des tâches documentaires grâce au traitement des ressources pour la constitution, la maintenance, et la mise à disposition du catalogue. Il s'agit de la veille stratégique, de l'indexation et de la catégorisation de ressources de santé. (Névéol, 2005).

Pour cet univers, les travaux sont basés sur le thésaurus MeSH, qui constitue la seule terminologie employée par CISMeF, au cours du processus d'indexation, d'une part, et lors la recherche d'information d'autre part.

En 2007, la stratégie de l'équipe CISMeF s'oriente vers un univers multi-terminologique en se basant sur les terminologies traduites en français telles que la SNOMED, la CIM-10, les terminologies de santé françaises telle que la CCAM, ou encore les substances chimiques et médicamenteuses et les dispositifs médicaux. Un tel passage permet d'avoir un résultat diversifié selon plusieurs contextes médicaux.

4.2. *Univers multi-terminologique*

La réalisation d'un Serveur Multi-Terminologique Médical, représente le centre de l'univers multi-terminologique, et se fait par la constitution d'une base terminologique diversifiée, l'établissement de correspondances entre ces ressources (médiation sémantique), la normalisation par l'adoption d'un format pivot et la mise en œuvre d'outils de maintenance. Ainsi, les services offerts de façon interactive aux utilisateurs et aux applications clientes permettront d'aider les professionnels de santé pour indexer des documents, de natures différentes, avec la terminologie appropriée au type de document, et d'aider les utilisateurs à rechercher des informations et des connaissances. Le but d'un tel serveur, est de rendre interopérables les terminologies médicales francophones usuelles.

4.2.1. Amélioration de la recherche des substances

Pour être plus exhaustif et répondre au mieux aux requêtes des utilisateurs dans un univers multi-terminologique, le module **Action Pharmacologique** constitue un chapitre complémentaire du catalogue CISMeF en introduisant à la terminologie les substances chimiques ainsi que leurs codes correspondants et leurs synonymes. Se référant à la définition de la National Library of Medicine (NLM), une action pharmacologique est une catégorie d'actions chimiques et d'utilisations qui ont comme conséquence la prévention, le traitement, ou le diagnostic de la maladie. Sont inclus les produits chimiques qui agissent en changeant des fonctions normales du corps et les effets des produits chimiques sur l'environnement. Au niveau de la terminologie CISMeF, les actions pharmacologiques permettent de regrouper les substances ayant une même action. Au sein du thésaurus MeSH, les noms des substances peuvent correspondre soit à des mots-clés hiérarchisés (mc) soit à des *supplementary concept records* non hiérarchisés (sc). Ceci permet d'effectuer une recherche sur tous les termes MeSH ayant une action pharmacologique donnée.

Actuellement, la recherche à l'aide d'un terme MeSH qui correspond à une action pharmacologique se restreint aux substances qui ont ce terme MeSH comme action pharmacologique. Par exemple : Si on recherche vitamines en tant qu'**a**ction **p**harmacologique (ap) on obtiendra toutes les ressources indexées avec les termes MeSH correspondant à des substances dont l'action pharmacologique est « vitamines » tels que acétylcarnétine, acide folique, acide lipoïque...

Jusqu'à ce jour, les substances reliées aux actions pharmacologiques descendants, n'étaient pas reliées à l'action pharmacologique racine. Ainsi, les substances reliées à l'action pharmacologique « complexe vitaminique B », n'étaient pas reliées à l'action pharmacologique « vitamines » (cf. Figure 3). Ce qui revenait à considérer, par exemple, que la « vitamine B6 » n'est pas une vitamine.

Figure 3. *les descendants du mc « vitamines »*

Ainsi, pour faire face à cette incohérence les substances reliées à l'action pharmacologique « complexe vitaminique B » sont désormais reliées à l'action pharmacologique « vitamines ».
Formellement,

T : terme MeSH, action pharmacologique
S_T : les substances ayant comme action pharmacologique T
F : l'ensemble des fils du terme T
S_f : les substances ayant comme action pharmacologique f, $f \in F$
La liste exhaustive est sous la forme :

$$ListeS_T = S_T \cup S_f \quad \forall f \in F$$

Par ailleurs, au sein de la liste $ListeS_T$ (qui correspond donc à l'ensemble substances possédant l'action pharmacologique T), il ne faut considérer que les substances filles possédant la même action pharmacologique que la substance racine (si la substance est un mot-clé).

Par exemple, la kallicréine (mc) possède deux actions pharmacologiques : « coagulants » et « fécondostimulants masculin ». Pour autant, les descendants de la kallikréine, à savoir « antigène spécifique prostate », « kallicreine plasmatique », « kallicreines tissulaires », « kallicreinogene » ne possèdent pas forcément les mêmes actions pharmacologiques. D'ailleurs, l'antigène spécifique de la prostate n'est ni un coagulant ni un fécondostimulant masculin.

Ainsi, l'explosion des mots-clés ne doit pas s'appliquer aux substances (mc) reliées à une action pharmacologique donnée. En d'autres termes, il ne faut pas intégrer les descendants des mots clefs substances à la liste précédente.

$$\text{Liste}S_T = S_{T(\text{sans explosion})} \cup S_{f\,(\text{sans explosion})}$$

4.2.2. Amélioration de la recherche pour les médicaments

S'intéressant au domaine médicamenteux, la recherche peut être effectuée grâce au système de Classification Anatomique, Thérapeutique et Chimique (ATC) qui permet de classer les médicaments selon l'organe ou le système sur lequel ils agissent et /ou leurs caractéristiques thérapeutiques et chimiques. Les résultats de la recherche dépendraient du niveau du détail auquel on s'intéresse. Quant à la nomenclature CAS, elle permet d'identifier chaque produit chimique d'une manière unique auprès de la banque de données Chemical Abstract Service (CAS), ainsi dans un contexte toxicologique, par exemple, une recherche par code CAS permet de retrouver la substance chimique correspondante.

La recherche par code CIS (Code Identifiant de Spécialité) fait référence à la codification CIS qui identifie les spécialités pharmaceutiques faisant ou ayant fait l'objet d'une Autorisation de Mise sur le Marché (AMM). Un médicament peut être identifié par plusieurs numéros CIS, qui font référence à un dosage et/ou forme galénique différents. Pour le même code CIS, on peut avoir plusieurs codes CIP (Code Identifiant de Présentation) selon les différentes présentations de la spécialité pharmaceutique existantes (la taille et/ou le conditionnement). Une recherche d'un médicament pourrait être affinée, par conséquent, suite à la conjonction des deux codes.

5. Discussion

La recherche d'information en sante repose d'une part sur un corpus de ressources médicales, et d'autre part sur une indexation de celui-ci au moyen de thésaurus structurés et standardisés qui définissent son aspect multi-terminologique. Les difficultés majeures de notre travail se manifestent au niveau de l'intégration des terminologies complémentaires au thésaurus poly-hiérarchique MeSH en respectant leurs structures d'origines, d'une part, et en améliorant la recherche d'information, d'autre part.

6. Conclusion et perspectives

La recherche d'information médicale sur le web est en perpétuelle gestation et ne cesse de se développer afin de répondre aux besoins des utilisateurs notamment les professionnels de santé ou encore les patients. A cette fin, notre travail consiste à améliorer la structure de recherche de CISMeF en élargissant les thématiques dans le cadre d'un univers multi-terminologique comprenant les thésaurus les plus utilisés dans le domaine médical. Certes le besoin d'avoir l'information selon un contexte spécifique se ressent avec l'expansion de l'information médicale sur Internet, cependant, il faut s'assurer que ceci n'influencerait pas négativement sur le

rappel, et si ca serait le cas il faut trouver les méthodes appropriées pour remédier à cette contrainte.

7. Bibliographie

Clarke M., Greaves L., James S., « MeSH terms must be used in Medline searches ». BMJ.1997 Apr 19;314(7088):1203.

Darmoni SJ., Leroy JP., Thirion B., et al. « CISMeF a Structured Health Resource Guide », *Methods Inf. Med*. 39(1):30-35, 2000.

Darmoni SJ., Thirion B., Leroy JP., Douyere M., Lacoste B., Godard C., Rigolle I., Brisou M., Videau S., Goupy E., Piot J., Quéré M., Ouazir S., Abdulrab H., « A search tool based on 'encapsulated' MeSH thesaurus to retrieve quality health resources on the Internet ». *Medical Informatics & The Internet in Medicine* 2001; 26(3):165 - 178.

Dekkers M., Weibel S., « State of the Dublin Core Metadata Initiative », *D-Lib Mag*. 2003 V9 N° 40.

Gehanno JF., Thirion B., Darmoni SJ., « Evaluation of Meta-concepts for Information Retrieval in a Quality-Controlled Health Gateway. AMIA Symp ». (in press).

Nelson SJ., Johson WD., Humphreys BL., « Relationship in Medical Subject Heading », *Kluwer Publishers*, 2001: pp171-84.

Névéol A., Automatisation des tâches documentaires dans un catalogue de santé en ligne. Thèse en informatique de l'INSA de Rouen. 2005.

Soualmia L., Etude et évaluation d'approches multiples d'expansion de requêtes pour une recherche d'information intelligente : application au domaine de la santé sur l'internet. Thèse en informatique de l'INSA de Rouen. 2004.

Soualmia LF., Dahamna B., Thirion B., Darmoni SJ., « Some Strategies for Health Information Searching ». *MIE 2006, Twentieth International Congress of the European Federation for Medical Informatics. Stud Health Technol Inform*. 2006;595-600