CISMeF dans l'Infrastructure du Web Sémantique

LF.Soualmia[a,b], A.Névéol[a,b], B.Dahamna[b],

M. Douyère[b], JP. Leroy[b], B. Thirion[b] et SJ. Darmoni[a,b]

[a] *Laboratoire Perceptions, Systèmes, Information, FRE 2645 CNRS, INSA de Rouen.*

[b] *Equipe CISMeF, CHU de Rouen.*

Abstract

We describe the CISMeF catalogue and index of health Internet resources, which have been designed for Information Retrieval in the domain of medicine. The main issues to describe the catalogue quality-controlled resources are the modeling of a terminology which « encapsulates » the MeSH thesaurus and a set of metadata. The catalogue content is evolving from HTML resource format passing by XML, which enables interoperability with other systems, and recently to RDF format, which is a basic Semantic Web language.

Mots-clés :

Terminologie ; Métadonnées ; Recherche d'Information ; Indexation Automatique ; Web Services.

1. Introduction

Etant donnée la quantité d'information disponible sur le Web, celui-ci doit faire face aux problèmes d'exhaustivité et de précision en recherche d'information. Le catalogue CISMeF[1] a été développé afin de faciliter l'accès à l'information de santé disponible sur l'Internet selon trois axes (médecine factuelle, ressources pédagogiques, ressources pour les patients et leur famille). La problématique d'aujourd'hui se veut aussi être une recherche d'information « intelligente » dans l'infrastructure du Web Sémantique[2], une extension du web actuel qui permettrait de rendre compréhensible le contenu des ressources par les hommes mais aussi par les machines. Un principe basique du Web Sémantique est de décrire les ressources d'information à l'aide de marqueurs exploitables par différents logiciels, par exemple des métadonnées au format RDF[3], pour guider une recherche.

2. Structure du catalogue CISMeF

La terminologie CISMeF : La terminologie CISMeF est composée des mots clés et des qualificatifs du MeSH. Un ensemble de métatermes (spécialités médicales) et une hiérarchie de types de ressources ont été rajoutés afin d'optimiser la recherche d'information dans le catalogue par une expansion automatique des requêtes dans les arborescences de mots clés, de qualificatifs et de types de ressources, mais également pour permettre une vision plus globale concernant une spécialité, ceci n'étant pas possible au niveau du MeSH.

Les métadonnées CISMeF : La recherche d'information est la première utilité des métadonnées[4]. Elles sont également essentielles pour l'interopérabilité (celle-ci a été testée avec succès sur la plate-forme d'e-learning du Campus Viruel d'Archimède). Chaque ressource est décrite par onze des quinze éléments du Dublin Core ainsi que huit autres éléments spécifiques à CISMeF. Pour les ressources pédagogiques les onze éléments de la catégorie « Educational » du format IEEE 1484, pour les ressources de médecine factuelles, CISMeF a défini les champs *indication du niveau de preuve* et la *méthode* pour le déterminer, enfin pour qualifier la qualité de l'information de santé, le langage HIDEL du projet européen MedCIRCLE (CISMeF devient tiers de confiance explicite). Le format de ces métadonnées est passé du langage HTML en 1995, au langage XML en septembre 2000 pour permettre l'interopérabilité et depuis décembre 2002 à RDF. RDF est un langage simple doté d'une syntaxe XML et il est aussi à la base du Web Sémantique pour la représentation des métadonnées dans un contexte de recherche d'information. Il a été prouvé dans la pratique que RDF améliorait entre autres les performances des moteurs de recherche par mot clé et la recherche par navigation [5].

3. Recherche d'Information : thèse L.Soualmia

Nous évaluons à une échelle réelle, en terme de recherche d'information dans le catalogue CISMeF, les améliorations possibles de trois approches issues de domaines différents ainsi que leur complémentarité.

Traitement Automatique du Langage Naturel : Une requête peut employer des termes proches mais pas nécessairement identiques à ceux de la terminologie. Le MeSH ne comporte ni variations morpho-syntaxiques ni termes synonymes (les termes synonymes du MeSH ne correspondent pas aux termes en usage courant par les cybercitoyens). Nous proposons d'utiliser une base de connaissances morphologiques de la terminologie CISMeF et de faire une étude des logs (pour déterminer les termes synonymes) pour une réécriture automatique des requêtes. Ce module s'inscrit dans le projet UMLF[1] (équivalent du Specialist Lexicon de l'UMLS).

Bases de Données : Nous proposons de découvrir de nouvelles règles d'association entre termes (à partir de la base de données CISMeF) pour une expansion automatique des requêtes.

Intelligence Artificielle : Nous proposons de modéliser sous forme de règles sémantiques une partie du réseau sémantique de l'UMLS pour permettre un raisonnement sur le contenu des ressources et d'effectuer des inférences grâce à des outils comme TRIPLE [6].

4. Indexation Automatique : thèse A.Névéol

A l'heure actuelle, la recherche de nouvelles ressources à intégrer dans CISMeF et l'indexation de ces ressources sont des taches effectuées manuellement par l'équipe de documentalistes de CISMeF. Nous proposons d'implémenter un système de veille et d'indexation semi-automatique qui, supervisé par les documentalistes, permettrait d'élargir la couverture du catalogue CISMeF tout en maintenant la qualité de l'indexation. Nous nous appuyons pour cela sur des méthodes relevant de plusieurs domaines:

Traitement Automatique du Langage Naturel : Utilisation d'automates dictionnaires pour la reconnaissance des mot-clefs et qualificatifs Mesh et de leurs synonymes. *Apprentissage:* Détermination de règles d'association à partir du corpus indexé des 10.000 ressources CISMeF à notre disposition.

5. Web Services

L'équipe CISMeF a développé un Web Service permettant l'intégration des métadonnées extraites de CISMeF par une requête quelconque afin d'être intégré avec n'importe quelle application (site Web, dossier électronique du patient, méta-catalogue).

6. Perspectives

Dans les années à venir, l'équipe souhaite travailler sur la navigation sémantique, plus riche que la navigation hiérarchique actuelle, ainsi que sur la création d'un méta-catalogue en partenariat avec le LERTIM de Marseille.

Références

1. Darmoni SJ, Thirion B, Leroy JP et al. A Search Tool Based on 'Encapsulated' MeSH Thesaurus to Retrieve Quality Health Resources on the Internet. *Medical Informatics & the Internet in Medicine*, vol 26 n°3, p.165-178, 2001.
2. Berners-Lee T., Heudler J., Lassila O. The Semantic Web. *Scientific American 2001*. <http://www.sciam.com/2001/0501issue/0501berners-lee.html>
3. Lassila O., Swick R. Resource Description Framework (RDF) Model and Syntax Specification. *W3C Candidate Recommendation 1999*. <http://www.w3.org/TR/REC-rdf-syntax>

[1] Lexique Unifié du Français Médical

4. Laublet P., Reynaud C., Charlet J. Sur quelques aspects du Web Sémantique. *Actes des deuxièmes assises nationales du GdRI3* , pp.59-78, 2002.
5. Ossenbruggen JR., Hardman HL., Rutledge L. Hypermedia and the Semantic Web : A Research Agenda. *Tech.Rep. INS-R0105, 2001*.
6. Sintek M, Decker S. TRIPLE- An RDF Query, Inference and Transformation Language. *Proceedings of the Deductive Databases and Knowledge Management Workshop*, 2001.