Caractérisation des contenus de l'Internet en santé : l'exemple CISMeF

Aurélie Névéol[1,2], Lina F. Soualmia[1,2], Alexandrina Rogozan[2], Magaly Douyère[1], Benoît Thirion[1], Stéfan J. Darmoni[1,2]

[1] *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

[2] *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 présentons ici une méthode de caractérisation de ressources de l'Internet utilisée dans le catalogue de santé CISMeF. Nous expliquons l'intérêt de cette caractérisation pour les différents types d'utilisateurs, et explicitons les techniques employées pour rendre ces informations accessibles et donc exploitables par l'homme et par la machine. Les travaux en cours au sein de l'équipe CISMeF s'orientent maintenant vers la description et l'indexation automatique des ressources et l'exploitation des données de caractérisation pour la recherche d'information.

1. Contexte

A l'heure actuelle, Internet est une source d'information importante dans tous les domaines, et en particulier celui de la santé. Les utilisateurs rencontrent d'énormes difficultés pour trouver précisément ce qu'ils cherchent dans la pléthore de documents mis à leur disposition. Les moteurs de recherche généralistes comme Google restent impuissants à résoudre ce problème car ils proposent souvent une sélection de documents trop large, ou encore mal ciblée. De plus, les utilisateurs sont livrés à eux même pour évaluer la qualité et le degré de confiance des documents qu'ils consultent. Dans ce contexte, le catalogue CISMeF (Catalogue et Index des Sites Médicaux Francophones) créé en 1995, répertorie et indexe les ressources d'information institutionnelles de santé en langue française afin d'y permettre un accès rapide et précis [1]. Les ressources indexées par CISMeF sont d'une grande diversité, tant au niveau des types de documents sélectionnés (recommandations de pratique clinique, cours, informations pour les patients, …) que de leur format (site ou page Web, document pdf, …). Le catalogue contient à l'heure actuelle 13,642 ressources, et il est mis à jour au rythme de 50 nouvelles ressources en moyenne indexées chaque semaine. L'ajout d'une nouvelle ressource au catalogue s'effectue en quatre étapes: le recensement des ressources potentielles par une veille quotidienne, la sélection des ressources selon des critères de qualité précis, la description et l'indexation, et la mise en ligne de notices descriptives.

2. Notice CISMeF

Contenu d'une Notice CISMeF

Les notices CISMeF contiennent plusieurs types d'information :

- Une présentation contenant des informations générales sur le contenu et la qualité de la ressource : le titre, le nom du ou des auteurs, un résumé succinct, la source, le niveau de preuve, le type de ressource.
- Une classification contenant des informations détaillées sur le contenu de la ressource : la liste des spécialités médicales, et des mots clés (ou paires mot clé / qualificatif) MeSH. Le MeSH est le thésaurus de référence du domaine biomédical, développé par la National Library of Medecine américaine pour la base documentaire Medline [2]. La terminologie CISMeF « encapsule » le MeSH avec les concepts de « métatermes » et « type de ressource » détaillés dans [3]
- Des informations pratiques sur la ressource: l'URL, le format, la langue, le type d'accès (libre, restreint, payant), la date de consultation...

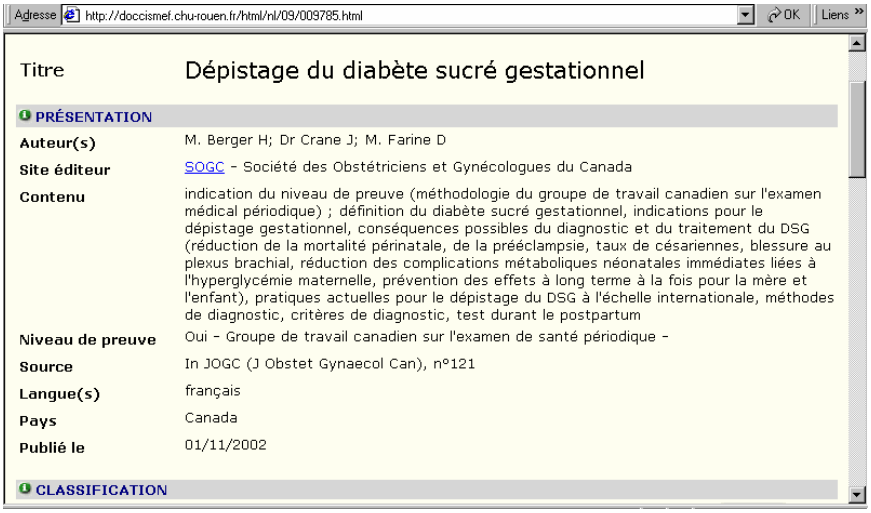

Fig 1. Extrait de la notice CISMeF n° 9785

Intérêt pour l'utilisateur

Le concept de métadonnée est apparu bien avant l'Internet, mais son intérêt a été décuplé par le nombre croissant de publications électroniques. A l'initiative du World-Wide-Web Consortium[1] (W3C), les métadonnées sont utilisées pour décrire le contenu des pages Web. En utilisant un formalisme approprié, les métadonnées permettent donc une caractérisation non ambiguë des ressources, et facilitent la recherche d'information dans les pages concernées.

La liste des champs à renseigner dans la notice CISMeF a été établie à l'aide des critères de qualité du Net Scoring[2] et des métadonnées du Dublin Core (DC) [4] pour les champs *auteur*, *date*, *description*, *format*, *identification*, *langue*, *éditeur*, *type de ressource*, *droits*, *sujet* et *titre*. Pour décrire les ressources pédagogiques, onze éléments de la catégorie "Education" du IEEE 1484 LOM (Learning Object Metadata) sont utilisés. Par ailleurs, des métadonnées spécifiques à CISMeF, ont été ajoutées pour décrire la qualité ou la localisation de la ressource : *institution*, *ville*, *province*, *pays*, *type d'accès*, *partenariat*, *coût* et *public ciblé*. Deux champs supplémentaires ont été créés pour les ressources destinées aux professionnels de santé: *indication du niveau de preuve* et la *méthode* utilisée pour l'établir [5]. Les métadonnées HIDDEL[3] ont été introduites dans CISMeF dans le cadre du projet européen MedCircle [6], qui a pour but d'évaluer la qualité de l'information de santé afin de guider les utilisateurs vers des sources fiables.

L'objet de cette démarche est de fournir des informations sur le contenu et la qualité d'une ressource de manière synthétique. La santé est l'un des domaines ou la qualité et la fiabilité des informations consultées sont les plus cruciales. Il est donc important de sensibiliser l'utilisateur aux critères qui peuvent attester de la validité des informations données. Les notices très détaillées de CISMeF fournissent les éléments de réponse nécessaires.

Accessibilité par l'Homme et par la Machine

Au début du projet CISMeF le standard HTML 2.0 était utilisé pour que le site soit lisible par la grande majorité des navigateurs. Le standard plus récent HTML 4.0 est maintenant employé. Depuis juin 2001, l'utilisation du standard XML permet une interopérabilité avec d'autres catalogues ou serveurs de ressources (e-learning dans le cadre du projet UMVF[4]). Depuis décembre 2002, CISMeF a adopté le format RDF[5] (into HTML) : les ressources sont maintenant décrites au format RDF d'après les concepts de l'ontologie HIDDEL.

3. Saisie Automatique des Notices

Saisie actuelle

[1] http://www.w3c.org/
[2] http://www.chu-rouen.fr/netscoring/
[3] High Information Description Disclosure Evaluation Language - cf. http://www.merdcircle.org/
[4] Université Médicale Virtuelle Francophone - cf. http://www.umvf.prd.fr/
[5] http://www.w3.org/RDF/

A l'heure actuelle, la plupart des champs de la notice sont saisis manuellement par les documentalistes de l'équipe CISMeF. Cependant, la classification en spécialités médicale est générée automatiquement à partir du contenu des champs « mots clés » et « type de ressource » grâce aux propriétés de la terminologie CISMeF [3] et à un algorithme de classification décrit dans [7]. 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, signalé dans la notice par des étoiles.
Par ailleurs, dans le cadre de *pré-CISMeF*, les auteurs ou éditeurs qui proposent régulièrement des ressources à ajouter au catalogue sont encouragés à transmettre une notice au format XML remplie par leur soins avec les métadonnées pour la ressource proposée. Dans l'attente d'une vérification et d'une validation des informations par l'équipe CISMeF, ces ressources sont automatiquement intégrées dans pré-CISMeF.

Saisie automatique des Mots Clés
Le temps passé à remplir une notice est en grande partie consacré à l'indexation (élaboration de la liste de mots clés). Ainsi, un système d'indexation automatique est en cours de réalisation dans le cadre d'une thèse [AN] afin d'alléger le travail des documentalistes. Ce système devra correspondre au cahier des charges de l'indexation manuelle, c'est à dire 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. De plus, le nombre de mots clés (ou paires) associé(e)s à une ressource n'est pas fixé à l'avance; il peut aller de zéro (pour les sites des hôpitaux par exemple) à plusieurs dizaines. Par ailleurs, la notion de descripteur obligatoire (*check tag*) doit être prise en compte. Les informations contenues dans la terminologie CISMeF 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 hiérarchiques 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. 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 à l'aide du logiciel INTEX[6]. Ces éléments textuels comprennent les flexions (et parfois des dérivations) des mots MeSH, mais également les synonymes de ces mots, contenus dans la terminologie ou dans le lexique développé dans le cadre du projet UMLF[7]. Ainsi, l'expression *femme enceinte* sera extraite, puis rapportée au mot clef *grossesse*. De même, certains éléments textuels peuvent être rapportés à des paires mot clé/qualificatif : l'expression générique *vaccin contre la maladie M* sera extraite puis rapportée à la paire *maladie M* / *prévention et contrôle.* Afin de prendre en compte les éléments graphiques contenus dans les documents, une thèse va débuter en janvier 2004 sur l'indexation automatique d'images et la recherche combinée texte-image.

Saisie automatique des autres champs
Le contenu des champs de la notice autres que « mots clés » et « type de ressource » peut généralement être extrait directement de la ressource. Cependant, les informations sont présentées de manière plus ou moins explicite. En effet, certaines ressources mentionnent fort à propos en en-tête ou en première page le titre, le nom du (des) auteur(s), ainsi que la date de publication. Pour d'autres documents, le nom de l'auteur se confond avec celui de l'éditeur du site, la date de publication doit être déduite de l'URL, certaines informations sont résumées par un logo… Ces particularités, ainsi que l'hétérogénéité des ressources traitées, font que la caractérisation de ressources de santé met en jeu une véritable "connaissance métier", et constitue un problème très complexe pour la saisie automatique des différentes données. De nombreux travaux abordent le problème de la saisie automatique de données semi-structurée à partir de textes libres. Dans le domaine de la santé, on peut

[6] Développé par M. Silberztein - cf. http://www.nyu.edu/pages/linguistics/intex/
[7] cf. www-test.biomath.jussieu.fr/umlf/

notamment mentionner [8] et le projet « Princip Health » actuellement en préparation en partenariat avec l'INaLCO pour attribuer automatiquement un indice de qualité aux ressources de santé sur l'Internet.

4. Exploitation du contenu des Notices : la Recherche d'Information

Différents modes de recherche d'information sont possibles: La recherche dite « simple » permet à l'utilisateur de saisir une requête (un terme ou une expression) en texte libre en français ou en anglais avec ou sans accent en majuscule ou en minuscule. La recherche dite « avancée » permet des requêtes plus pointues à l'aide d'un formulaire contenant des listes déroulantes et permet de combiner plusieurs champs des métadonnées (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. Elle est principalement destinée aux bibliothécaires médicaux.
La recherche « simple » telle qu'en place aujourd'hui est la plus utilisée (75%). Elle se fonde sur les relations hiérarchiques entre mots clés de la terminologie. 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.
Les résultats sont affichés en fonction de la date de publication, de la plus récente à la moins récente pour permettre aux utilisateurs d'avoir des informations les plus "fraîches" possibles. Par exemple une ressource datant de 1996 va indiquer 3 niveaux de gravité de l'asthme chez l'enfant alors qu'une ressource datant de 2002 va en indiquer 4. Si le terme saisi par l'utilisateur n'est pas un terme réservé, une recherche sur tous les autres champs de métadonnées est effectuée. Si le système est silencieux, une recherche en texte intégral sur tous les documents indexés est réalisée.
Un autre élément important est le *type de ressource*, qui permet de savoir à qui est destiné le document. Pour un professionnel de santé ce sera le type de ressource *ligne directrice et consensus,* pour un étudiant en médecine ce sera *enseignement et éducation* et enfin pour les néophytes ce sera *patient.*
Des travaux en cours dans le cadre d'une thèse (LS) s'orientent vers l'amélioration du moteur de recherche pour permettre une *recherche d'information intelligente* et ont conduit au développement du système KnowQuE (Knowledge-based Query Expansion) qui se fonde sur l'utilisation conjointe d'une base de connaissances morphologiques [9] 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).

Références

[1] Darmoni SJ, Leroy JP, Thirion B, Baudic F, Douyère M, Piot J. (2000) CISMeF: a structured Health resource guide. - *Methods of Information in Medicine,* Jan;39(1) 30-

[2] National Library of Medicine. Fact Sheet Medline. 6 July 1998 [Document Web, accédé le 3 Déc 2003] URL: http://www.nlm.nih.gov/pubs/factsheets/medline.html

[3] Soualmia LF, Barry-Greboval C, Abdulrab H, Darmoni SJ. (2002) Modélisation et représentation des connaissances dans un catalogue de santé. *IC'2002*, pp 139-149.

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

[5] Thirion B, Loosli G, Douyere M, Darmoni SJ. (2003) Metadata element set in a quality-controlled subject gateway: a step to a health semantic Web. *Medical Informatics Europe*, pp ?

[6] Mayer, MA., Darmoni, SJ., Fiene, M. et al. (2003) MedCIRCLE - Modeling a Collaboration for Internet Rating, Certification, Labeling and Evaluation of Health Information on the Semantic World-Wide-Web. *Medical Informatics Europe* p.667-672

[7] Névéol A., Soualmia LF., Douyère M., Rogozan A., Thirion B., Darmoni SJ. (2003) Using CISMeF MeSH Encapsulated Terminology and a Categorization Algorithm for Health Resources. *International Journal of Medical Informatics*, à paraître.

[8] Bekhouche D. (2003) Extraction Sémantique des Données de Patients dans un Réseau de Soins en Cancérologie. *Recueil des journées doctorant PSI*, pp 5-14.

[9] Grabar N., Zweigenbaum P., Soualmia LF., et Darmoni SJ. (2003) Matching Controlled Vocabulary. *Medical Informatics Europe*, p.445-450.