Indexation automatique de ressources de santé à l'aide d'un vocabulaire contrôlé

Aurélie Névéol

Laboratoire PSI – INSA et Université de Rouen
BP8 - Avenue de l'Université - 76801 Saint-Etienne-du-Rouvray Cedex
aneveol@insa-rouen.fr

Equipe CISMeF et L@STICS – Faculté de Médecine - CHU de Rouen
1, rue de Germont – 76031 Rouen

Résumé – Abstract

Nous présentons ici le système d'indexation automatique actuellement en cours de développement dans l'équipe CISMeF afin d'aider les documentalistes lors de l'indexation de ressources de santé. Nous détaillons l'architecture du système pour l'extraction de mots clés MeSH, et présentons les résultats d'une première évaluation. La stratégie d'indexation choisie atteint une précision comparable à celle des systèmes existants. De plus, elle permet d'extraire des paires mot clé/qualificatif, et non des termes isolés, ce qui constitue une indexation beaucoup plus fine. Les travaux en cours s'attachent à étendre la couverture des dictionnaires, et des tests à plus grande échelle sont envisagés afin de valider le système et d'évaluer sa valeur ajoutée dans le travail quotidien des documentalistes.

This paper presents the automatic indexing system currently developed in the CISMeF team to assist human indexers. The system architecture, using the INTEX platform for MeSH term extraction is detailed. The results of a preliminary experiment indicate that the automatic indexing strategy is relevant, as it achieves a precision comparable to that of other existing operational systems. Moreover, the system presented in this paper retrieves keyword/qualifier pairs as opposed to single terms, therefore providing a significantly more precise indexing. Further development and tests will be carried out in order to improve the coverage, and validate the efficiency of the system in the librarians' everyday work.
.

Mots Clés – Keywords

Indexation Automatique, Terminologie Médicale, Vocabulaire Contrôlé.
Automatic Indexing, Medical terminology, Controlled Vocabulary.

1 Contexte de travail

1.1 Le Catalogue et Index des Sites Médicaux Francophones (CISMeF)

L'Internet est devenu une source abondante de documents de toutes sortes, en particulier dans le domaine de la santé. 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 l'Internet. Le catalogue décrit et indexe les principales ressources institutionnelles de santé en français. A ce jour, 13 642 ressources ont été sélectionnées par une équipe de documentalistes et la mise à jour du catalogue s'effectue au rythme de 50 nouvelles ressources par semaine. Une ressource peut être un site ou une page Web, un cours, un article, 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 fondant sur un ensemble de métadonnées et une terminologie structurée semblable à une ontologie documentaire du domaine médical détaillée dans (Soualmia et al., 2002). Un exemple de notice est donné figure 1. Lors de l'élaboration d'une notice, la tâche la plus coûteuse en temps est l'indexation de la ressource à l'aide de mots clés. Crucial pour la recherche d'information dans le catalogue, ce travail met en jeu de solides connaissances métier. Ainsi, étant donné le nombre croissant de ressources mises en ligne chaque jour, il est nécessaire de pouvoir disposer d'outils automatiques pour l'indexation. De nombreux travaux ont été réalisés dans ce sens pour traiter des textes de tous domaines comme le système FASTR (Jacquemin, Royauté, 1994) ou plus spécifiquement dédiés à la santé, comme les projets Indexing Initiative (Aronson et al. 2000) et Indexing Aid (Humphrey, Miller 1987), ou les systèmes HONselect (Gaudinat, Boyer 2002) et NOMINDEX (Pouliquen 2002), ou encore des combinaisons de systèmes d'indexation (Ruch et al. 2003).

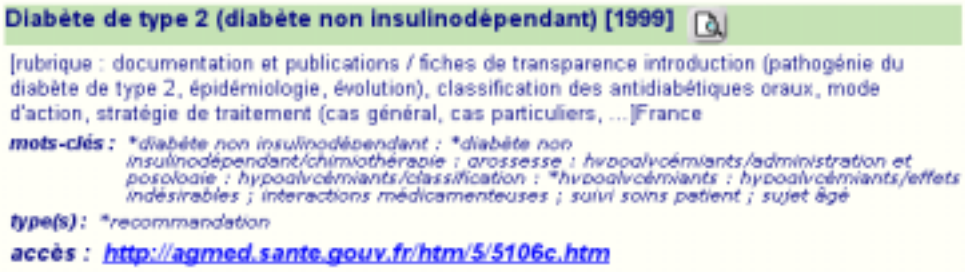

Figure 1 : Notice courte CISMeF pour la ressource « Diabète de type 2 »

1.2 Objectif des travaux présentés

L'objectif de nos travaux est de développer un système d'indexation automatique qui permettrait d'étendre la couverture du catalogue tout en maintenant une indexation de qualité, et en assurant des taux de précision et de rappel élevés lors de la recherche d'information dans CISMeF. Pour ce faire, le système d'indexation automatique doit impérativement intégrer les normes d'indexation manuelle en vigueur. D'après les résultats d'une évaluation du système NOMINDEX (Pouliquen 2002) réalisée sur un échantillon représentatif des ressources du catalogue, on se fixe comme objectif d'extraire les mots clés les plus importants, et de réduire le bruit autant que possible, même si cela implique de ne pas proposer une indexation

exhaustive. En effet, ce système est destiné à aider l'indexation manuelle, afin de réduire les délais actuels, tout en validant la qualité de l'indexation automatique proposée. Le système ne constituera pas une alternative à l'indexation manuelle. Cependant, tandis que les systèmes d'indexation cités précédemment en 1.1 permettent d'extraire des mot clés et des qualificatifs isolés, nous souhaitons porter une attention particulière sur l'extraction de *paires* mot clé/qualificatif, ce qui est indispensable pour produire une indexation précise - et à notre connaissance innovant.

2 Indexation des ressources dans CISMeF

2.1 Normes d'indexation

Les 13 000 ressources actuellement référencées dans le catalogue CISMeF ont été indexées manuellement par les documentalistes de l'équipe selon des critères très spécifiques, fondés sur les normes d'indexation de la base de données bibliographiques Medline[1]. Ainsi, à chaque ressource est associée une liste de termes (mots clés ou paires mot clé/qualificatif) empruntés au MeSH (Medical Subject Headings), le thésaurus de référence du domaine biomédical développé par la NLM (National Library of Medicine) américaine. Plus précisément, CISMeF étant un catalogue francophone, c'est la traduction française des termes MeSH établie par l'Institut National de la Santé et de la Recherche Médicale (INSERM) qui est utilisée[2]. Le MeSH dans sa version 2003 est composé d'environ 22 000 mots clés (exemples: *grossesse*, *diabète*) qui peuvent être associés à l'un des 84 qualificatifs (exemples: *diagnostic*, *prévention & contrôle…*) afin de préciser leur sens. Ainsi, une indexation utilisant des paires mot clé/qualificatif permet une description plus précise qu'une liste de termes isolés. Par exemple, dans une ressource détaillant les mesures prises par les autorités pour le dépistage du sida, il est plus pertinent d'utiliser la paire *sida/prévention & contrôle* que le mot clé isolé *sida* pour l'indexation. De cette manière, les utilisateurs intéressés par les aspects autres que *prévention & contrôle* peuvent d'emblée savoir qu'ils doivent consulter une autre ressource.

Afin d'être encore plus précis dans la description d'une ressource, un poids majeur est associé aux mots clés (ou paires mot clé/qualificatif) si le concept représenté est traité tout au long de la ressource de manière détaillée, et mineur si le concept est peu détaillé, ou évoqué seulement dans une partie de la ressource. L'indexation prend également en compte les descripteurs obligatoires, ou *check tags*. Il s'agit de mots clés MeSH sélectionnés par la NLM comme étant des termes prioritaires pour l'indexation. CISMeF utilise tous les *check tags* relatifs à la médecine humaine.

De plus, il faut remarquer que le nombre de mots clés (ou paires) à retenir pour l'indexation d'une ressource n'est pas fixe. Il peut varier de zéro (pour les sites des hôpitaux par exemple) à plusieurs dizaines.

---

1 cf. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi

2 cf. http://disc.vjf.inserm.fr:2010/basismesh/mesh.html

Toutes les caractéristiques évoquées ci-dessus doivent donc être prises en compte dans la conception du système d'indexation automatique. Nous allons maintenant décrire comment elles ont été intégrées dans notre système expérimental.

2.2 Description du système d'indexation automatique CISMeF

Après une présentation globale du système, les trois grandes étapes de son fonctionnement seront détaillées : l'identification d'éléments textuels, le mapping vers les termes MeSH et l'exploitation des bases MeSH et CISMeF, le calcul du score et la sélection de l'index final.

2.2.1 Présentation globale du système

Comme le montre la figure 2, le système d'Indexation Automatique CISMeF intègre la plateforme linguistique INTEX (Silberztein 1993). INTEX est un outil d'analyse de corpus puissant, dont les différentes fonctionnalités peuvent être intégrées dans d'autres systèmes. L'Indexeur utilise également deux types de bases de données: un dictionnaire MeSH, et trois bases de connaissances contenant des informations sur les relations hiérarchiques entre les termes MeSH, la liste des descripteurs obligatoires (*check tags*), et un historique des associations mot clé/qualificatif tiré des notices CISMeF.

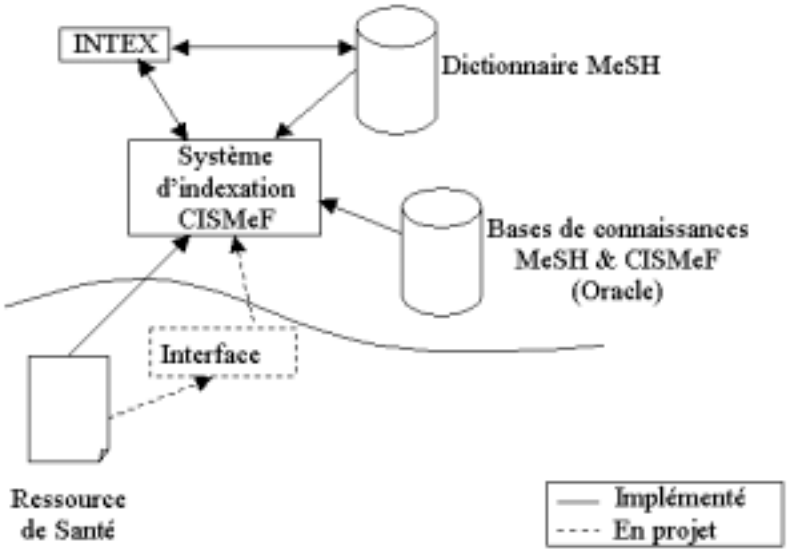

Figure 2 : Architecture du système d'indexation automatique CISMeF

2.2.2 Identification d'éléments textuels

La première étape de l'indexation automatique présentée ici est l'identification d'éléments textuels utiles à l'indexation qui sont répertoriés dans un dictionnaire MeSH. Ces éléments textuels peuvent être des termes (mots clés ou qualificatifs) MeSH (par exemple, *sujet âgé* ou *diagnostic*), des formes fléchies (par exemple, "sujets âgés") ou dérivées (par exemple, "diagnostiquer") de ces termes, des synonymes (par exemple, "personne âgée" est un synonyme MeSH de *sujet âgé*) ou des formes fléchies de ces synonymes (par exemple, "personnes âgées"). La taille importante du dictionnaire nécessaire (rappelons que le MeSH compte plus de 22 000 mots clés) nous a conduits à envisager l'utilisation d'automates pour extraire ces éléments textuels. En effet, les automates à états finis peuvent être utilisés pour la

détection des entrées d'un dictionnaire, et ce en un seul parcours du texte (Crochemore et al. 2001). De plus, le temps d'exécution requis ne dépend pas de la taille du dictionnaire, mais est proportionnel à la longueur du texte analysé. Cette technique est couramment utilisée dans de nombreuses applications du TAL. Le logiciel INTEX propose une boîte à outils complète des fonctionnalités qui nous intéressent – c'est pourquoi nous avons décidé de l'intégrer à notre système. INTEX permet de manipuler des dictionnaires au format DELA (Courtois, Silberztein, 1990). Ainsi, le dictionnaire MeSH utilisé dans le système d'indexation automatique est composé de deux dictionnaires: un au format DELAF pour les mots simples, et un au format DELACF pour les mots composés.

Un certain nombre de règles d'indexation ont été mises au point avec le documentaliste "superindexeur", qui supervise l'indexation de toutes les ressources du catalogue CISMeF. Ces règles permettent d'extraire des paires mot clé/qualificatif à partir d'expressions récurrentes. Par exemple, à partir de l'expression "dépistage de la maladie M" on peut extraire la paire *maladie M/prévention & contrôle*. Ces règles sont implémentées sous forme de grammaires locales, à l'aide de graphes INTEX. La figure 3 donne le graphe associé à la règle évoquée dans l'exemple ci-dessus. Suite aux résultats de l'évaluation décrite dans (Gaudinat et Boyer 2002), des grammaires similaires sont utilisées pour extraire efficacement les mots clés liés aux groupes d'âge (par exemple, *sujet âgé* ou *adulte âge moyen*).

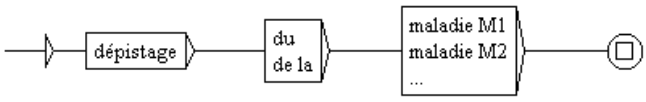

Figure 3 : Grammaire locale extrayant les paires du type *maladie M/prévention & contrôle*

2.2.3 Mapping vers les termes MeSH et exploitation des bases de connaissance

La deuxième étape consiste à rapporter les éléments textuels repérés aux mots clés MeSH correspondants, afin de créer une liste de termes MeSH indiquant la position de chaque occurrence dans le texte. L'information sur les positions respectives des termes MeSH identifiés dans la ressource permet de proposer des associations mot clé/qualificatif non répertoriées dans les grammaires locales. Ainsi, un mot clé et un qualificatif apparaissant dans la même phrase ont de grandes chances d'être une instance de la paire mot clé/qualificatif correspondante. Au contraire, s'ils apparaissent dans des paragraphes différents, il est peu probable qu'ils soient associés. On considère cependant qu'un qualificatif isolé (c'est à dire, qui n'a pu être associé à aucun mot clé proche) peut être associé à l'un des deux mots clés les plus fréquents du texte, ceux-ci pouvant avoir été sous entendus dans la phrase en question. Les paires ainsi constituées sont validées à l'aide du MeSH et de l'historique des associations rencontrées dans le catalogue CISMeF. Toutes les associations ne sont pas autorisées par le MeSH - Ainsi, le qualificatif *prévention & contrôle* pourra être associé au mot clé *diabète*, mais pas au mot clé *pouce*.

2.2.4 Calcul du score et sélection de l'index

L'indexation finale résulte du classement des mots clés (ou paires) en fonction du score qui leur est attribué. Une fonction de rupture classique permet de fixer une limite adaptative pour déterminer quels sont les éléments retenus. Par la suite, un seuil expérimental (à déterminer) sera utilisé pour départager les mots clés mineurs et majeurs.

Les descripteurs obligatoires (*check tags*) sont systématiquement sélectionnés: leur score devient égal à celui du(des) mot(s) clé(s) le(s) plus fréquent(s) dès qu'ils apparaissent plus d'une fois dans le texte. Pour les autres mots clés (ou paires), le score est calculé à partir du nombre d'occurrences dans la ressource. S'il existe une relation hiérarchique entre plusieurs mots clés (comme sur la figure 4), un partage du score du mot clé père entre les mots clés fils est effectué, afin de satisfaire la règle d'indexation par les mots clés les plus précis.

endocrinopathies
diabète
diabète insulinodépendant
diabète lipoatrophique
diabète gestationnel
diabète non insulinodépendant

Figure 4 : Extrait d'une hiérarchie du MeSH

Ainsi, si on compte 10 occurrences de *diabète*, 15 de *diabète insulinodépendant* et 4 de *diabète non insulinodépendant*, les relations hiérarchiques entre *diabète* et *diabète insulinodépendant* d'une part et entre *diabète* et *diabète non insulinodépendant* d'autre part conduiront aux scores 15+(10/2) = 20 pour *diabète insulinodépendant*, et 4+(10/2) = 9 pour *diabète non insulinodépendant* tandis que *diabète* ne figurera plus dans la liste des mots clés retenus. En accord avec la méthode d'indexation préconisée par Salton (Salton, Mc Gill, 1983) nous calculons à partir du nombre d'occurrences ainsi obtenu le poids tf*idf, en utilisant le catalogue CISMeF comme collection de référence. Le score obtenu est ensuite multiplié par une seconde mesure, qui tient compte de la longueur du document (Robertson, Sparck Jones, 1997). La version finale du système d'indexation automatique devrait également attribuer un poids supplémentaire aux mots clés (ou paires) apparaissant dans les titres des paragraphes, ou dans le résumé. En effet, cette méthode, déjà utilisée par NOMINDEX (Pouliquen 2002) trouve sa justification dans la manière de procéder des documentalistes qui consultent principalement les titres de paragraphes pour vérifier l'indexation d'une ressource.

En résumé, on peut dire que le calcul des scores attribués à chaque mot clé (ou paire) est une fonction du nombre d'occurrences de chaque mot clé, de la longueur de la ressource, de l'historique des notices CISMeF, de la liste des descripteurs obligatoires, et de la structure hiérarchique du MeSH.

3 Evaluation du système

Une évaluation préliminaire du système d'indexation automatique présenté ci-dessus a été réalisée à l'aide d'un dictionnaire MeSH contenant approximativement 500 entrées, relatives aux descripteurs obligatoires, qualificatifs, et mots clés liés au diabète. L'objectif de cette évaluation est de valider globalement la procédure d'indexation, et en particulier la méthode utilisée pour la construction des dictionnaires. Les tests ont donc porté sur un échantillon de 10 ressources sur le thème du diabète extraites du catalogue CISMeF. La taille du corpus ainsi constitué était d'environs 300 Ko, soit 45 241 mots. L'indexation automatique établie par le système a été comparée à l'indexation manuelle figurant dans les notices du catalogue. Les performances du système ont été évaluées par la précision et le rappel, qui sont les mesures usuelles en traitement du langage naturel. Précisons que certaines des ressources utilisées dans

cette évaluation contenaient des termes qui ne se rapportaient pas au diabète (par exemple, *suivi soins patient*), et donc ne figuraient pas dans le dictionnaire, ce qui peut avoir diminué les performances.

Pour chaque ressource, nous avons procédé à une indexation "plein texte" et à une indexation basée sur le résumé ou les titres des paragraphes –notée TOC par la suite. Une analyse comparative des deux types d'indexation devrait mettre en évidence l'importance à accorder aux mots du titre ou du résumé. Nous prévoyons d'intégrer le résultat de ces observations dans la version finale du système. Afin d'illustrer les résultats obtenus, la figure 5 permet de comparer l'indexation automatique plein texte (colonne 1) et l'indexation automatique TOC (colonne 2) avec l'indexation manuelle CISMeF (colonne 3) pour une même ressource.

Indexation Plein Texte	Indexation TOC	Indexation Manuelle CISMeF
grossesse diabète non insulinodépendant /thérapeutique **sujet âgé** **hypoglycémiants /administration & posologie** diabète non insulinodépendant /génetique diabète non insulinodépendant /complication **hypoglycémiants/effets indésirables** **diabète non insulinodépendant** **hypoglycémiants**	**diabète non insulinodépendant** **hypoglycémiants**	* diabète non insulinodépendant /chimiothérapie * diabète non insulinodépendant grossesse hypoglycémiants / effets indésirables hypoglycémiants /administration & posologie hypoglycémiants /classification * hypoglycémiants interactions médicamenteuses (suivi soin patient)[3] sujet âgé

Figure 5 : Indexation de la ressource "diabète de type 2" disponible à l'URL http://agmed.sante.gouv.fr/htm/5/5106c.htm (le 10/01/04)

Les mots clés (ou paires) correctement extraits par le système automatique figurent en gras. Les mots clés (ou paires) considérés comme majeurs par les documentalistes sont signalés dans la colonne 3 par une étoile. La figure 6 indique les mesures de précision et de rappel obtenues sur le corpus pour l'indexation automatique plein texte et TOC.

[3] Ce mot clé MeSH ne figure pas dans le dictionnaire utilisé.

	Précision au rang 2	Précision au rang 4	Rang de la rupture	Précision à la rupture	Rappel à la rupture
Indexation Plein Texte	55%	50% / 43%[4]	4,7 / 4,3	58% / 53%	45% / 33%
Indexation TOC	67%	58% / 47%	4 / 2,4	65% / 68%	35% / 25%

Figure 6 : Précision et rappel du système d'indexation automatique.

4 Discussion

Les résultats montrent que l'indexation TOC permet une indexation plus précise que l'indexation plein texte, alors que le rappel est plus élevé pour l'indexation plein texte que pour l'indexation TOC. Cette observation met en évidence la différence entre les deux types de documents : les titres de paragraphes donnent une bonne vue d'ensemble du contenu, mais pour être exhaustif, il est nécessaire d'utiliser le texte entier. Il faut cependant remarquer que la forme standard de certains titres de paragraphes ne reflète que très vaguement le contenu du texte. Par exemple, les paragraphes intitulés « Introduction » ou « Matériaux et Méthodes » sont muets pour l'indexation. De plus, certaines ressources sont très peu structurées, et ne comportent pas de titres de paragraphes. Ce n'était le cas d'aucune des ressources utilisées pour notre évaluation, mais cet élément est à prendre en compte dans l'utilisation de l'information apportées par les titres.

La précision obtenue sur les 4 premiers termes est de 50% pour l'indexation plein texte et de 58% pour l'indexation TOC, ce qui de l'ordre des 51% réalisés par le système HONselect. La précision sur les 2 premiers termes est de 55% pour l'indexation plein texte et de 67% pour l'indexation TOC, ce qui est dans les deux cas inférieur aux 88% obtenus par le système hybride décrit dans (Ruch et al. 2003). Cette moins bonne performance de notre système semble dûe pour une bonne part à une sélection trop systématique des descripteurs obligatoires. Il faudrait donc étudier les indexations proposées (automatique et manuelle) afin de déterminer si ce phénomène est du à un sur-apprentissage du système ou à un silence de l'indexation manuelle. Plus généralement, une étude menée par la NLM[5] sur la variabilité de l'indexation entre différents indexeurs humains montre que, selon les catégories de descripteurs (*check tags*, associations mot clé/qualificatif *etc.*) l'homogénéité peut varier de 76% à 33% seulement, c'est à dire que, au pire, les indexeurs ne proposent les mêmes descripteurs que dans un cas sur trois.

Contrairement aux autres systèmes, notre approche est fondée sur l'extraction de paires mot clé/qualificatif, ce qui signifie que l'extraction d'un terme isolé au lieu d'une paire est considéré comme erronée. Ainsi, si le système propose le mot clé *diabète* au lieu de la paire

---

[4] Si différent, le chiffre obtenu sans les associations mot clé/qualificatif est indiqué après le signe « / »

[5] cf. Indexing Consistency in Medline , ME. Funk, CA. Reid, LS. Mc Googan à l'URL : http://www.pubmedcentral.gov/picrender.fcgi?tool=pmcentrez&action=stream&blobtype=pdf&artid=227138 (le 10/01/04)

diabète/chimiothérapie, nous considérons que le descripteur correct n'a pas été extrait. De même, si le système propose la paire *diabète/thérapeutique* au lieu de *diabète/chimiothérapie*, l'indexation est incorrecte, bien que ces deux descripteurs renvoient à des notions très proches – la *chimiothérapie* indique spécifiquement un traitement médicamenteux, alors que *thérapeutique* peut évoquer tous types de traitement. Malgré cela, sur l'exemple présenté dans la figure 5, nous constatons que deux des quatre paires mot clé/qualificatif attendues sont extraites, et on observe également que deux des trois mots clés (ou paires) majeurs attendus sont proposés par le système d'indexation automatique. Nous observons également pour l'indexation TOC que les mots clés extraits sont des mots clés majeurs dans l'indexation manuelle.

Les taux de précision et de rappel donnés sur la figure 6 montrent d'une manière générale la plus value de cette approche, puisque les résultats obtenus en recherchant les associations mot clé/qualificatif sont meilleurs que les résultats de l'indexation par terme isolé. Cela est en partie du à l'application de la norme tf/idf aux descripteurs, car elle est dépendante de l'indexation des documents contenus dans la collection. En effet, les paires ont en général une fréquence moins élevée que les mots clés isolés, ce qui fait qu'elles sont plus promues dans le classement final. Cependant, cette évaluation a été réalisée sur un nombre réduit de ressources, et ne peut être considérée comme représentative des performances du système. Par ailleurs, l'utilisation de dictionnaires de plus large couverture pourrait certainement amplifier le bruit, ou au contraire, augmenter la précision en permettant de repérer un plus grand nombre de mots clés pertinents.

Les travaux en cours visent à étendre la couverture du dictionnaire à tout le MeSH, en utilisant diverses méthodes de traitement du langage naturel. Certaines entrées peuvent être générées à l'aide des dictionnaires DELA du français réalisés au LADL (Courtois, Silberztein, 1990). Le lexique médical français développé dans le cadre du projet UMLF (Zweigenbaum et al. 2003) permet également d'enrichir notre base. Les améliorations futures doivent porter sur la distinction des mots clés mineur et majeur, grâce aux informations sur leur nombre d'occurrence et leurs positions dans le texte. Une interface conviviale devra également être développée, afin de faciliter le paramétrage du système par les documentalistes. Dans les semaines à venir, on effectuera en parallèle l'indexation de 60 ressources à l'aide du système d'indexation CISMeF et du système hybride décrit dans (Ruch et al. 2003).

5 Conclusion

Nous avons présenté l'architecture d'un système d'indexation MeSH utilisant la plateforme INTEX ainsi que des bases de connaissances MeSH et CISMeF. Conçu pour alléger la tache des documentalistes de l'équipe CISMeF, ce système peut extraire des paires mot clé/qualificatif plutôt que des termes isolés, et cherche à maximiser la précision, de préférence au rappel. Les premiers résultats obtenus montrent que la stratégie générale d'indexation est pertinente, et que la couverture du dictionnaire employé sur les termes relatifs au diabète est bonne. La précision atteinte est comparable à celle des systèmes existant. Cependant, des améliorations, notamment en ce qui concerne la couverture des dictionnaires, doivent encore être apportées avant que le système puisse être effectivement opérationnel sur toutes les ressources de santé.

Références

ARONSON AR., BODENREIDER O., CHANG F., HUMPHREY SM., MORK JG., NELSON SJ., RINDFLESCH TC., WILBUR WJ. (2000) The NLM Indexing Initiative. Actes de *AMIA Symposium*, (20 Suppl): 17-21.

COURTOIS B., SILBERZTEIN M. (1990) *Dictionnaires électroniques du français*, Paris, éditions Larousse.

CROCHEMORE M., HANCART C., LECROQ T. (2001) *Algorithmique du texte*, Paris, éditions Vuibert.

DARMONI SJ., LEROY JP., THIRION B., BAUDIC F., DOUYÈRE M., PIOT J. (2000) CISMeF: a structured Health resource guide. *Meth Inf Med*, 39(1): 30-5.

GAUDINAT A., BOYER C. (2002) Automatic Extraction of MeSH terms from Medline Abstracts. Atelier de *NLPBA2002* sur le *Traitement du Langage Naturel dans les applications Biomédicales*.

HUMPHREY SM., MILLER NE. (1987) Knowledge-based indexing of the medical literature: The Indexing Aid Project. *J Am Soc Inf Sci* , May 38(3): 184-96.

JACQUEMIN C., ROYAUTÉ J. (1994) Retrieving Terms and their Variants in a Lexicalised Unification-Based Framework , Actes de *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 132-141.

POULIQUEN B. (2002) Indexation de textes médicaux par indexation de concepts, et ses utilisations. *Thèse de Doctorat*, Université Rennes 1, 2002.

ROBERTSON S.E., SPARCK JONES K. (1997) Simple, proven approaches to text-retrieval. Rapport technique 356, *Laboratoire d'informatique de l'Université de Cambridge*.

RUCH P., BAUD R., Geissbühler A. (2003) Learning-Free Text Categorization. Actes de *Artificial Intelligence in Medicine in Europe*, 199-204.

SALTON G., MC GILL MJ. (1983) *Introduction to Modern Information Retrieval*, New York, McGraw-Hill.

SILBERZTEIN M. (1993) *Dictionnaires électroniques et analyse automatique de textes : le système INTEX*, Paris, Masson.

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

ZWEIGENBAUM P., BAUD R., BURGUN A., NAMER F., JAROUSSE E., GRABAR N., RUCH P., LE DUFF F., THIRION B., DARMONI SJ. (2003) UMLF : construction d'un lexique médical francophone unifié. Actes de *Journées Francophones d'Informatique Médicale*, sous presse.