Université de Rouen U.F.R des sciences et techniques

Ecole doctorale Sciences Physiques, Mathématiques et de l'information pour l'Ingénieur

Recherche d'information et indexation automatique des médicaments à l'aide de plusieurs terminologies de santé

THÈSE

présentée et soutenue publiquement le 23 Novembre 2010 par

Saoussen SAKJI

Pour obtenir le grade de

Docteur de l'université de Rouen

Discipline : Informatique

Composition du Jury :

Stéfan DARMONI	Directeur de thèse
Michel JOUBERT	Co-encadrant
Jean-François GEHANNO	Co-encadrant
Alain VENOT	Rapporteur
Lynda TAMINE-LECHANI	Rapporteuse
Thierry PAQUET	Examinateur

Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes (LITIS EA4108)

RESUME L'utilisation des données médicales et l'accès à une information concise sont devenus des enjeux majeurs, non seulement, pour les professionnels de santé mais aussi pour le grand public. Pour faciliter à atteindre cet objectif, plusieurs terminologies médicales ont été développées. Ces dernières sont employées, généralement, pour des finalités différentes. Par exemple, la classification ATC est utilisée pour classer les médicaments, la nomenclature SNOMED pour le codage clinique, les classifications CIM-10 et CCAM pour le codage épidémiologique puis médico-économique, le thésaurus MeSH pour la bibliographie...

Ainsi, dans un contexte appliqué au domaine de la recherche d'information médicale, les objectifs de cette thèse ont été la création d'un modèle de recherche utilisant plusieurs terminologies médicales, dans un premier temps. Cet univers multi-terminologique permet d'améliorer la qualité de l'information restituée selon les propres connaissances des utilisateurs. Ensuite, nous avons été amenés à concevoir une approche d'indexation automatique, par la classification ATC, pour les ressources du Portail d'Information sur les Médicaments (PIM), conçu dans le cadre du projet européen PSIP. Cette indexation a pour but d'améliorer l'indexation des médicaments afin de fournir à l'utilisateur une information plus fine et détaillée. Enfin, nous avons modifié notre algorithme de recherche afin de l'adapter à notre nouvelle structure multi-terminologique.

MOTS-CLEFS : indexation et rédaction du résumé comme sujet ; médicaments ; recherche et stockage d'information ; terminologie ; traitement langage naturel ; vocabulaire contrôlé.

ABSTRACT The use of medical data and the access to concise information has become of major importance, not only, for health professionals but also for the general public. To facilitate this goal, several medical terminologies have been developed. The latter are employed, generally, for different purposes. For example, the ATC classification is used to classify drugs, SNOMED nomenclature for clinical coding, ICD-10 and CCAM classifications for epidemiologic coding then medico-economic, the MeSH thesaurus for the bibliography etc.

Thus, in the context of medical information retrieval, the objectives of this thesis were the creation of a research model using several medical terminologies, as a first stage. This multi-terminological universe allows to improve the quality of the retrieved information according to users' own knowledge. Then, we developed an automatic indexing approach, by ATC classification, for the resources of the Drug Information Portal (DIP), designed within the framework of PSIP European project. The purpose of this study is to improve the indexing of drugs in order to provide to the user more accurate and detailed information. Lastly, we modified our algorithm of research in order to adapt it to our new multi-terminological structure.

KEY WORDS: Abstracting and indexing as topic; Drugs; Information Storage and Retrieval; Terminology; Natural language processing; Vocabulary, controlled.

REMERCIEMENTS

Je tiens, tout d'abord, à remercier le professeur Stéfan Darmoni pour m'avoir accueillie au sein de son équipe CISMeF et pour avoir dirigé ma thèse. Son énergie et son dynamisme ont développé en moi le sens du travail approfondi et de la recherche. Je remercie, également, mes encadrants Michel Joubert qui m'a fait profiter de sa compétence ainsi que, le docteur Jean-François Gehanno pour son aide.

Je remercie le professeur Alain Venot et madame Lynda Tamine-Lechani qui ont accepté d'être mes rapporteurs, ainsi que Thierry Paquet qui a évalué mon travail.

Je tiens à remercier très sincèrement l'ensemble des membres du jury qui me font le grand honneur d'avoir accepté de juger mon travail.

J'adresse un grand merci chaleureux à tous les membres de l'équipe CISMeF pour la bonne ambiance et leur aide ; par ordre alphabétique : Ahmed, Aurélie, Badisse, Benoit, Catherine, Élise, Gaétan, Ivan, Josette, Julien, Lina, Romain, Sandrine, Suzanne, Tayeb et Zied et sans oublier Hocine du LERTIM et Thierry Locroq de l'équipe TIBS.

Je tiens à remercier, tout particulièrement, et à témoigner toute ma reconnaissance aux personnes suivantes :

- Christian Kala-Lobé pour son assistance lors de la mise en place des outils sémantiques d'Oracle ;
- le professeur Peter Elkin et toute son équipe pour leur accueil durant mon séjour aux États-Unis et leur sérieux travail qui m'a permis de passer un stage laborieux et agréable ;
- la pharmacienne-documentaliste Catherine Letord, le docteur Laetitia Rollin et le docteur Philippe Massari pour leur aide à l'évaluation des études réalisées ;
- Richard Medeiros pour ses conseils linguistiques.

Mes remerciements à Josette, Benoit, Catherine, Ivan et le docteur Massari pour la lecture de mon manuscrit et leurs remarques qui m'ont permis d'améliorer la qualité de mon rapport.

Par ailleurs, un merci chaleureux à tous mes proches qui ont cru en moi et n'ont cessé de m'encourager : la confiance de mes parents, la bienveillance et le support de ma sœur qui m'a poussé vers l'avant, mon Doudou qui m'a fait oublié les coups durs, l'intérêt de mon frère ainsi que tous les membres de ma chère famille.

Je remercie, également, mes amis qui m'ont accompagnée et soutenue durant cette thèse, et tout particulièrement Nadine, Rania et Hany.

Finalement, je tiens à remercier toutes les personnes qui ont rendu possible la réalisation de cette thèse et m'ont encouragé à la finaliser.

Je dédie cette thèse à toute ma famille.

Table de matières

TABLE DES FIGURES

LISTE DES TABLEAUX

INTRODUCTION GENERALE

Il est de la responsabilité de tous de veiller à ce que les nouveaux moyens de diffusion de l'information se traduisent par un enrichissement, et non un appauvrissement du patrimoine culturel mondial.

Pierre Joliot

La recherche d'information est aujourd'hui une activité d'autant plus importante qu'elle s'inscrit dans un contexte dans lequel les technologies de l'information et de la communication (TIC) évoluent rapidement. Pour cela, il faut pouvoir, parmi l'abondance de documents disponibles, trouver l'information correspondant à nos besoins en un minimum de temps. En effet, sur la Toile, le meilleur cohabite souvent avec le pire, ce qui nous incite à développer des stratégies de recherche de plus en plus complexes et simplifiées en même temps, afin de trouver l'information souhaitée.

Des logiciels de traitement de l'information permettent de retrouver des informations dans des corpus riches en documents. La question qui se pose au sujet de ces systèmes de recherche d'informations se rapporte essentiellement à leur efficacité : pertinence, exhaustivité, ergonomie…

Un système de recherche d'information possède trois fonctions principales fondamentales : représenter le contenu des documents d'un corpus donné, représenter le besoin de l'utilisateur exprimé sous la forme d'une requête et comparer ces deux représentations pour en extraire le meilleur. La représentation des documents et de la requête se fait à l'issue de la phase d'indexation qui consiste à choisir les termes les plus représentatifs des documents dans un espace de représentation. Le résultat de la recherche d'information devrait être aussi pertinent que possible afin de satisfaire l'utilisateur. La satisfaction des utilisateurs peut influencer la grille d'évaluation des systèmes de recherche d'information.

Si, en effet, Internet a beaucoup à offrir à qui sait ce qu'il cherche, le même Internet est tout aussi capable de compléter l'abrutissement de ceux et celles qui y naviguent sans boussole.

Laurent Laplante

S'intéressant au domaine médical, et avec le développement du Web et la croissance du volume des données diffusées sur Internet, la recherche d'information médicale devient de plus en plus difficile en termes de qualité et requiert davantage de techniques et connaissances pour avoir une information fiable qui répond au mieux aux besoins des utilisateurs.

Quelles que soient leurs expériences du Web et leurs compétences en recherche d'information, les utilisateurs rencontrent des difficultés à rechercher une information de santé sur l'Internet (Keselman et al., 2008). La plupart de ces derniers entament leurs

recherches via les moteurs de recherche généralistes (tel que Google), plutôt que les bases de données médicales spécialisées (Jansen et al., 2006). Cependant, avoir recours à un moteur de recherche spécialisé peut, dans la plupart des cas, donner de meilleurs résultats.

Au cours de son processus de recherche, la difficulté majeure rencontrée par l'utilisateur est de bien exprimer ses besoins informationnels et de trouver les termes adéquats correspondant à l'objet de sa recherche. Les principaux répertoires et sites médicaux de la Toile indexent les documents à l'aide du thésaurus MeSH développé par la National Library of Medicine (NLM) des Etats-Unis. Ceci peut obliger, en quelque sorte, tous les utilisateurs, quels que soient leurs niveaux en médecine, d'utiliser les termes du MeSH pour transcrire leurs requêtes dans le but d'avoir un résultat aussi performant que possible. À ce niveau, la question qui se pose « Sommes-nous (professionnels de santé, étudiants et patients) contraints de connaître le thésaurus MeSH pour avoir une réponse pertinente à notre quête d'information médicale ?».

Motivation et objectifs

Pour répondre à la question précédente, notre objectif durant cette thèse a été de fournir un univers multi-terminologique (fondé sur plusieurs terminologies médicales, outre le thésaurus MeSH) à l'utilisateur du catalogue CISMeF (Catalogue et Index des Sites Médicaux de langue Française) afin de satisfaire au mieux son besoin informationnel selon ses propres connaissances terminologiques. CISMeF est un site Web relatif au domaine de santé permettant de fournir aux utilisateurs les ressources disponibles en français répondant à leurs requêtes. Jusqu'en 2005, les ressources du catalogue ont été indexées exclusivement à l'aide du thésaurus MeSH permettant ainsi une recherche mono terminologique.

La problématique cruciale qui se posait, au fur et à mesure, était de permettre un accès « intelligent » à l'information médicale. De ce point de vue, les terminologies médicales prennent de plus en plus d'importance. En effet, elles fournissent un vocabulaire commun et une description de la signification des termes d'un domaine ainsi que les relations qui les relient. Elles sont, non seulement exploitables de manière informatique, mais aussi elles jouent un rôle important pour la nouvelle génération du Web sémantique car elles sont indispensables pour décrire le contenu des ressources du Web et faciliter ainsi leurs exploitation.

Dans ce travail, nous cherchons principalement à implémenter une structure multi-terminologique (fondée sur plusieurs terminologies médicales) au sein de CISMeF. Les terminologies à intégrer dans la nouvelle base de données peuvent être employées selon plusieurs contextes d'utilisateurs. Par exemple, un pharmacien pourrait accéder au catalogue en utilisant la classification Anatomique Thérapeutique et Chimique (ATC) vu qu'il aurait plus de connaissances de cette terminologie. D'autre côté, un médecin de réadaptation souhaiterait les références des ressources relatives aux diagnostics, au fonctionnement et handicap et indexées par ailleurs par la terminologie spécifique à savoir la CIF (Classification Internationale du Fonctionnement, du Handicap et de la Santé)…

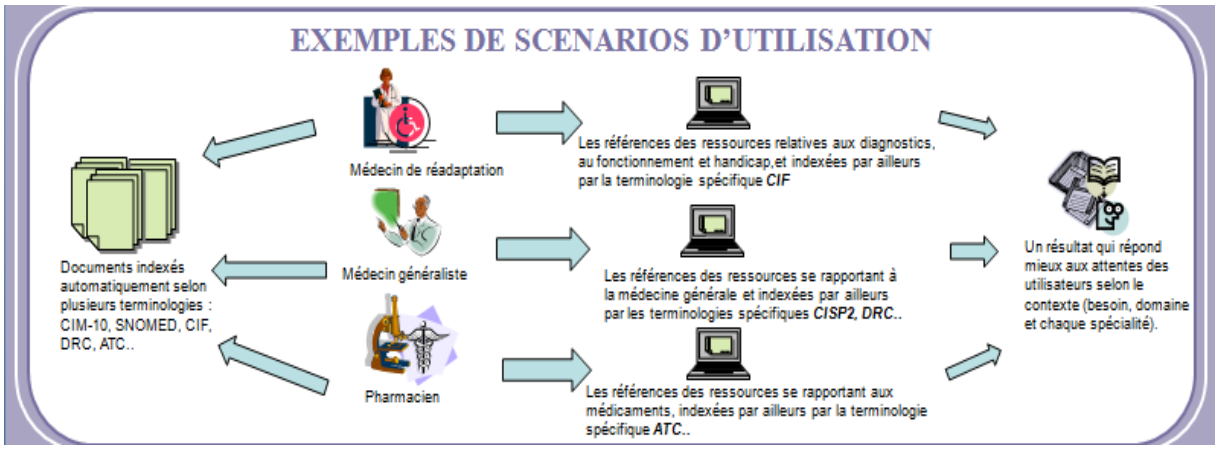


Figure IG. Les différents contextes d'utilisation de plusieurs terminologies médicales

Pour ce faire, nous devons se procurer des terminologies médicales disponibles en français et qui correspondent aux connaissances des utilisateurs de CISMeF. Ensuite, il faut les étudier afin de comprendre leurs structures et leurs spécificités et, les modéliser pour pouvoir les intégrer dans une même structure homogène.

Ce manuscrit est organisé en sept chapitres. Nous exposons en premier lieu le contexte du travail dans lequel s'est déroulée la thèse. Nous commençons par une brève présentation du LERTIM et de l'équipe CISMeF. Nous décrivons par la suite le catalogue CISMeF autour duquel se déroulent nos travaux de recherche. Nous décrivons, par la suite, notre participation au projet PSIP qui finance cette thèse.

Le deuxième chapitre a pour objectif de présenter quelques concepts de base utiles pour la compréhension du domaine de la recherche d'information. Après une brève présentation de la recherche documentaire, nous définissons quelques systèmes de recherche d'information (SRI), leurs particularités et leurs fonctionnements. Nous présentons par la suite les notions de l'indexation, puis nous passons en revue les modèles piliers de la RI et les critères et mesures d'évaluation des SRI.

À travers le troisième chapitre, nous définissons le vocabulaire utilisé en tant que terminologies médicales en se focalisant sur celles qui ont été les plus impliquées dans notre travail. Dans la deuxième partie de ce chapitre, nous mettons en relief le passage vers une structure multi-terminologique fondée sur plusieurs terminologies médicales en mettant en avant le processus d'intégration de toutes ces terminologies selon un modèle générique.

Nous présentons, dans le quatrième chapitre, la deuxième réalisation faite autour de l'univers multi-terminologique ; à savoir la création d'un Portail d'Information bilingue sur les Médicaments (PIM). Cette réalisation nous a permis, par la suite, une exploitation plus analytique des informations concernant les médicaments en mettant en place une approche d'indexation automatique par la classification ATC. Enfin, nous concluons ce chapitre par une description de l'étude mettant en relief l'amélioration de la recherche d'information grâce à la correspondance entre le thésaurus MeSH et la classification ATC.

Le cinquième chapitre décrit notre algorithme de recherche d'information multi-terminologique. Nous présentons, tout d'abord, une panoplie de travaux et de systèmes de

recherche d'information fondés sur l'expansion de requêtes et de la sémantique. Nous détaillons par la suite l'algorithme de recherche d'information multi-terminologique au sein du catalogue CISMeF ainsi que l'évaluation qui a été faite, afin de mettre en relief la valeur ajoutée de notre approche.

Dans l'avant dernier chapitre, nous décrivons les travaux connexes aux principaux thèmes de la thèse, notamment le passage du monde mono-terminologique vers l'univers multi-terminologique, la recherche d'information multi-terminologique et l'indexation automatique bi-terminologique des médicaments. Toutefois, ils restent au centre du domaine de la recherche d'information multi-terminologique. Notre participation à ces travaux a donné suite à d'autres perspectives prometteuses pour améliorer l'indexation et la recherche d'information médicale.

Enfin, à travers le dernier chapitre, nous mettons en relief nos perspectives et nos projets de recherche en continuation avec les travaux de la thèse.

Nous concluons ce manuscrit par une conclusion générale récapitulative des différentes réalisations de notre travail.

.

CHAPITRE 1

CONTEXTE DU TRAVAIL

INTRODUCTION

Dans ce chapitre, nous exposons le contexte du travail dans lequel s'est déroulée la thèse. Nous commençons par une brève présentation du LERTIM et de l'équipe CISMeF. Nous décrivons par la suite le catalogue CISMeF autour duquel se déroulent nos travaux de recherche. Finalement, nous décrivons notre participation au projet PSIP qui finance cette thèse.

1.1 CONTEXTE DU TRAVAIL

1.1.1 LE LERTIM

La thèse est co-encadrée par Michel Joubert membre du Laboratoire d'Enseignement et de Recherche sur le Traitement de l'Information Médicale (LERTIM) de Marseille. Les principaux thèmes de recherche du LERTIM sont l'élaboration des systèmes d'informations hospitaliers tels que les systèmes d'informations médicaux et de santé ou encore les systèmes d'information pour la formation à distance…Par ailleurs, les activités de recherche s'appliquent à la bio statistique, la représentation des connaissances, l'aide à la décision et le soutien méthodologique en recherche clinique.

L'objectif de l'activité de recherche du LERTIM est de comprendre, représenter et utiliser la connaissance pour faciliter et/ou permettre l'accès aux connaissances et leur acquisition. Cette recherche vise à élaborer des méthodes et développer des outils permettant un couplage entre connaissance médicale et information sur le patient, afin d'améliorer la décision médicale et la prise en charge du patient.

Les projets de l'équipe se situent dans les champs de recherche concernant les outils d'interopérabilité, d'aide à la décision, des références médicales… et exigent une approche intégratrice :

- de différents domaines de recherche classiques (description des concepts médicaux, ontologies, référentiels sémantiques, méthodes d'intelligence artificielle et psychologie cognitive, élaboration de modèles de raisonnement, modèles cognitifs d'interaction homme-machine) ;
- de développement de composants logiciels de présentation, de traitement et de communication des informations et des connaissances ainsi que des technologies du multimédia ;
- des technologies du génie logiciel offertes par le marché pour réaliser des outils de couplage interopérables en pratique (technologies de l'Internet, approche composant, architectures de systèmes d'information,…).

Outre une activité de soutien à la recherche clinique, l'équipe développe une activité de recherche propre portant sur la biostatistique.

Ces travaux de recherche clinique concernent le plus souvent la recherche de facteurs pronostiques notamment en cancérologie. Un autre champ de recherche est consacré au paludisme, en collaboration avec d'autres centres de recherche et hôpitaux.

Par ailleurs, au sein du LERTIM, plusieurs travaux ont vu le jour dans un but de faciliter l'accès à des bases d'information du domaine médical. Parmi ceux-ci, nous pouvons citer les projets[1] WARPIN (Joubert et al., 2007) dédiés principalement aux citoyens et, ARIANE (Joubert et al., 2002) et CoMeDIAS (Joubert et al., 2003) qui ont été conçus afin de permettre aux professionnels de santé d'accéder plus facilement à des bases de données patients, à des banques de données sur les médicaments, à des guides de bonnes pratiques ou encore à des bibliographies.

1.1.2 L'EQUIPE CISMEF

Sous la codirection du responsable des technologies de l'information et de la communication (Stéfan Darmoni) et du conservateur de la bibliothèque médicale (Benoît Thirion) du Centre Hospitalier Universitaire de Rouen, l'équipe CISMeF est composée de trois documentalistes experts du domaine médical, de trois ingénieurs de recherche, d'un post-doc et de trois doctorants (cf. figure 1.1.2).

[1] Projets de recherche du LERTIM. URL : http://cybertim.timone.univ-mrs.fr/recherche/projets-recherche

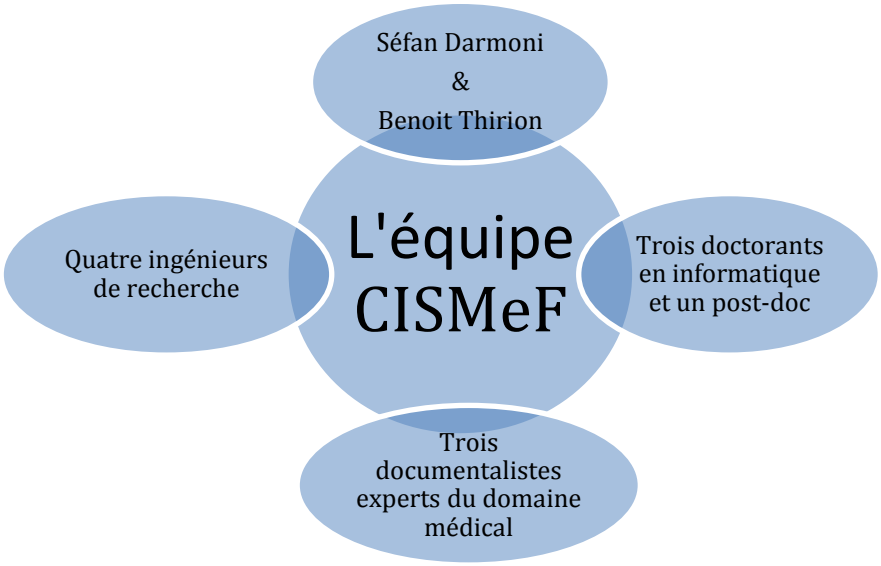


Figure 1.1.2. L'organisation de l'équipe CISMeF

De nombreux travaux ont été entrepris par l'équipe CISMeF dans le domaine de l'indexation et la recherche d'information en santé. La principale application réalisée par l'équipe est la création du catalogue CISMeF (cf. section 1.1.2.1) qui permet de faciliter la recherche documentaire et l'accès aux ressources de santé sur l'Internet.

1.1.2.1 LE CATALOGUE ET INDEX DES SITES MEDICAUX DE LANGUE FRANÇAISE : CISMEF

Depuis sa création, en 1995, CISMeF (acronyme de Catalogue et Index des Sites Médicaux de langue Française) est hébergé par le Centre Hospitalier Universitaire (CHU) de Rouen. Ce site s'adressait avant tout aux professionnels de santé et aux étudiants en médecine. Au fil du temps, une partie destinée aux patients et leur famille a été mise en place afin de leurs apporter des informations de qualité, d'ordre documentaire, produites principalement par des institutions comme le ministère de la santé et les différentes agences de santé, par des sociétés savantes ou encore par des professionnels de santé.

CISMeF respecte un grand nombre de critères de qualité de l'information de santé sur l'Internet, en particulier, ceux émis par le Net Scoring [2] et MedCIRCLE[3].

Labellisé HONcode[4] depuis dix ans, CISMeF recense en priorité les principaux sites et documents francophones tels que les sites institutionnels, les sites non commerciaux en libre accès... Tout site ou document à caractère commercial (site de société pharmaceutique ou autre organisme de vente, site contenant de la publicité. . .) est rejeté, quel que soit son contenu. Les documents retenus sont ceux qui présentent, de préférence, un caractère

[2] NetScoring : Critères de qualité de l'information de santé sur l'Internet. URL : http://www.chu-rouen.fr/netscoring
[3] MedCIRCLE: The Collaboration for Internet Rating, Certification, Labeling and Evaluation of Health Information. URL : http://www.medcircle.org
[4] Health On the Net Foundation. URL : https://www.hon.ch/HONcode

institutionnel et d'une manière générale les documents émanant des sites gouvernementaux (Ministère de la Santé, Ministère de la Justice, Sénat etc.), des facultés de médecine, des hôpitaux, des agences nationales reconnues dans le domaine médical (HAS, etc.) et des sociétés savantes en médecine. Par ailleurs, des sites d'associations et quelques sites personnels, ne comportant pas de publicité et non affiliés à des organismes commerciaux, peuvent être retenus. Il s'agit souvent de sites mis en ligne par des patients qui peuvent faire bénéficier d'autres patients de leur expérience.

En effet, les critères de sélection des ressources CISMeF, s'appuient principalement sur la source et la qualité de la ressource. Sensible à la qualité des ressources du catalogue CISMeF, l'équipe CISMeF a participé à la mise au point d'une grille d'évaluation fondée sur les critères de qualité du Net Scoring (au total 49 critères) (Darmoni et al., 1999). Avec ces critères, une attention particulière est portée à la mention explicite du nom des éditeurs, des auteurs ainsi que les dates de publication et de mise à jour des ressources. Ces critères concernent essentiellement le contenant plus que le contenu.

Afin de compléter cette évaluation de la qualité des ressources disponibles sur l'Internet, l'équipe CISMeF a retenu un critère majeur dénotant la qualité du contenu. Il s'agit de l'indication du niveau de preuve[5] selon la définition de la FNCLCC (Fédération Nationale des Centres de Lutte Contre le Cancer) (Darmoni et al., 2003).

Des efforts considérables sont mis en œuvre permettant une sélection des ressources qui respectent ces critères de qualité et une indexation fine de ces ressources avec des métadonnées standardisées (Thirion et al., 2004).

Chaque ressource du catalogue est décrite et indexée par son contenant en utilisant plusieurs ensembles de métadonnées et par son contenu en utilisant les terminologies médicales, notamment la terminologie CISMeF (cf. section 3.1.2.4). Les métadonnées se réfèrent aux informations descriptives des ressources Web et ont pour finalité de faciliter et d'améliorer la recherche d'information. Dans CISMeF, les métadonnées sont essentiellement celles du Dublin Core (Thirion et al, 2004).

Les ressources incluses dans CISMeF sont décrites par 11 champs (auteur ou créateur, date de publication, description, format, identifiant, langue, éditeur, type de ressource, droit, sujet et mots-clés et titre) parmi 15 éléments de la version 1.1 du DCMES. CISMeF n'emploie pas les 4 autres éléments de DCMES (contribuant, assurance, relation, source) parce qu'ils n'étaient pas nécessaires pour décrire des ressources de santé à inclure dans CISMeF (Dekkers et al., 2003). En plus, onze éléments de la catégorie « Education » d'IEEE 1484 LOM (Learning Object Metadata)[6], sont utilisés pour représenter les ressources pédagogiques. Les métadonnées *indication du niveau de preuve* et *méthode utilisée pour calculer le niveau de preuve* ont été créées pour les ressources destinées aux professionnels de santé (Darmoni et al., 2003). Les métadonnées HIDDEL (Eysenbach et al., 2001) ont été introduites dans le

[5]Médecine fondée sur la preuve. URL : http://www.chu-rouen.fr/ssf/profes/evidencebasedmedicine.html

[6] IEEE 1484 Learning Objects Metadata (IEEE LOM).
URL :http://projects.ischool.washington.edu/sasulton/IEEE1484.html

cadre du projet européen MedCircle (Mayer et al., 2003) afin d'évaluer la qualité de l'information de santé. Par ailleurs, des métadonnées spécifiques à l'équipe CISMeF ont été ajoutées pour décrire la qualité ou la localisation de la ressource telles que *institution*, *ville*, *province*, *pays*, *type d'accès*, *partenariat*, *coût* et *public ciblé*. Certains de ces champs (par exemple : coût) sont également présents dans LOM (Bourda et al., 1999).

❖ Le degré d'importance des ressources

Selon le degré d'importance des ressources collectées dans le catalogue CISMeF, trois niveaux d'indexation sont appliqués : le premier niveau (N1) pour une indexation manuelle, le deuxième niveau (N2) pour une indexation supervisée et le troisième niveau (N3) pour une indexation automatique. L'indexation manuelle concerne les ressources jugées importantes et prioritaires telles que : les recommandations nationales, les lectures critiques d'articles, les sites institutionnels (ministériels ou gouvernementaux) et les sites d'associations patients. L'indexation supervisée concerne les ressources qui sont moins importantes que celles du premier niveau, cependant jugées assez importantes pour qu'elles ne soient pas indexées qu'automatiquement. Ainsi, l'indexation supervisée est, d'abord, automatique, puis revue manuellement dans un second temps par les indexeurs de l'équipe CISMeF. On retrouve les rapports techniques, les études d'évaluation, les cours de campus numériques, les articles de périodiques concernant les formations continues médicales et les ressources sur l'information sur les médicaments. Quant à l'indexation automatique, elle concerne les ressources qui ont une importance mineure telles que : les rapports sur la politique de santé et de santé publique, les cours ne venant pas de campus numériques, quelques ressources sur l'information sur les médicaments.

À ce jour[7], le catalogue CISMeF recense 38.712 ressources indexées manuellement, 9.659 ressources supervisées et 24.982 ressources indexées automatiquement.

Ainsi, CISMeF est un catalogue décrivant et indexant les principales sources d'information institutionnelles de santé françaises (N ≈ 73.353)[8], ayant également un système de recherche d'information médicale (Doc'CISMeF). Ce dernier a été longtemps fondé exclusivement sur un monde mono-terminologique, reposant exclusivement sur le thésaurus MeSH (cf. section 4.1.2.3).

CISMeF propose un accès aux ressources de santé du catalogue selon trois contextes utilisateur et selon cinq modes de recherche différents (cf. Figure 1.1.2.1.1).

[7] Statistiques datant du 6 Juillet 2010
[8] À la date du 06/07/2010

Figure 1.1.2.1.1. Page d'accueil du catalogue CISMeF

En effet, CISMeF offre un accès contextuel pour les professionnels de santé à travers la rubrique « Recommandations et consensus », pour les patients via la rubrique « Informations pour les patients » et pour les étudiants en médecine selon la rubrique « Enseignements et formation » en limitant la recherche générale à chacun de ces domaines. Ces trois catégories sont répertoriées selon le type des ressources et la nature d'indexation de ces dernières. Par exemple, dans la catégorie « Recommandations et consensus », nous retrouvons les ressources de types : *conférence de consensus*, *recommandations de bon usage du médicament*…alors que dans la catégorie « Enseignements et formation », nous retrouvons celles qui concernent les *documents pédagogiques*, *les périodiques*…

Concernant les cinq modes d'accès aux ressources, il s'agit d'un accès :

- ✓ par le *moteur de recherche Doc'CISMeF* qui offre trois possibilités de recherche :
  - ➢ la recherche simple s'effectue par un seul mot ou par une expression de mots, en langage naturel ou à l'aide de termes appartenant à la terminologie CISMeF. La recherche peut s'effectuer, aussi, d'une manière booléenne à travers des opérateurs logiques, ce qui requiert une bonne connaissance pour la manipulation des opérateurs booléens (ET, OU, SAUF) et des codes des

champs de recherche. Exemple (asthme.ti) pour chercher une ressource ayant le mot « asthme » dans le titre ;

- la recherche avancée permet d'effectuer des recherches précises, sur tous les champs d'une notice (titre, mots d'indexation de la ressource…), à l'aide ou non des opérateurs booléens (cf. Figure 1.1.2.1.2).

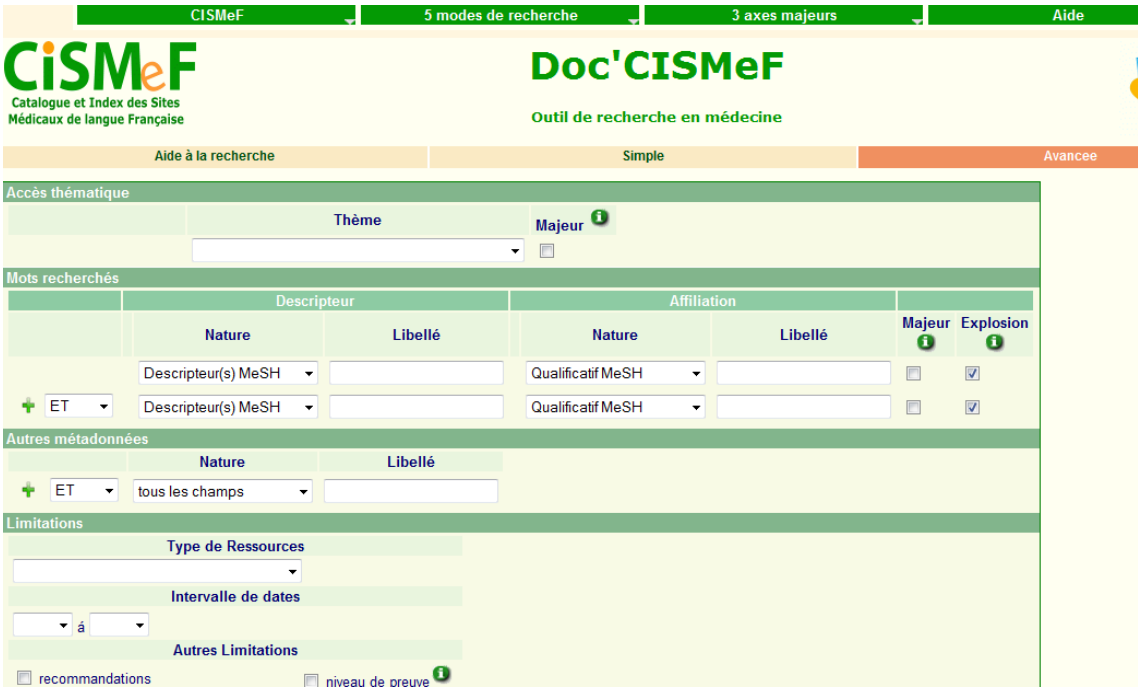


Figure 1.1.2.1.2. Exemple de recherche avancée dans CISMeF

- ✓ par l'*Index alphabétique*. Il s'agit d'un classement alphabétique de la traduction française des termes du thésaurus MeSH ainsi que les qualificatifs et les types de ressources de la terminologie CISMeF. À chaque terme correspond une page présentant le terme en anglais, sa définition, ses synonymes MeSH, l'arborescence du thésaurus MeSH contenant le terme et des requêtes préconstruites[9]. Ces dernières définissent des stratégies de recherche pour améliorer la recherche d'information sur des notions qui n'ont pas d'équivalents dans la terminologie CISMeF. Par exemple, la requête « insulinothérapie » est interprétée comme suit : rechercher les ressources indexées par le descripteur MeSH « *insuline* » et le qualificatif « *usage thérapeutique* ». Ces requêtes donnent accès aux ressources selon un contexte utilisateur : pour les professionnels de santé ou pour les patients ou encore pour les étudiants ;
- ✓ par l'*Index thématique*. Il s'agit d'un classement thématique par spécialité médicale. A chaque spécialité correspond une page définissant le terme en anglais, tous les termes

[9] Se référer au Chapitre3 ; Section 3.1.2.4 pour plus de détails concernant les requêtes préconstruites

CISMeF[10] (descripteur, qualificatif ou type de ressource) qui lui sont liés sémantiquement ainsi que des requêtes préconstruites permettant d'accéder aux ressources relatives soit à la spécialité médicale choisie, soit à l'un des termes qui lui sont sémantiquement liés. Ces liens sémantiques ont été réalisés manuellement par le responsable de la bibliothèque médicale (Benoit Thirion) ;

- ✓ par le *portail terminologique MeSH*. Les requêtes en langage naturel renvoient des informations concernant la définition du terme, ses synonymes MeSH français, ses synonymes MeSH anglais, les qualificatifs associés à ce terme, les types de ressources[11] affiliés au terme, les métatermes[12] auxquels il appartient ainsi qu'aux arborescences du terme. Pour chaque terme, des requêtes préconstruites permettent d'accéder aux ressources correspondantes en français dans le catalogue CISMeF ou en anglais dans la base MEDLINE ;
- ✓ par les *types de ressources*. Ce mode d'accès permet d'avoir des ressources selon le contexte des utilisateurs : les professionnels de santé, les étudiants en médecine ou les patients. A chaque type de ressource correspond une annotation définissant ses synonymes, les métatermes auxquels il appartient ainsi que sa définition complète. Chaque type de ressource est représenté par son équivalent anglais et les types de ressources qui le subsument.

❖ Présentation du résultat de la recherche d'information

Selon ces cinq modes de recherche et ces trois modalités d'accès contextuels, le résultat de la recherche est un ensemble de notices courtes (cf. Figure 1.1.2.1.3), associées aux ressources retournées répondant au besoin informationnel de l'utilisateur, et affichées par ordre chronologique et par degré d'importance (les ressources du N1, puis celles du N2 et enfin celles du N3).

Toujours suivant ce principe d'affichage (les ressources du N1, puis celles du N2 et enfin celles du N3) et depuis 2009, le résultat de la recherche d'information est présenté selon un ordre qui fait référence à l'ordre chronologique et à la pertinence des ressources[13]. La pertinence est mesurée suivant le nombre de termes de la requête identifiés, comme étant des termes d'indexation de la ressource ou identifiés au niveau du titre.

[10] Se référer au Chapitre3 ; Section 3.1.2.4 pour plus de détails concernant la terminologie CISMeF
[11] Se référer au Chapitre3 ; Section 3.1.2.4 pour plus de détails
[12] Se référer au Chapitre3 ; Section 3.1.2.4 pour plus de détails
[13] Se référer au Chapitre 5 pour plus de détails concernant le nouveau classement des résultats.

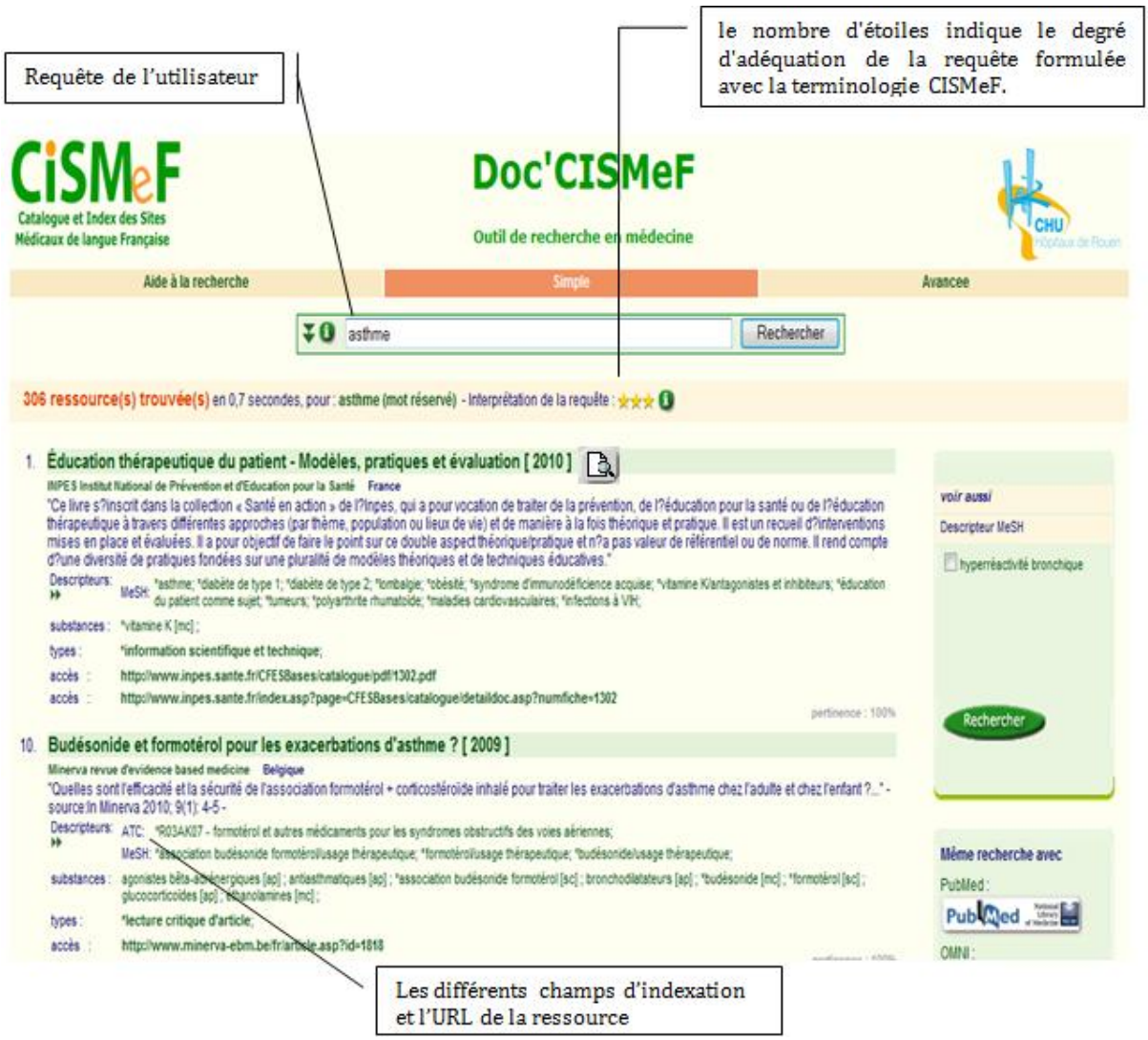


Figure 1.1.2.1.3. Le résultat de recherche pour le terme « asthme »

À chaque notice est associé un ensemble de métadonnées décrites par les documentalistes de l'équipe (les indexeurs), essentiellement issu du Dublin Core.

Cette représentation décrit :

- ✓ les informations sur le contenant de la ressource : le titre, la date de publication, le site éditeur, le type de la ressource, l'URL ;
- ✓ les informations sur le contenu de la ressource : un résumé succinct élaboré par les indexeurs, les mots clefs majeurs d'indexation décrivant les notions principales abordées dans le document ainsi que les mots clefs mineurs représentant les notions complémentaires.

Pour une ressource indexée manuellement, le clic sur le lien à droite du titre de la ressource permet d'afficher la notice détaillée contenant des informations supplémentaires concernant la ressource telles que : la langue, le pays, le mode d'accès (format de la ressource, tarif, accès), la date de création, la date de consultation…

1.1.2.2 POSITIONNEMENT DE LA THESE DANS L'EQUIPE CISMEF

Comme la plupart des systèmes de recherche d'information de qualité, de nombreux travaux ont été menés autour du catalogue CISMeF afin de préserver sa pérennité, d'améliorer la recherche d'information médicale et de faciliter la tâche de l'utilisateur.

Dans un cadre de travail interne, plusieurs thèses se sont succédées permettant l'enrichissement et le développement des stratégies entreprises au sein de l'équipe CISMeF. Par ordre chronologique des thèses en relation avec l'indexation et la recherche d'information, nous pouvons citer les travaux de L. Soualmia (Soualmia, 2004), d'A. Névéol (Névéol, 2005), de S. Pereira (Pereira, 2008) et de T. Merabti (Merabti, 2010).

Pour faciliter la tâche des utilisateurs, une recherche d'information implicite a été mise en œuvre avec le système KnewQuE (Knewledge-based Query Expansion) afin de corriger, préciser et enrichir les requêtes des utilisateurs (Soualmia et al., 2003)(Soualmia, 2004).

Concernant l'indexation des ressources du catalogue CISMeF, une tâche d'automatisation de ce processus a été étudiée pour faciliter la tâche des indexeurs face à l'explosion des documents médicaux disponibles sur le net. L'élaboration du système MAIF (MeSH Automatic Indexing in French) a été l'aboutissement de ce travail (Névéol, 2005) (Névéol et al., 2005).

Depuis peu, la stratégie de l'équipe CISMeF a été de passer d'un monde mono-terminologique vers un univers multi-terminologique[14] (cf. figure 1.1.2.2). La première réalisation dans cet univers est le développement de l'outil F-MTI (French Multi-Terminology Indexer). Il s'agit d'un outil d'aide à l'indexation automatique multi-terminologique, multi-documents et multitâches capable de produire une proposition d'indexation pour les documents de santé. Il a été appliqué notamment aux dossiers médicaux avec trois terminologies médicales supplémentaires au thésaurus MeSH (Pereira, 2008). Ce travail est poursuivi par la thèse en cours d'A. Dirieh Dibad (Dirieh Dibad et al., 2009) pour indexer les dossiers médicaux en utilisant les techniques de la sémantique d'Oracle et principalement les outils d'interrogation basés sur le SPARQL (voir chapitre 6).

Dans la même perspective de la multi-terminologie, un récent travail (Merabti, 2010) a vu le jour permettant un alignement automatique des terminologies francophones du domaine de la santé. Une telle correspondance entre ces dernières permet, en effet, d'élargir les connaissances recueillies et d'intervenir dans l'amélioration du processus de recherche d'information.

Cette thèse constitue ainsi, entre autres, une passerelle entre la mono terminologie et la multi-terminologie permettant d'avoir une indexation et une recherche d'information multi-terminologique au sein du catalogue CISMeF. De plus, nous nous sommes intéressés à une indexation bi-terminologique (le thésaurus MeSH et la classification ATC pour les médicaments) dans un cadre focalisé sur les médicaments.

[14] Voir Chapitre 3 pour plus d'explications et de détails

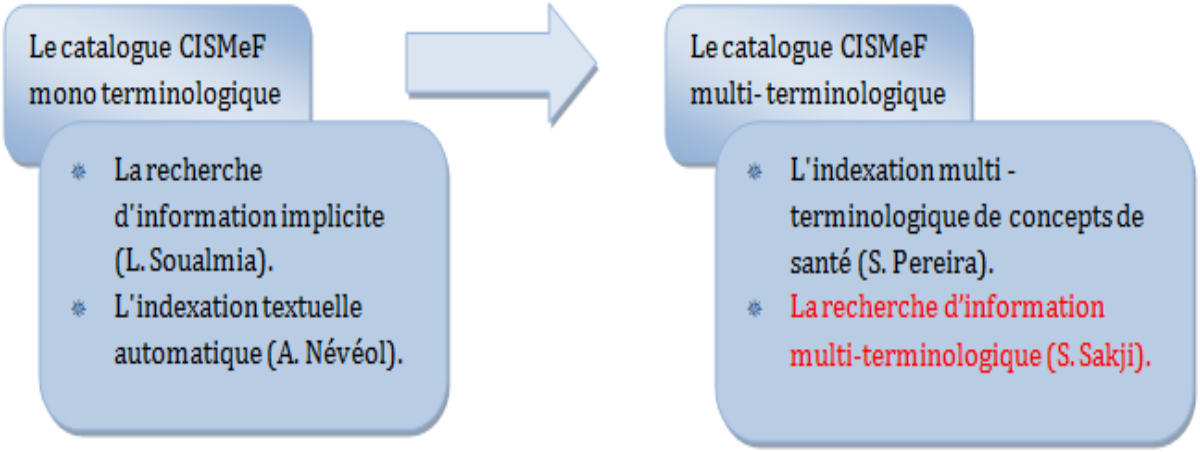


Figure 1.1.2.2. Positionnement de la thèse dans l'équipe CISMeF

1.1.2.3 QUELQUES PROJETS DE L'EQUIPE CISMEF

Dans un cadre de travail collaboratif externe, le partenariat avec des industriels met l'accent sur le développement réalisé au sein de CISMeF en termes de services et de qualité. En effet, de nombreux projets en collaboration avec des industriels ont vu le jour tels que : le portail PIH [15] (Portail Institutionnel du Handicap) crée en collaboration avec la société TEMIS [16] leader européen de la fouille des données (text mining) permettant de rechercher des informations sur le handicap.

Dans la même perspective, deux portails pour l'industrie pharmaceutique ont été réalisés avec le laboratoire Lilly[17] puis le laboratoire GSK[18].

Le moteur de recherche Doc'UMVF[19] a été créé en coopération avec l'UMVF (Université Médicale Virtuelle Francophone), comme outil de recherche en enseignement médical (Cuggia et al., 2007).

Depuis 2007, en parallèle avec cette thèse, des travaux orientés vers la problématique de la multi-terminologie ont été entrepris tel que le projet ANR InterSTIS[20] (Interopérabilité Sémantique des Terminologies dans les Systèmes d'Information de Santé français) qui a pour but de rendre interopérables les principales terminologies médicales au sein d'un serveur terminologique multi-sources.

Début 2009, un partenariat avec des laboratoires de recherche, des industriels et une société savante de médecine générale a permis de mettre en place le projet L3IM[21] (Langage Iconique et Interfaces Interactives en Médecine) qui a pour finalité d'offrir un accès rapide à des

[15] URL : http://doccismef.chu-rouen.fr/servlets/PIH
[16] URL : http://www.temis.com
[17] URL : http://www.lilly.fr/lilly/laboratoire-pharmaceutique.cfm
[18] URL: http://www.gsk.fr/
[19] URL : http://doccismef.chu-rouen.fr/servlets/ECN
[20] URL: http://www.interstis.org/
[21] URL: http://projet4-limbio.smbh.univ-paris13.fr/

informations médicales[22]. Cette approche est rendue possible grâce au langage iconique (VCM : Visualisation de Connaissances Médicales) qui permet de représenter un ensemble de concepts médicaux comme des maladies, des médicaments ou encore des examens complémentaires (Lamy et al., 2010).

1.2 LE PROJET PSIP : PATIENT SAFETY THROUGH INTELLIGENT PROCEDURES IN MEDICATION

Le projet PSIP est un projet de recherche européen, déposé le 8 mai 2007 dans le cadre de l'appel à projets « Technologies et Sciences de l'Information », pour une durée de quarante mois. Il a été labellisé en juillet 2007 et est formé de treize partenaires comprenant notamment le CHRU et l'université de Lille, le CHU de Rouen et les équipes de recherche associées, notamment notre équipe CISMeF, Vidal, Oracle, les dix centres hospitaliers de la «Région Capitale de Copenhague »…

Suite à la constatation remarquée concernant un problème majeur de santé concernant : « des effets indésirables liés aux médicaments s'observent dans 6% des séjours hospitaliers entrainant au moins 10.000 décès en France (et 98.000 aux USA) », le but de la mise en œuvre de ce projet est de proposer des méthodes innovantes destinées à contextualiser l'information et les alertes publiques (Chazard et al. 2009).

Le projet PSIP a pour objectif général de développer des services (des procédures, des systèmes de décision, des prototypes…) qui permettent de :

- ✓ identifier, grâce aux techniques d'extraction sémantique, des situations de santé quand la sécurité du patient est en danger ;
- ✓ améliorer les outils d'aide à la décision concernant les cycles de médication ;
- ✓ livrer aux professionnels de santé et aux patients, des alertes efficaces et contextuelles et des informations pertinentes au moment désiré ;
- ✓ démontrer une réduction significative du risque patient de certaines maladies et pratiques au sein d'un centre hospitalier ;
- ✓ mettre en application des outils basés sur la connaissance normalisée.

Dans un cadre scientifique, les principaux objectifs sont :

- ✓ obtenir une meilleure connaissance des effets indésirables liés aux médicaments et leurs caractéristiques, selon l'hôpital, la région et le pays ;
- ✓ développer des méthodes et des concepts pour réaliser la contextualisation des fonctions des systèmes d'aide à la décision clinique ;
- ✓ modéliser une architecture assurant l'indépendance et l'interdépendance entre la connaissance et les applications mises en jeu.

Le projet se déroule selon quatre phases :

1- génération des connaissances ;

[22] Nous utiliserons ce langage iconique pour effectuer une recherche d'information au sein de Doc'CISMeF.

2- développement d'un système de prescription contextuel intelligent ;
3- intégration et tests ;
4- évaluation.

La figure 1.2 met en évidence l'organisation du projet PSIP.

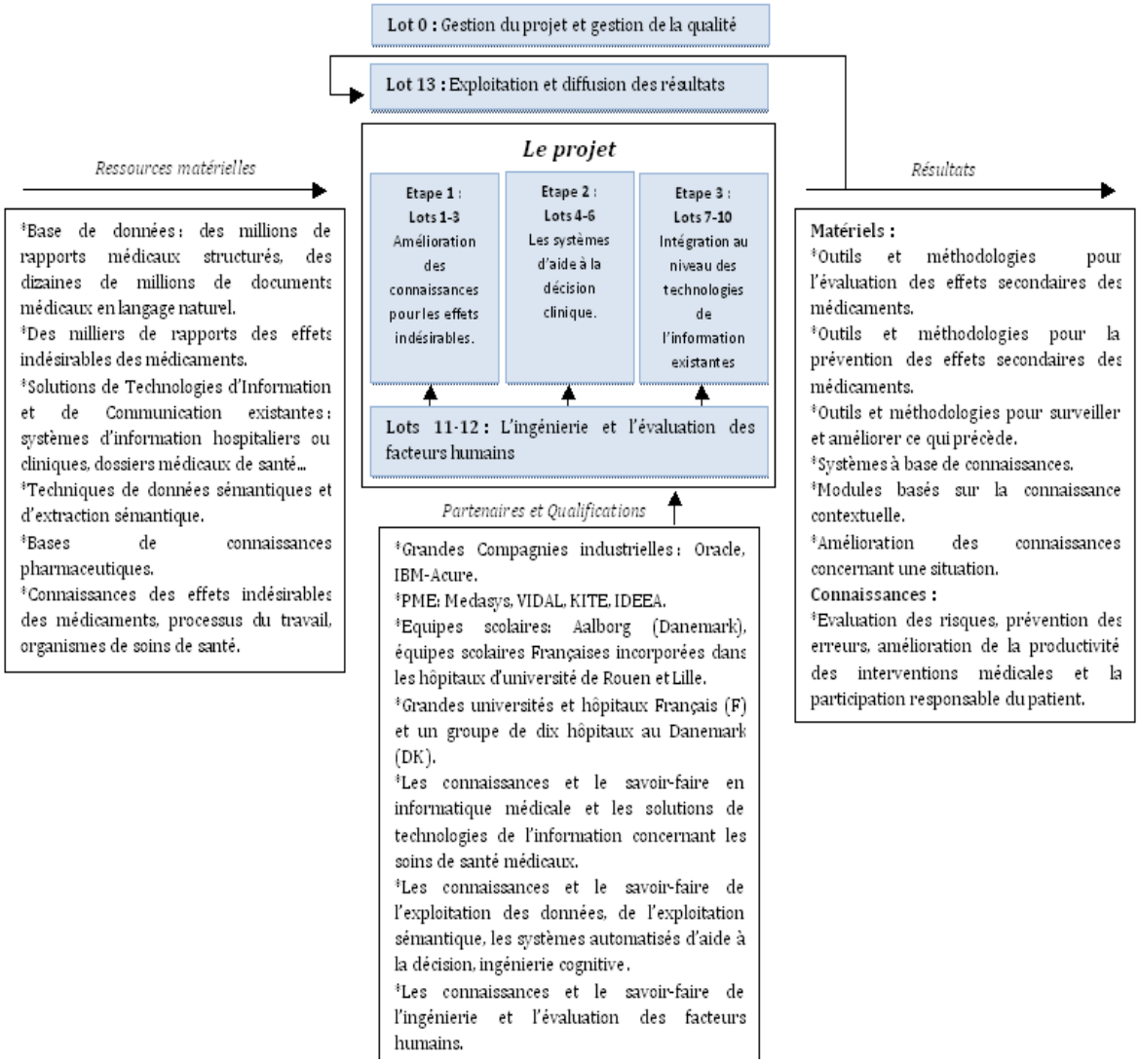


Figure 1.2. L'organisation du projet PSIP

Dans le cadre du projet PSIP, l'équipe CISMeF est en charge du semantic mining et de la création d'un portail terminologique de santé.

Dans le cadre de cette thèse, nous sommes intervenus principalement au niveau de la tâche concernant la modélisation des terminologies médicales impliquées dans ce projet à savoir la CIM-10, la classification ATC, la nomenclature IUPAC et la ICPS (Darmoni et al., 2010). Nous avons fourni aussi le modèle générique englobant toutes les terminologies médicales (même celles qui ne sont pas sollicitées dans le projet PSIP). Ce modèle est réalisé dans le cadre de cette thèse[23] et a constitué le point d'entrée vers la structure multi-terminologique du

[23] Se référer au Chapitre 3, Section 3.2 pour plus de détails

catalogue CISMeF. En effet, la généricité du modèle délivré permet une flexibilité du traitement en ajoutant, supprimant ou mettant à jour une terminologie donnée.

Au fil du déroulement du projet, nous étions amenés à tester l'outil d'extraction de concepts (Pereira, 2008) sur les comptes rendus médicaux récupérés des différents sites participants au projet. Pour les besoins du projet, nous avons amélioré l'outil en termes de performance (temps de traitement) et de couverture en terminologies en ajoutant notamment celles qui concernent les médicaments (Darmoni et al., 2009).

CONCLUSION

Nous avons présenté, dans ce chapitre introductif, le contexte général de cette thèse. Nous avons décrit les différents centres d'intérêts en termes de recherche de l'équipe CISMeF et du LERTIM auxquels j'appartiens.

Nous avons présenté, par la suite, le projet PSIP qui finance cette thèse, et qui nous a permis d'élargir nos travaux de recherche particulièrement dans le domaine des médicaments.

CHAPITRE 2
ÉTAT DE L'ART : LA RECHERCHE D'INFORMATION

INTRODUCTION

L'objectif de ce chapitre est de présenter quelques concepts de base utiles pour la compréhension du domaine de la recherche d'information (RI). Celle-ci peut être définie comme une branche de l'informatique qui s'intéresse à l'acquisition, l'organisation, le stockage, la recherche et la sélection d'information (Salton et al., 1983). Du point de vue de l'utilisateur, l'accès à l'information peut être effectué de manière délibérée à travers un système de recherche d'information (SRI) ou bien de manière passive à travers un système de filtrage d'information. Après une brève présentation de la recherche documentaire, nous définissons, dans ce chapitre, quelques systèmes de recherche d'information, leurs particularités et leurs fonctionnements. Nous présentons par la suite quelques notions d'indexation, puis nous passons en revue les modèles piliers de la RI et les critères et les mesures d'évaluation des SRI.

2.1 LE PRINCIPE DE LA RECHERCHE DOCUMENTAIRE

La recherche documentaire vise à retrouver des documents textuels répondant à un besoin informationnel spécifié par une requête.

Parmi les définitions exposées dans la littérature, nous pouvons citer celle de (Lewis, 1992) qui résume les étapes de la recherche d'information comme suit :

- ✓ l'indexation des textes : elle permet de représenter le contenu des documents et la requête de l'utilisateur (en langage naturel, requête booléenne, document entier comme un exemple du résultat, un graphe de concepts…) afin qu'ils soient exploitables par le système de recherche d'information ;
- ✓ la comparaison entre la représentation de la requête et celle des documents. La comparaison se fait généralement en utilisant une fonction de similarité. Le processus de comparaison permet de choisir les documents répondant au besoin d'information de l'utilisateur, en comparant la base des index du corpus à la représentation de la requête dans le même espace. Cette phase vise à extraire des caractéristiques sur le contenu sémantique des informations textuelles ;
- ✓ le feedback : le résultat retourné par le système peut ne pas correspondre aux résultats attendus de l'utilisateur, ce qui amène ce dernier à reformuler sa requête.

2.2 LES SYSTEMES DE RECHERCHE D'INFORMATION

Plusieurs définitions des SRI ont été établies et sont plus ou moins semblables. Parmi lesquelles, nous citons celle de (Smeaton, 1992) (Smeaton, 1999) « l'objectif d'un système de recherche d'information est de trouver des documents en réponse à une requête d'utilisateur tel que le contenu des documents soit pertinent par rapport au besoin initial de l'utilisateur ». Une autre définition (Strzalkowski, 1993) suggère que « la tâche typique de la recherche d'information est de sélectionner des documents dans une base de données, en réponse à une requête de l'usager, et de les ranger par ordre de pertinence ».

Pour résumer, nous pouvons dire que la tâche principale d'un système de recherche d'information est de sélectionner dans une collection de documents ceux qui sont susceptibles de répondre aux besoins en information de l'utilisateur. Son but est de retourner à ce dernier le maximum de documents pertinents pouvant satisfaire son besoin et le minimum de documents non pertinents. Dans son livre (Blair, 1990), Blair met l'accent sur la complexité des systèmes de recherche d'information pour fournir un bon résultat dans la mesure où ils nécessitent un langage précis pour mettre les termes dans leurs contextes ce qui manque, d'après lui, aux SRI. (Tamine-Lechani et al., 2007) définissent les systèmes centrés utilisateurs. Dans leur travaux, ils mettent en évidence l'adaptation du cycle de vie d'un processus d'accès à l'information, à un utilisateur spécifique, en vue de lui délivrer une information pertinente relativement à ses besoins précis, son contexte et ses préférences.

Ainsi, pour répondre aux besoins en information de l'utilisateur, un SRI met en œuvre un certain nombre de processus pour réaliser la mise en correspondance des informations contenues dans le fonds documentaire d'une part, et les besoins en information des utilisateurs, d' autre part. Parmi les représentations des SRI qu'on peut trouver dans la littérature, nous pouvons citer celle de (Van Rijsbergen, 1979) qui les représente sous forme de trois principales composantes : *input*, *processor* et *output*. Nous nous intéressons du plus près à la représentation de (Boughanem et al., 2008) qui les définit sous la forme de

« processus en *U* » (cf. Figure 2.2). La première étape consiste à indexer les documents et les requêtes. Le résultat de l'indexation est une représentation paramétrée qui couvre au mieux le contenu sémantique des documents et des requêtes. L'ensemble des termes reconnus par le SRI constitue le langage d'indexation.

La deuxième étape est réalisée grâce au processus d'appariement qui permet de comparer la représentation des documents d'une collection donnée et celle de la requête de l'utilisateur dans un même espace de représentation. Cette comparaison a pour finalité de permettre de choisir les documents répondant au besoin d'information de l'utilisateur.
L'appariement requête-documents consiste à calculer un score, supposé représenter la pertinence du document vis-à-vis de la requête. Le score est souvent calculé à partir d'une fonction de similarité qui tient compte du poids des termes dans les documents. L'assignation d'un score de pertinence à un document permet d'ordonner les documents renvoyés à l'utilisateur, et ce qui peut influencer le jugement de l'utilisateur vis-à-vis du SRI.

La troisième étape est la reformulation de la requête de l'utilisateur (en cas de besoin) afin de faire correspondre au mieux la pertinence-utilisateur et la pertinence système. La reformulation de la requête consiste généralement à rajouter de nouveaux termes à la requête initiale, et/ou à re-pondérer ses termes dans la nouvelle requête.

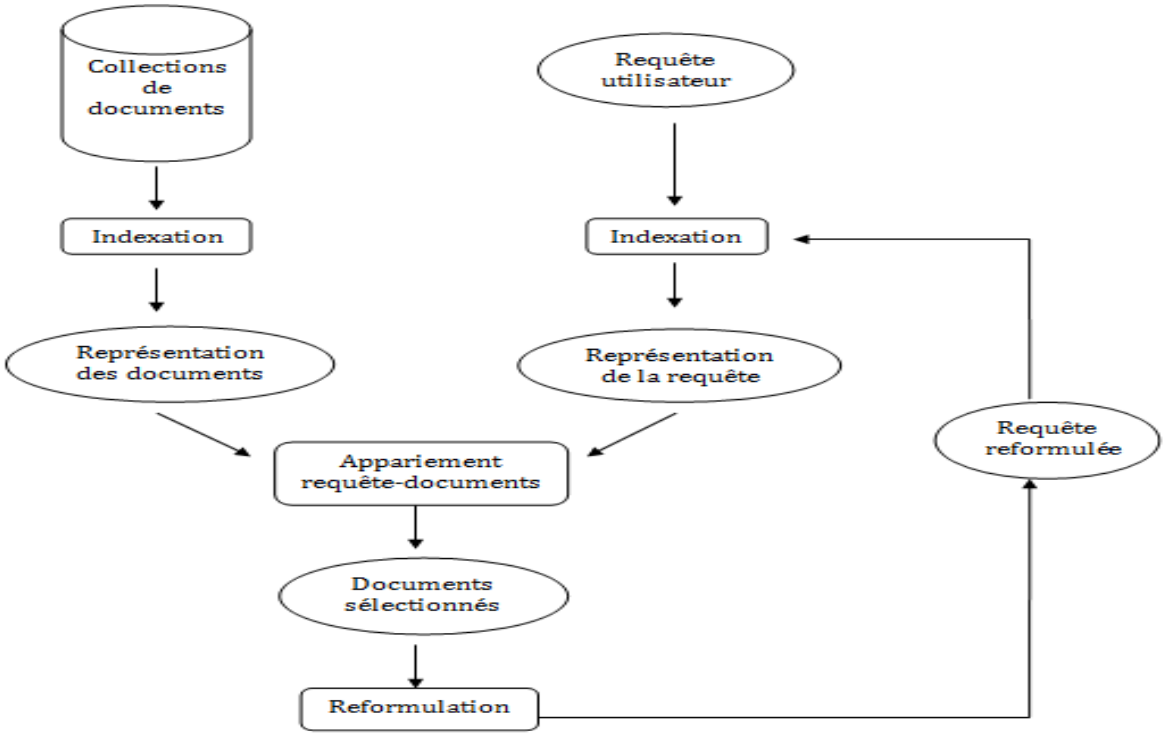


Figure 2.2. Processus en *U* de recherche d'information

2.3 L'INDEXATION

Le processus d'indexation consiste à extraire des documents les termes (ou concepts) jugés les plus significatifs et pertinents afin d'en construire une représentation médiatrice entre les documents et les utilisateurs. Il s'agit de décrire leurs contenus et de les représenter par des

index. Prie (Prie, 2000) définit l'index comme « quelque chose » permettant d'accéder à « autre chose ». Ainsi, l'indexation consiste à mettre en place des objets permettant d'accéder à d'autres objets tels que des documents, des parties de documents ou encore des ensembles de documents. Cette représentation a pour but de retrouver les ressources documentaires qui répondent au mieux à la requête de l'utilisateur parmi un grand nombre de ressources.

L'indexation est une étape préliminaire à la recherche documentaire. C'est une tâche importante car elle définit l'espace de représentation de l'information contenue dans les textes et influence par conséquent la fonction de comparaison.

Pour pouvoir être comparées, il faut que la représentation de la requête et la représentation d'un document soient exprimées dans le même espace d'indexation. Ceci permet de répondre à la question suivante : *comment retrouver un document pertinent pour une requête alors qu'ils ne sont pas indexés (donc représentés) par le même vocabulaire ?*

Le processus d'appariement permet de comparer la représentation des documents d'une collection donnée et celle de la requête de l'utilisateur dans un même espace de représentation. Cette comparaison a pour finalité de permettre de choisir les documents répondant au besoin d'information de l'utilisateur.

2.3.1 LES LANGAGES D'INDEXATION

Le rôle des descripteurs représentatifs d'un document lors de la phase d'indexation est double (Salton et al., 1983). D'une part, ils doivent être descriptifs, pour bien décrire le contenu du document et d'autre part, discriminants afin de mettre en évidence ce qui distingue le document dans la collection des documents.

Nous pouvons distinguer deux types d'indexation, utilisant des langages d'indexation différents :

- ✓ *l'indexation libre* utilise librement tous les mots d'une langue naturelle donnée : il s'agit d'un ensemble ouvert de termes. L'indexation d'une ressource consiste en une liste de tous les mots du *langage naturel* contenus dans la ressource, auquel un filtrage ou une certaine normalisation pourront être appliqués (Salton et al., 1983) ;
- ✓ *l'indexation contrôlée* utilise des termes appartenant à une liste de référence prédéfinie (un langage connu) : un ensemble fermé de termes. Ce *langage contrôlé* définit la forme des termes d'indexation utilisés. Il peut s'agir de termes ou d'expressions de la langue naturelle ou bien de symboles choisis pour représenter un concept de manière normative et unique.

  Dans le cas d'une indexation contrôlée, une connaissance approfondie du vocabulaire est nécessaire pour une indexation de qualité. De plus, une mise à jour du vocabulaire devrait être accompagnée par une révision de l'indexation déjà faite sur les documents.

Plusieurs études et discussions ont été faites sur le type d'indexation qu'il faut choisir. Certains travaux concluent que l'utilisation d'un vocabulaire contrôlé lors du processus d'indexation donne des résultats équivalents ou légèrement supérieurs pour la recherche d'information. Les études de (Leonard, 1977) et (Markey, 1984) montrent que la consistance de l'indexation augmente en moyenne de 15% avec l'utilisation d'un vocabulaire contrôlé.

(Leininger, 2000) estime que le choix d'un vocabulaire contrôlé pour l'indexation des ressources d'une base documentaire permet de favoriser la précision lors de la recherche d'information, alors qu'une indexation libre favoriserait le rappel. Il observe également que l'utilisation d'un vocabulaire contrôlé est conditionnée par l'existence d'un thésaurus adapté à la base documentaire considérée.
Dans la même perspective, le grand nombre de terminologies médicales fait de la médecine un domaine particulièrement propice à l'indexation contrôlée. Une étude de la National Library of Medicine (NLM)[24] (Wilbur et al., 2003) met en évidence l'avantage de l'utilisation des termes appartenant au thésaurus MeSH[25] par rapport à l'utilisation des termes en langage naturel.

2.3.2 LES TYPES D'INDEXATION

L'indexation consiste à identifier, dans un document, certains éléments significatifs qui serviront de clé pour retrouver ce document au sein d'une collection. Le choix du type d'indexation dépend des applications et de la taille du corpus étudié. Cette identification peut être :

- ✓ manuelle : chaque document de la collection est analysé par un documentaliste ou un spécialiste du domaine d'application ;
- ✓ automatique : le processus d'indexation est entièrement informatisé ;
- ✓ supervisée (dite aussi semi-automatique) : suite à l'indexation automatique appliquée aux documents, l'indexeur (le documentaliste ou le spécialiste du domaine) intervient pour valider le choix des termes représentatifs des documents.

2.3.2.1 L'INDEXATION MANUELLE

L'indexation manuelle est effectuée par des experts qui ont pour tâche d'analyser les documents, comprendre et identifier leurs contenus afin de construire une bonne représentation. Cette indexation permet d'obtenir une caractérisation assez performante mais subjective car elle dépend des compétences de l'indexeur en termes de connaissances et d'esprit analytique. En effet, même quand l'indexation s'appui sur un langage contrôlé, la représentation d'un même document (l'index généré) peut être différente selon l'interprétation personnelle des indexeurs ou encore à des moments différents pour le même indexeur (Le Loarer, 1994). Bien que les indexeurs suivent tous les mêmes procédures et les règles éditoriales propres à la collection documentaire pour analyser les documents, leurs critères d'appréciation de ce qui constitue une bonne indexation (la décision de conserver ou de rejeter un descripteur) semblent varier (David et al., 1995).

Par ailleurs, l'indexation manuelle est très coûteuse en temps. Comme exemple, la NLM dispose d'une moyenne de 120 indexeurs pour 712.675 articles indexés pour MEDLINE et l'équipe CISMeF de 4 indexeurs pour 39.874 ressources indexées manuellement.

[24] United States National Library of Medicine. URL : http://www.nlm.nih.gov/
[25] Se référer au Chapitre 3, Section 3.1.2.3 pour la définition du thésaurus MeSH

Ainsi, face à des bases de données de grande taille, l'indexation manuelle peut être une entrave au bon fonctionnement du processus d'indexation ce qui peut diminuer les performances des SRI.

2.3.2.2 L'INDEXATION AUTOMATIQUE

L'indexation automatique, processus entièrement informatisé, regroupe un ensemble de traitements automatisés sur un document ce qui le rend avantageux, par rapport à l'indexation manuelle, en terme de régularité de l'index. En effet, pour le même document, nous avons toujours la même représentation. Ceci peut être aussi un inconvénient du fait qu'il n'y aurait pas une adaptation aux nouveaux éventuels vocabulaires appliqués. Néanmoins, face à des bases de données de très grande taille, l'indexation automatique se révèle la seule possible pour le bon fonctionnement des SRI.

L'indexation automatique repose sur des algorithmes associant automatiquement des descripteurs à des parties de document. Chaque mot est, potentiellement, un index du paragraphe qui le contient.

L'indexation automatique tend donc plutôt à rechercher les mots qui correspondent au mieux au contenu informationnel d'un document. On admet, généralement, qu'un mot qui apparaît souvent dans un texte représente un concept important. Ainsi, la première étape consiste à déterminer les mots représentatifs par leur fréquence. Cependant, on s'aperçoit que les mots les plus fréquents sont des mots fonctionnels (mots vides) tels que *de*, *un*, *les*... Ainsi, après l'élimination de ces mots vides, un traitement est ensuite, couramment, appliqué lors de l'indexation pour effacer les terminaisons des mots (flexions de nombre, genre, conjugaison, déclinaison) et retrouver leurs racines. Il s'agit soit de la désuffixation soit de la lemmatisation. Ce procédé permet de relever les fréquences en cumulant les nombres d'occurrence des variations des mêmes mots.

Les techniques de *désuffixation* permettent de supprimer pour une bonne part les variations morphologiques. Elles visent à supprimer les suffixes qui sont souvent utilisés pour créer des dérivées d'un terme ce qui permet de trouver les racines lexicales. Une comparaison entre différents algorithmes développés pour cet effet a été menée dans (Hull, 1996). Pour chaque langue, des règles différentes peuvent être appliquées, d'où la nécessité d'une adaptation algorithmique : par exemple l'algorithme le plus connu pour la langue anglaise est celui de (Porter, 1980). Pour la version française, nous citons celui de Carry (Paternostre et al. 2002) ou celui de Lucene (Hatcher et al., 2004). Le dictionnaire formé suite à cette phase d'analyse sera donc composé de radicaux.

La *lemmatisation* consiste chercher le « lemme » des mots. En somme, nous débarrassons les mots de leur genre, leur nombre, leur personne (toi, moi, etc.), leur mode (impératif, indicatif, etc.). Nous transformons, donc, tous les verbes à l'infinitif et les mots au masculin singulier. Pour y arriver, il faut déterminer le mode, le genre, etc., des mots et trouver les verbes et les autres catégories de mots ; ce qui exige une connaissance de la grammaire. Un algorithme efficace, nommé TreeTagger (Schmid, 1994), a été développé pour les langues anglaise, française, allemande et italienne.

Dans le même cadre de travail, le projet UMLF (Unified Medical Lexicon for French) (Zweigenbaum et al., 2003) a vu le jour pour effectuer la collecte, la synthèse, la complétion et la validation de ressources lexicales pour le français médical. Par une approche monolingue, il vise à produire un lexique contenant les variantes flexionnelles et dérivationnelles des mots du domaine. Ces informations doivent être encodées dans un format informatique standard afin de favoriser leur intégration dans des systèmes de traitement automatique de la langue médicale. Ce besoin est survenu pour pallier le manque du lexique médical informatisé français.

La désuffixation repose sur des contraintes linguistiques bien moins fortes du fait qu'elle se base sur la morphologie flexionnelle (par exemple les formes conjuguées d'un verbe avec son infinitif) et dérivationnelle (par exemple un adjectif avec le substantif associé lent/lenteur) (De Loupy, 2001). De ce fait, les algorithmes sont beaucoup plus simplistes et rapides que ceux permettant la lemmatisation qui est beaucoup plus complexe. Il n'est pas certain que la lemmatisation soit toujours requise : la désuffixation, bien que moins efficace, peut suffire.

Par ailleurs, des formules de pondération sont appliquées pour affecter, généralement, un poids élevé aux termes non-distribués uniformément au sein du corpus. Il existe plusieurs formules de pondération dont le but est de distribuer le poids pour contribuer à la différentiation informationnelle des documents. Certaines formules de pondération harmonisent les poids en fonction de la longueur des documents où la fréquence des termes est, globalement, plus élevée. D'autres formules s'appuient sur la fréquence maximale des termes afin de concilier l'aspect multi-thématique d'un document avec des documents mono-thématiques. Les formules de pondération les plus connues sont TF-IDF (Term Frequency. Inverse Document Frequency) (Salton et al., 1983).

Les principales limites de l'indexation automatique est que, les algorithmes exploitent l'information contenue *dans* les documents alors que l'interprétation doit se guider depuis l'information contextuelles accessible *hors* des documents.

Se comparant à l'indexation manuelle, on obtient de manière automatique des descripteurs qui reflètent le contenu *physique* des documents. En effet, l'indexation manuelle permet d'obtenir des concepts interprétant le document dans son contexte.

2.3.2.3 L'INDEXATION SUPERVISEE

L'indexation supervisée tient compte de l'indexation automatique réalisée d'une manière informatisée et est vérifiée, par la suite, par les indexeurs (documentalistes ou spécialistes du domaine d'application) afin de valider la représentation proposée. Cette méthode d'indexation doit être considérée comme un compromis entre l'indexation manuelle et l'indexation automatique.

Plusieurs études et évaluations ont été faites comparant l'indexation manuelle et l'indexation automatique, mettant en relief les avantages et les limites de l'une par rapport à l'autre. Basés sur la collection INSPEC[26] (de 12.684 documents, 84 requêtes) Rajashekar et Croft

[26] INSPEC (***In****formation* ***S****ervice for* ***P****hysics,* ***E****lectronics, and* ***C****omputing)* a été lancée en 1969 par l'IEE (**I**nstitution of **E**lectrical **E**ngineers) à partir de la collection *Science Abstracts*. En 2006, elle

(Rajashekar et al., 1995) jugent que l'indexation automatique présente une performance moyenne supérieure à l'indexation manuelle. La comparaison était réalisée sur le titre et le résumé des documents.

Dans la même perspective, Savoy (Savoy, 2005) compare la performance de l'indexation automatique (sur la base du titre et du résumé d'articles scientifiques) et manuelle, dont les termes appartiennent à un vocabulaire contrôlé. En se basant sur un corpus relativement important de notices bibliographiques[27] (148.688 documents et 25 requêtes) et des requêtes courtes (en moyenne 3,7 termes par requête) ou de longueur moyenne (15,6 termes), l'auteur juge que l'indexation manuelle permet une meilleure précision moyenne par rapport à l'indexation automatique

2.4 LES MODELES DE RECHERCHE D'INFORMATION

Un modèle de recherche d'information a pour rôle de fournir une formalisation du processus de recherche d'information.

Dans la littérature, nous trouvons plusieurs modèles décrits permettant, entre autres, une recherche d'information dite « classique » (Baeza-Yates et al., 1999). Parmi ces modèles, nous trouvons le modèle booléen, le modèle booléen étendu (Salton et al., 1983), le modèle vectoriel (Salton et al., 1975), le modèle vectoriel étendu (Martinet et al., 2002), le modèle logique (Van Rijsbergen, 1986) (Nie, 1990) et le modèle probabiliste (Van Rijsbergen, 1979).

2.4.1 LE MODELE BOOLEEN & LE MODELE BOOLEEN ETENDU

Le modèle booléen doit son nom à l'utilisation des opérateurs logiques de l'algèbre de Boole « et » « ou » et « non » pour la représentation des documents et des requêtes. Un document (ou une requête) est représenté par une conjonction de termes. La fonction de comparaison retrouve les documents dont les index correspondent à la représentation logique de la requête. Ainsi, nous aurons comme résultat un ensemble de documents qui correspondent à la requête et un deuxième ensemble de documents qui ne correspondent pas à la requête. Ce modèle booléen est reconnu pour sa force pour faire une recherche très restrictive et obtenir, pour un utilisateur expérimenté, une information exacte et spécifique.

Les inconvénients de ce modèle se résument dans le fait que les documents pertinents dont la représentation ne correspond qu'approximativement à la requête ne sont pas sélectionnés. En plus, tous les termes d'indexation ont la même importance et, par conséquent, ce modèle est incapable de trier les documents résultats selon leur degré de pertinence.

---

propose des références bibliographiques issues de 3 850 journaux scientifiques et techniques et d'environ 2 200 actes de conférence, plus des livres, rapports et thèses du domaine de la physique, de l'électronique et du génie électrique, du génie informatique et de la télématique, des technologies de l'information.

[27] Le corpus utilisé fait partie de l'évaluation CLEF 2002 qui se compose de 148.688 références bibliographiques rédigés en français et appartenant aux collections FRANÇAIS (pour les sciences sociales et humaines) et PASCAL (pour les sciences naturelles, la technologie et la médecine) de l'INIST (INstitut de l'Information Scientifique et Technique).

Pour remédier à ces inconvénients, le modèle booléen étendu (Salton et al., 1983) a été proposé. Il tient compte de l'importance des termes dans la représentation des documents et dans la requête, et ce, en affectant des poids aux termes d'indexation. Ainsi, ce modèle permet un ordonnancement des documents par rapport à la valeur de correspondance avec la requête.

2.4.2 LE MODELE VECTORIEL & LE MODELE VECTORIEL ETENDU

Le modèle vectoriel représente un document (ou une requête) par un vecteur dans un espace à *n* dimensions, *n* étant le nombre de termes du langage d'indexation. Les coordonnées des vecteurs sont les poids indiquant l'importance du descripteur par rapport au document.

La fonction de comparaison évalue la correspondance entre les deux vecteurs (du document et de la requête) et cherche à retrouver les vecteurs des documents qui s'approchent le plus du vecteur requête. Ainsi, les documents sont triés et classés selon une mesure de similarité.

Ce modèle est l'un des modèles de RI classique les plus influents, les plus étudiés et les mieux adaptés. Le système SMART (Salton, 1971) est un des premiers systèmes de recherche d'information basé sur ce modèle.

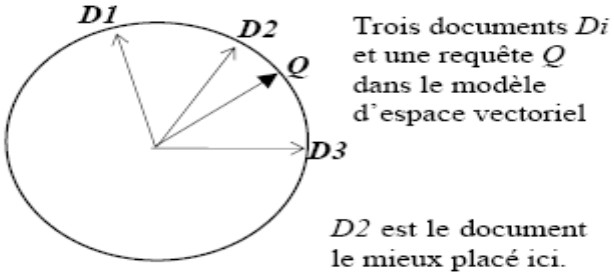


Figure 2.4.2. Le modèle vectoriel

Chaque document est représenté par un vecteur : $D_j = (d_{t1j}, d_{t2j}, d_{t3j}, ..., d_{tnj})$,
Chaque requête est représentée par un vecteur : $Q = (q_{t1}, q_{t2}, q_{t3}, ..., q_{tn})$,
Avec : d_{tij} : poids du terme ti dans le document D_j,

q_{ti} : poids du terme ti dans la requête Q.

Les coordonnées des vecteurs d_{tij} sont calculées à partir de la fréquence des termes dans les documents par la formule tf-idf.

tf : la fréquence du terme dans le document.
Idf : l'importance du terme dans tout le corpus de documents, qui est la fonction inverse du nombre de documents indexés par ce terme.

Ainsi,

$d_{tij} = tf_{ti} * idf_{ti}$ avec tf= fréquence du ti dans d_j

$$\text{et} \quad \text{idf} = \frac{1}{\text{nb de documents} \quad \text{index és par ti}}$$

Un terme d'indexation avec une forte pondération est un terme fréquent dans un document et absent des autres documents. Cette pondération amplifie considérablement l'importance des termes étrangers, des noms propres. Comme l'indexation est complètement automatisée, les termes rares qui risquent d'être peu utilisés pour la recherche d'information sont privilégiés.

Dans son approche, Salton (Salton et al., 1975) fait l'hypothèse que les mots sont indépendants les uns des autres. La fonction de comparaison se base sur le cosinus de l'angle formé entre les deux vecteurs : plus l'angle est petit, plus les vecteurs sont similaires.

Ainsi, la fonction de similarité entre un document et la requête s'écrit sous cette forme :

$$Sim\left(D_j, Q\right) = \frac{\sum_{i=1}^{n} q_{ti}\, d_{ij}}{\sqrt{\sum_{i=1}^{n} {q_{ti}}^2 \sum_{i=1}^{n} {d_{ij}}^2}}$$

Dans la même perspective, les travaux de Martinet (Martinet et al., 2002) se sont basés sur une extension du modèle vectoriel concernant la nature des termes d'indexation, la représentation multi-vectorielle des documents ainsi que la fonction de correspondance adaptée à cette représentation. Ils ont appliqué leurs travaux aux documents images et ont implanté le modèle vectoriel étendu à l'aide de SMART (Salton, 1971). La mise en œuvre du modèle a été faite sur une base d'images décrites par des concepts et de relations.

2.4.3 LE MODELE PROBABILISTE

Le modèle de recherche probabiliste utilise un modèle mathématique fondé sur la théorie de la probabilité (Robertson et al., 1976) (Salton et al., 1983) (Maron et al., 1960) (Robertson, 1997). Le principe de base consiste à présenter le résultat dans un ordre fondé sur la probabilité de pertinence d'un document par rapport à la requête.

Étant donné une requête utilisateur *Q* et un document *D*, la question qui se pose est : « Pour chaque document *D* et chaque requête *Q*, quelle est la probabilité que ce document soit pertinent pour cette requête ? »

Deux possibilités se présente R : D est pertinent pour *Q*

$\overline{R}$: D est non pertinent pour Q

(Boughanem et al., 2008) mettent en relief ce modèle probabiliste et explicitent les hypothèses et les différents postulats à tenir en compte pour pouvoir estimer le degré de pertinence des documents par rapport à la requête de l'utilisateur. En effet, le modèle probabiliste tente d'estimer la probabilité que le document *D* appartienne à la classe des documents pertinents (non pertinents). Un document est alors sélectionné si la probabilité qu'il soit pertinent pour *Q*, notée P(R/D), est supérieure à la probabilité qu'il soit non pertinent pour *Q*, notée P ($\overline{R}$/Q). Le score d'appariement entre le document *D* et la requête *Q*, noté RSV (Q, D) (Robertson et al., 1994) est donné par :

$$\text{RSV (Q, D)} = \frac{P(R/D)}{P(\bar{R}/D)}$$

Si l'on applique la formule de Bayes, nous avons :

$$P\ (R/D) = \frac{P(D/R)P(R)}{P(D)} \text{ et } P\ (\bar{R}/D) = \frac{P(D/\bar{R})P(\bar{R})}{P(D)}$$

En supposant que les documents aient tous la même probabilité d'être sélectionnés et que la sélection d'un document soit indépendante d'un autre, le terme P(D) peut être supprimé. Nous obtenons alors :

$$RSV\ (Q, D) = \frac{P(D/R)P(R)}{P(D/\bar{R})P(\bar{R})}$$

Le terme $\frac{P(R)}{P(\bar{R})}$ est le même pour tous les documents de la collection, un classement de document avec RSV (Q, D) revient donc au classement suivant :

$$RSV\ (Q, D) = \frac{P(D/R)}{P(D/\bar{R})}$$

Plusieurs méthodes ont été utilisées pour estimer les différentes variables utilisées par les modèles probabilistes. nous pouvons trouver le modèle d'indépendance binaire qui considère que la variable document $d(t_1=x_1, t_2=x_2, \ldots, t_n=x_n)$ est représenté par un ensemble d'événements qui dénotent la présence ($x_i=1$) ou l'absence ($x_i=0$) d'un terme dans un document. Les probabilités de pertinence (non pertinence) d'un document, notées P (D/R) (resp. P (D/$\bar{R}$)) sont données par :

$$P\ (D/R) = \prod i\ P(t_i = x_i/R)$$

$$P\ (D/\bar{R}) = \prod i\ P(t_i = x_i/\bar{R})$$

t_i est le $i^{ème}$ terme utilisé pour décrire le document D et x_i est sa valeur 0 si le terme est absent, 1 si le terme est présent dans un document. La distribution des termes suit une loi de Bernoulli P (D/R) et peut s'écrire comme suit:

$$P\ (D/R) = \prod_{i=1}^{n} P(t_i = x_i/R) = \prod_{i=1}^{n} P(t_i = 1/R)^{x_i} * P(t_i = 0/R)^{1-x_i}$$

Nous réalisons le même développement pour P (D/$\bar{R}$). Notons P (t_i=1/R) par p_i et P(t_i=1/$\bar{R}$) par q_i, RSV peut s'écrire, après transformation, comme suit :

$$RSV\ (Q, D) = \prod_{i=1}^{n} \frac{p_i^{x_i}(1-p_i)^{(1-x_i)}}{q_i^{x_i}(1-q_i)^{(1-x_i)}}$$

En se ramenant à la fonction log et après un petit développement, la fonction RSV s'écrit alors :

$$RSV\ (Q, D) = \sum_{i:x_i=1} log \frac{p_i(1-q_i)}{q_i(1-p_i)}$$

Différents postulats pour l'estimation de p_i et q_i produisent différentes fonctions d'ordonnancement des documents. Par exemple Croft (Croft et al., 1979) considère que p_i est la même pour tous les termes de la requête et $\frac{p_i}{1-p_i}$ est constante et peut être ignorée lors du classement. En plus, il considère que la plupart des documents de la collection sont non pertinents et estime par conséquent q_i égale $\frac{n_i}{n}$ avec n est la taille de la collection et n_i est le nombre des documents qui contiennent le terme i, ce qui nous amène à :

$$\text{RSV (Q, D)} = \sum_{i:x_{i=1}} log \frac{n-n_i}{n_i}$$

Un obstacle majeur avec les modèles de recherche d'information probabilistes est de trouver des méthodes pour estimer les probabilités utilisées pour évaluer la pertinence qui soient théoriquement fondées et efficaces au calcul (Crestani et al., 1998). Pour des raisons de simplicité, l'hypothèse de l'indépendance des termes est utilisée en pratique pour implémenter ces modèles.

2.4.4 LE MODELE LOGIQUE

(Van Rijsbergen, 1986) modélise la pertinence d'un document répondant à une requête par une implication logique. Soit x(d) l'information contenue dans le document *d* et x(q) le besoin informationnel de l'utilisateur formulé par la requête *q*. L'expression de l'information est faite grâce aux formules logiques. Ainsi, le système cherche à évaluer l'ajout minimal d'information nécessaire pour obtenir l'implication $x(d) \rightarrow x(q)$, permettant de classer les documents résultats.

De nombreux modèles logiques ont été proposés depuis. Chevallet (Chevallet, 2004) propose un certain nombre d'hypothèses pour modéliser la pertinence avec la logique. La première hypothèse est que le processus de RI est formalisable. Les documents et les requêtes peuvent être formalisés et la formalisation d'un document est une opération bijective. Nous supposons qu'il existe un mécanisme de correspondance qui calcule l'ensemble des documents qui sont pertinents pour une requête.

Par ailleurs, nous supposons qu'il existe une relation de pertinence entre un document et une requête s'il existe une chaîne de déductions logiques incertaines commençant par le document pour aboutir à la requête. Le calcul de pertinence se résume alors à prouver que $d \rightarrow q$, $\rightarrow$ représente un lien *logique incertain de pertinence* entre le document *d* et la requête *q*.

La modélisation la plus simple consiste à utiliser la logique des propositions. Si l'on considère que l'ensemble des termes d'indexation (et que l'ensemble des termes atomiques de la logique) est $\{t_1,..., t_n\}$ et que le document *d* est indexé par $\{t_1, ...t_i\}$ alors nous pouvons représenter *d* par la formule suivante: $t_1 \wedge ... \wedge t_i \wedge \neg t_{i+1} \wedge ... \wedge \neg t_n$ (*d* est une interprétation logique des termes atomiques). *d* est pertinent pour *q* si et seulement si $\models d \rightarrow q$; avec $\rightarrow$ l'implication logique classique.

Le problème de la logique classique est qu'elle n'offre aucune souplesse (un terme est présent ou non dans un document) et, dès que l'on a moins d'information sur le document, le système

de RI ne peut plus répondre à des requêtes précises. D'autres modélisations logiques ont été proposées. Par exemple, nous pouvons utiliser une logique modale. Les documents sont alors représentés par des mondes et les requêtes sont des formules qui peuvent être vérifiées dans les mondes. Il existe une relation d'accessibilité entre les mondes. Cette dernière peut par exemple représenter le fait que deux documents (deux mondes) sont accessibles s'ils contiennent des termes synonymes ou s'il existe un hypertexte pour passer d'un document à l'autre. Nous avons alors *d* pertinent pour *q* si et seulement si d |= ◊q (◊: possibilité).

Nous pouvons rajouter une couche de probabilité à ce modèle modal. Cela ajoute une notion d'incertitude sur la pertinence : par exemple, nous pouvons passer d'un monde à un autre avec une certaine probabilité. Dans (Crestani et al., 2001), différents modèles logiques incertains sont décrits.

2.4.5 AUTRES MODELES DE RECHERCHE D'INFORMATION

- **Le modèle Latent Semantic Indexing (LSI)**

Latent Semantic Indexing (LSI) est une technique mathématique/statistique pour extraire et représenter le sens entre les termes. Comparativement au modèle vectoriel, la technique LSI réduit la dimension de l'espace de représentation aux seuls vecteurs de représentation de l'information sémantique, et ce, en réduisant l'effet de variation d'utilisation des termes. Dans ce modèle, les documents sont représentés dans un espace de dimension réduit issu de l'espace initial des termes d'indexation (Deerwester et al., 1990). Les documents partageant des termes co-occurrents ont des représentations proches dans l'espace de représentation. Ceci permet de sélectionner des documents pertinents même s'ils ne contiennent aucun terme de la requête. Ainsi, le LSI est défini comme étant une technique qui tend à implanter partiellement la recherche sémantique ou orientée concepts (Dumais, 1995) (Bradford, 2006).
L'avantage de la méthode est qu'elle arrive à une représentation pseudo-conceptuelle des documents de la base, permettant de retrouver des documents même s'ils ne contiennent pas les mots des requêtes.

Son inconvénient est qu'elle est sensible à la quantité et à la qualité des données traitées. Si par exemple, le nombre de documents est faible, alors le calcul d'une approximation de la taille de la collection pourrait aboutir à des faux résultats.

- **Le modèle connexionniste**

Le premier modèle connexionniste pour la RI a été présenté en 1989. Il constitue un support formel opportun pour la modélisation de l'apprentissage dans un système de recherche d'information. Ce type de modèle se base sur le formalisme des réseaux de neurones (Kwork, 1989) (Boughanem, 1992) (Mothe, 1994) (Laskri et al., 2002). Les réseaux de neurones supportent de nombreux modèles dont l'objectif est d'imiter les fonctions de représentation et traitement de l'information du système nerveux humain. Un réseau de neurones est composé de nœuds et de liens. A chaque nœud sont associées des entrées et des sorties pondérées. A chaque lien est associé un poids traduisant le degré d'interconnexion des nœuds qu'il relie. Le

fonctionnement du réseau est basé sur la propagation des signaux d'activation depuis les entrées jusqu'aux sorties.

Le fonctionnement du réseau se fait par propagation de signaux de la couche d'entrée vers la couche de sortie. Chaque neurone de la couche d'entrée calcule une valeur de sortie et la transmet aux neurones de la couche suivante. Chaque neurone intermédiaire calcule à son tour une valeur d'entrée, une valeur de sortie et la transmet à la couche suivante… Ce processus se reproduit jusqu'à la couche de sortie. Les valeurs dans la couche de sortie servent de critères de décision (pertinence de documents, expansion de requêtes) (Boughanem et al., 1992) (Mothe, 1994) (Crestani, 1995).

L'une des propriétés fondamentales d'un réseau de neurones est la dynamique de ses états. Celle-ci traduit l'apprentissage du réseau par changement de son comportement grâce à l'évolution des poids de ses connexions en cours du temps.

Les systèmes de RI basés sur l'approche connexionniste utilisent les fondements des réseaux de neurones tant pour la modélisation des unités textuelles que pour la mise en œuvre du processus de recherche d'informations. Le modèle offre en effet des atouts intéressants pour la représentation des relations entre termes (synonymie, voisinage...) entre documents (similitude, référence...) et entre termes et documents (fréquence, poids...). En outre, sa propriété intrinsèque d'apprentissage permet de supporter de manière inhérente à son fonctionnement, le processus de reformulation de requête et/ou réinjection de pertinence utilisateur.

Il n'existe pas une représentation unique d'un réseau de neurones pour la RI. Cependant, l'architecture la plus répandue est celle fondée sur l'interconnexion de couches représentant les éléments d'un système de recherche d'informations (Boughanem et al., 2004).

❖ Le modèle RI basé-concepts

Un Système de Recherche d'Information basé-concepts se caractérise par la notion d'espace conceptuel dans lequel les documents et les requêtes sont représentés, par opposition à l'espace mots simples qu'on trouve dans les modèles classiques (Baeza-Yates et al., 1999).

Depuis la fin des années 1990, les ontologies offrent cet espace conceptuel sur lequel ces systèmes s'appuient. Ceci permet de saisir une partie de la sémantique présente dans les documents et les requêtes. Cette sémantique vient de l'utilisation des représentants des concepts (termes) de l'ontologie comme vocabulaire de référence qui englobe aussi bien le vocabulaire de l'utilisateur que celui de l'auteur du document. Ceci permet, à l'utilisateur qui exprime un besoin en information et à l'auteur du document, de "parler le même langage".

Les travaux de (Vallet et al., 2005) se basent sur une recherche d'information basé-concepts en utilisant une indexation appuyée sur les techniques d'annotation pondérée. Leur approche peut être considérée comme une évolution du modèle vectoriel classique dans la mesure où les indices basés sur des mots clefs sont remplacés par une base de connaissance fondée sur une ontologie. L'annotation semi-automatique des documents et la procédure de pondération sont équivalentes au processus d'indexation et d'extraction des mots clefs du document.

Dans (Baziz, 2005), le travail s'est focalisé sur l'utilisation (restreinte/partielle ou avancée/étendue) des ontologies pour une représentation conceptuelle de l'information en RI.

Dans le premier cas, les ontologies sont utilisées en amont d'un moteur de recherche et servent de ressource sémantique externe pour améliorer la formulation du besoin en information avant de le soumettre au SRI. Cette méthode peut s'avérer efficace, notamment lorsqu'il s'agit d'information traitant d'un domaine spécifique (médical par exemple), dans la mesure où elle permet à l'utilisateur d'exprimer son besoin dans le langage de l'ontologie. Un module de modification de la requête se charge de la reformulation ou de l'expansion de celle-ci avec des termes liés sémantiquement aux concepts de l'ontologie, puis de la retranscrire dans le langage d'indexation du SRI.

Dans le deuxième cas, l'ontologie peut être utilisée de façon plus poussée. Elle sert dans ce cas, d'espace de représentation conceptuelle dans lequel les documents et les requêtes sont exprimés par rapport à un référentiel commun : l'information est représentée non pas par rapport aux mots qu'ils contiennent mais par rapport aux concepts de l'ontologie auxquels ils renvoient. Dans la même perspective, Safran (Safran, 2005) propose une approche orientée utilisateur pour améliorer le processus de transfert des connaissances. Cette approche permet d'assister l'utilisateur dans le processus de conceptualisation en lui fournissant des requêtes personnalisées et contextualisées lors de la recherche. Ces requêtes sont établies à un niveau basé-concepts afin de satisfaire le besoin en information de l'utilisateur concernant les concepts de la base de connaissances.

La section 2.4 décrit une liste non exhaustive des modèles de recherche d'information. A titre d'exemple, nous pouvons citer les modèles cognitifs (Tricot, 2006), le modèle basé sur les réseaux possibilistes (Brini, 2005)…

Pour les modèles de recherche d'information, nous ne pouvons pas dire qu'un modèle est meilleur qu'un autre. Le choix du modèle à utiliser pour un système de recherche d'information dépend des objectifs et des paramètres utilisés lors de sa conception.

Le tableau qui suit représente un exemple de comparaison entre deux modèles de base de la RI et met en relief le fait qu'un avantage d'un modèle peut être un inconvénient de l'autre et vice versa.

Modèles de recherche	**Avantages**	**Inconvénients**
Modèle vectoriel	• Le langage de requête est plus simple (liste de mots clés). • Les performances sont meilleures grâce à la pondération des termes. • Les documents restitués sont triés et classés par pertinence.	• Le modèle considère que tous les termes sont indépendants (inconvénient théorique) • Le langage de requête est moins expressif

Modèle booléen	• Le modèle est transparent et simple à comprendre pour l'utilisateur. • Raison de sélection d'un document claire : il répond à une formule logique. • Adapté pour les spécialistes et les vocabulaires contraints	• Il est difficile d'exprimer des requêtes longues sous forme booléenne. • Le critère binaire peu efficace : il est admis que la pondération des termes améliore les résultats. • Le résultat est binaire (les documents contiennent les termes demandés ou ne les contiennent pas). Pas de classement.

2.5 EVALUATION DES SYSTEMES DE RECHERCHE D'INFORMATION

L'évaluation consiste à mesurer la différence entre un résultat obtenu et un résultat attendu.

La performance des systèmes de recherche d'information peut être évaluée à partir de la pertinence des documents renvoyés. En effet, les SRI ont pour but de retrouver les documents pertinents et d'éliminer ceux non pertinents.

La notion de pertinence peut être définie selon deux points de vue : pertinence objective et pertinence subjective. Dans le premier cas, la pertinence est mesurée par rapport au résultat de la recherche alors que dans le second cas, un document peut être jugé pertinent pour une requête à un instant *t* et pour un utilisateur donné.

Ayant une base documentaire et suite à une requête posée par l'utilisateur, nous pouvons classer, d'une façon générale, le résultat de la recherche d'information comme suit :

	Documents pertinents	**Documents non pertinents**	**Total**
Documents sélectionnés	Documents trouvés en contexte (a)	Documents trouvés hors contexte : *bruit* (b)	a+b
Documents non sélectionnés	Documents oubliés : *Silence* (c)	Documents non pertinents non trouvés (d)	c+d
	Total documents pertinents dans la base	Total documents non pertinents dans la base	a+b+c+d=N

	documentaire a+c	documentaire b+d	

Deux métriques prédominent dans la littérature pour évaluer les SRI :

- ✓ la précision du résultat correspond au pourcentage de documents pertinents, trouvés en contexte par rapport aux documents sélectionnés par le système.

$$\text{précision} = \frac{\text{Documents trouvés en contexte}}{\text{Documents sélectionnés}} = \frac{a}{a+b}$$

La précision est la capacité d'un système à ne sélectionner **que** des documents pertinents.

D'un point de vue opposé, nous pouvons définir la notion du « bruit » qui représente le pourcentage de termes non pertinents extraits par le système (faux positif):

$$\text{bruit} = 1\text{-précision}$$

- ✓ le rappel désigne le pourcentage de documents pertinents renvoyés par le système par rapport au nombre total de documents pertinents qui se trouvent dans la base documentaire.

$$\text{rappel} = \frac{\text{Documents trouvés en contexte}}{\text{Total documents pertinents}} = \frac{a}{a+c}$$

Le rappel est la capacité du système à sélectionner **tous** les documents pertinents de la collection.

D'un point de vue opposé, nous pouvons définir la notion du « silence » qui représente le pourcentage de termes pertinents n'ayant pas été extraits (faux négatifs).

$$\text{silence} = 1\text{- rappel}$$

Un système de recherche d'information est jugé performant s'il réussit à trouver l'équilibre, le juste milieu, entre ces deux critères de mesure (précision-rappel).

Autres mesures (Nakache et al., 2005) peuvent être calculées à partir du tableau de contingence présenté ci-dessus tels que:

$$\text{pertinence} = \frac{a+d}{N} \quad ; \qquad \text{erreur} = \frac{b+c}{N} ; \qquad \text{spécificité} = \frac{d}{b+d} \quad \ldots.$$

❖ La F-mesure

À partir de ces mesures, plusieurs indicateurs de synthèse ont été créés, le plus célèbre est la F-mesure qui est la moyenne pondérée de la précision et du rappel (Van Rijsbergen, 1979).

$$\mathrm{F-mesure} = \frac{2 * \text{précision} * \text{rappel}}{\text{précision} + \text{rappel}}$$

Ceci est connu comme mesure F_1, car précision et rappel sont pondérés de façon égale. Il s'agit d'un cas particulier de la mesure générale F_β (pour des valeurs réelles positives de β):

$$\mathrm{F_\beta - mesure} = \frac{(1+\beta^2) * \text{précision} * \text{rappel}}{\beta^2 * \text{précision} + \text{rappel}}$$

Si par exemple β=2, cela veut dire que la F-mesure donne un poids deux fois plus important au rappel qu'à la précision. Ceci correspond au fait que l'utilisateur tolère la présence de quelques termes inappropriés tant qu'il dispose de plusieurs autres termes pertinents (Gay et al., 2005).

❖ La précision et le rappel à n documents

Pour étudier la qualité de l'ordonnancement du résultat de la recherche d'information, il est intéressant de calculer précision P_n ou le rappel R_n du sous-ensemble des documents des n premiers. Ces deux mesures reflètent, ainsi, la similarité de chaque document avec la requête. Elles se notent respectivement P@n et R@n.

Ainsi, il est utile d'examiner la précision à 10 documents restitués si l'on s'intéresse à la capacité du système de restituer des documents pertinents en tête de liste. La précision à 5, 10, 15, 20, 30, … documents restitués présente néanmoins des limites : par exemple, si pour une requête donnée, nous avons seulement 8 documents pertinents, et que le SRI restitue bien ces 8 documents en tête de liste, le SRI aura une précision à 10 documents restitués égale à 0,8, ce qui n'illustre pas que tous les documents pertinents disponibles ont été trouvés. De plus, dans cet exemple, une précision à 10 documents égale à 0,8 ne permet pas de déterminer où se situent les deux documents non pertinents parmi les dix restitués. Pour pallier ce défaut, nous pouvons avoir recours à la R-précision

❖ La précision exacte ou R-précision

La précision exacte représente celle obtenue à l'endroit où elle vaut le rappel. Si la requête admet n documents pertinents, la R-précision est celle calculée pour les n premiers documents de la liste ordonnée des documents restituée, où n est égal au nombre total de documents pertinents de la requête (Boughanem, 2008). Cette mesure est plus réaliste pour l'étude de l'ordonnancement en tête de liste. Cependant, pour l'avoir, il est nécessaire de connaître au préalable le nombre de documents pertinents disponibles dans le corpus pour une requête donnée.

❖ La précision moyenne

La précision moyenne est une mesure de performance globale, c'est la moyenne des valeurs de précision à chaque document pertinent de la liste ordonnée.

La précision moyenne MAP (Median Average Precision) se calcule comme suit :

$$MAP = \frac{1}{n}\sum_{i=1}^{N} P@i * R(i)$$

Avec : R(i) = 1 si le i^{ème} document restitué est pertinent ;
R(i) = 0 si le i^{ème} document restitué est non pertinent ;
P@i : la précision à i documents restitués ;
n ; le nombre de documents pertinents restitués ;
N ; le nombre total de documents retournés.

❖ La courbe de rappel-précision

La précision mesurée indépendamment du rappel et inversement est peu significative. Pour pouvoir examiner les résultats efficacement, nous calculons la paire précision-rappel à chaque document restitué.

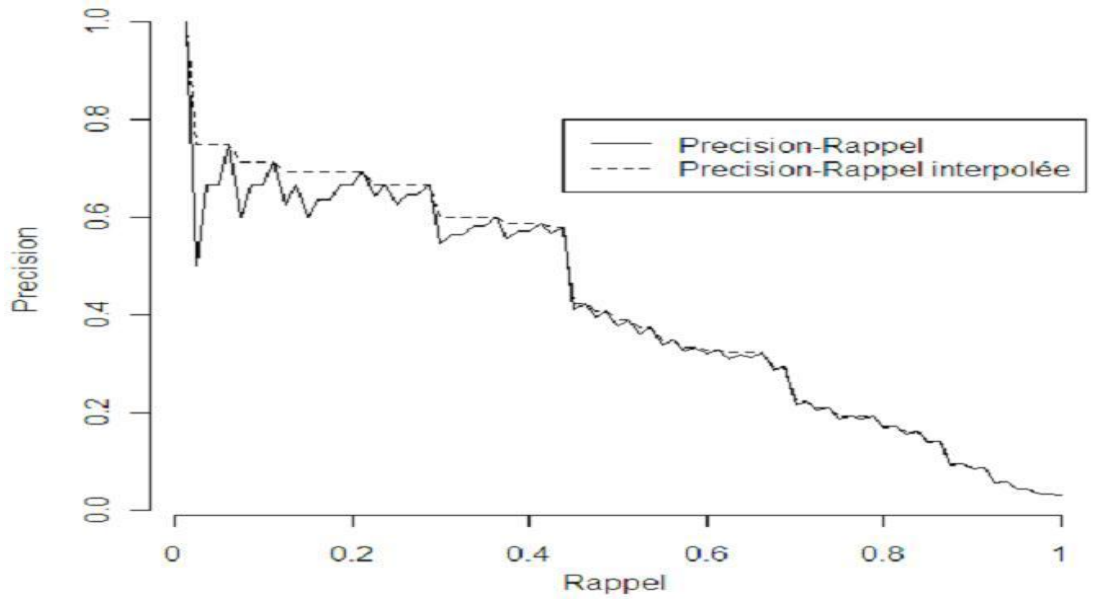


Figure 2.5. courbe précision-rappel pour la requête 157 du corpus Cranfield[28] avec la méthode SimRank

Nous observons généralement que la croissance du rappel entraîne la diminution de la précision.

❖ Les campagnes et collections de test

Un des objectifs des campagnes d'évaluation est d'évaluer et de mesurer l'efficacité des systèmes de recherche d'information, développer la communication entre l'industrie, l'université et l'état en mettant en place un forum ouvert pour faciliter les échanges d'idées sur la recherche…

[28] Cranfield University. URL: https://dspace.lib.cranfield.ac.uk/community-list

Parmi les projets les plus ambitieux pour cet effet, nous pouvons citer les campagnes d'évaluation de CLEF (Cross Language Evaluation Forum)[29] qui ont pour objectif de promouvoir la recherche et le développement dans le domaine de la recherche d'information multilingue, d'une part en offrant une infrastructure pour tester et évaluer les systèmes de recherche d'information sur des supports écrits dans les différentes langues européennes, en mode monolingue, multilingue ou inter langue, et d'autre part en mettant au point des séries de tests composés de données qui peuvent être réutilisées par les développeurs de systèmes, pour l'évaluation.

Dans le cadre de ces campagnes CLEF et depuis 2004, ImageCLEFmed (une tâche de recherche médicale de CLEF) (Müller et al., 2007) a vu le jour permettant l'évaluation de la performance des systèmes de recherche d'information médicale, fondés sur des collections d'images décrites en mono ou multilingues. La collection CLEFMedical (Müller et al., 2007), composée de comptes-rendus médicaux multilingues associés à des images. Ces comptes-rendus peuvent être rédigés en anglais, en français ou en allemand.

Le corpus utilisé en 2005 et 2006 comporte 50.412 documents, et celui utilisé en 2007 comporte 55.485 documents. Sur ces trois années, 85 requêtes, avec jugements de pertinence faits au niveau des images, sont disponibles (chaque année comporte respectivement 25, 30 et 30 requêtes).

La campagne TREC (Text REtrieval Conference)[30] est une série d'évaluations annuelles des méthodes et des outils pour la recherche d'information qui propose un cadre expérimental pour évaluer différentes applications. Pour chaque session de TREC, un ensemble de documents et de requêtes sont proposés aux participants. Ces derniers exploitent leurs propres systèmes sur ces données. Ensuite, ils envoient au NIST[31] une liste ordonnée de documents afin d'être évaluée. À la fin, les participants disposent de la liste des documents pertinents pour chaque requête ce qui leur permet d'évaluer la performance de leurs systèmes de recherche d'information.

Cette liste des mesures d'évaluation ne constitue pas une liste exhaustive. En effet, nous pouvons citer la courbe rappel-précision restreinte à un ensemble de requêtes. Par ailleurs, les campagnes d'évaluation ne cessent de voir le jour pour juger de l'efficacité des systèmes et ainsi faire évoluer leur performance, technologiquement mais également par rapport aux attentes des utilisateurs.

CONCLUSION

Nous avons présenté dans ce chapitre les principales notions et concepts de la recherche d'information ainsi que les principales étapes d'un processus de recherche d'information. Nous avons rappelé ce qu'est l'indexation dans ces systèmes : une projection des documents et des requêtes dans un espace de représentation. Par la suite nous avons mis en relief les

[29] URL : http://www.clef-campaign.org/
[30] URL: http://trec.nist.gov/
[31] National Institute of Standards and Technology

principaux modèles de la recherche d'information existants dans la littérature. Et pour finir, nous avons décrit les différentes mesures d'évaluation des systèmes de RI.

Dans le cinquième chapitre, nous présentons le modèle que nous avons utilisé pour la recherche d'information dans le catalogue CISMeF. Notre modèle s'inspire largement du modèle basé–concepts utilisant comme espace conceptuel les principales terminologies médicales disponibles en français.

CHAPITRE 3

LES TERMINOLOGIES MEDICALES ET LA MISE EN PLACE DE L'UNIVERS MULTI-TERMINOLOGIQUE

INTRODUCTION

Dans ce chapitre, nous définissons le vocabulaire utilisé en tant que terminologies médicales en se focalisant sur celles qui ont été les plus utilisées dans notre travail. Dans la deuxième partie du chapitre, nous mettons en relief le passage vers une structure fondée sur plusieurs terminologies en mettant en avant le processus d'intégration, dans la structure de base de CISMeF, de toutes ces terminologies selon un modèle générique.

3.1 ONTOLOGIES, CLASSIFICATIONS, THESAURUS, TERMINOLOGIES, DICTIONNAIRE, NOMENCLATURE

3.1.1 DEFINITIONS

Les langages documentaires permettent de mettre au point l'organisation des connaissances et de faciliter l'accès à l'information. Leur nécessité dépend de la croissance des volumes d'information disponible et l'apparition de nouvelles modalités de communication de

l'information. Bon nombre d'ouvrages consacrés aux langages documentaires qui font autorité à l'heure actuelle ont été publiés quasi simultanément à la fin des années quatre-vingt. Parmi lesquels, nous citons (Chaumier, 1988a), (Maniez, 1987), (Van Slype, 1987)…

Vocabulaire contrôlé : Un vocabulaire contrôlé est une liste établie de termes normalisés (vocabulaire qui n'a pas l'ambiguïté du langage naturel) à utiliser dans l'indexation et la recherche documentaire. Un vocabulaire contrôlé assure qu'un sujet sera décrit en utilisant le même terme préférentiel chaque fois qu'il est indexé, facilitant la recherche d'information sur un sujet spécifique. Lorsque cette liste de vocabulaire est organisée et régie par des relations sémantiques, nous parlons de thésaurus.

Thésaurus : D'après (Rector, 1998) un thésaurus est un langage documentaire fondé sur une structuration hiérarchisée, alphabétique au premier niveau puis thématique. Les termes normalisés étant reliés à des termes plus précis.

Une autre définition a été relatée par Chaumier (Chaumier, 1988b) pour définir le thésaurus comme étant un langage documentaire fondé sur une structuration hiérarchisée des termes. Ils y sont organisés de manière conceptuelle et reliés entre eux par des relations sémantiques. Du point de vue de sa structure, le thésaurus est un vocabulaire, contrôlé et dynamique, de termes ayant entre eux des relations sémantiques et génériques et qui s'applique à un domaine particulier de la connaissance.

Certains thésaurus (par exemple le thésaurus MeSH) sont utilisés dans des systèmes de recherche d'information, ils permettent d'associer les mots de la requête de l'utilisateur avec des termes connus pour avoir des sens voisins. Ils permettent, du fait de leur organisation hiérarchique, de spécialiser les requêtes et de proposer des structures support pour la navigation dans la base textuelle.

De nombreux thésaurus ont été construits par introspection et consultation d'experts du domaine, soit sans aucune référence aux applications dans lesquelles le thésaurus sera utilisé. En fait, historiquement comme l'écrit Lerat (Lerat, 1995), « un thésaurus est surtout un outil linguistique pour l'indexation des documents dans lequel on peut puiser des mots pour construire un index par exemple ». Il n'a pas vocation à représenter les connaissances terminologiques du domaine telles qu'elles sont exprimées dans les textes.

De fait, le thésaurus ne propose pas une vue d'ensemble du domaine que délimitent les textes d'un système d'information.

Dictionnaire : D'après le dictionnaire Larousse, un dictionnaire est un « *Ouvrage didactique constitué par un ensemble d'articles dont l'entrée constitue un mot, indépendants les uns des autres et rangés dans un ordre déterminé, le plus souvent alphabétique* ». Un Dictionnaire médical contient des définitions, des termes médicaux et scientifiques, abréviations, maladies… relatifs au domaine médical.

Classification : D'après le dictionnaire Larousse, la classification est l'« *Action de distribuer par classes, par catégories* ».

(Rector, 1998) suggère qu'une classification répartit systématiquement en classes, des termes désignant des êtres, choses ou notions ayant des caractères communs afin d'en faciliter l'étude.

Quant à (Runciman et al., 2009), il définit une classification comme étant un arrangement des concepts (ayant ou exprimant un sens ou une signification) dans des classes (des groupes ou des ensembles de choses similaires) et de leurs subdivisions liées pour exprimer les rapports sémantiques entre eux (la manière dont ils sont associés les uns avec les autres selon leurs significations). Chaque classe organise hiérarchiquement des subdivisions composées de concepts. Les concepts peuvent être représentés par un certain nombre de termes qui tiennent compte des dialectes régionaux, de différentes langues ou de différentes disciplines.

Nomenclature : D'après le dictionnaire Larousse, une nomenclature est l'« *Ensemble des mots en usage dans une science, un art, ou relatifs à un sujet donné, présentés selon une classification méthodique* » ou encore « *une liste, catalogue détaillé et ordonné des éléments d'un ensemble, permettant de classer celui-ci : La nomenclature des monuments français* ».

Dans (ISO, 2000) la nomenclature est définie comme « *un ensemble de termes techniques, présentés selon un classement méthodique* ».

Ainsi, une nomenclature peut être définie comme étant un système de mots (ou de concepts) utilisés dans une discipline particulière, comme dans la médecine et la chirurgie, l'anatomie et la biochimie, etc. Un système standard de nomenclature présuppose l'existence d'une classification organisée des entités reliées à ce domaine.

Terminologie : Une terminologie est une liste des termes techniques ou des expressions utilisées dans un domaine spécifique. Une définition plus précise a été avancée par (Lefevre, 2000) présentant les terminologies comme « des listes de termes d'un domaine ou d'un sujet donné représentant les concepts ou notions les plus fréquemment utilisés ou les plus caractéristiques ».

Taxonomie : D'après le dictionnaire Larousse, une taxonomie est « *une Classification, une suite d'éléments formant des listes qui concernent un domaine ou une science* ».

Dans la littérature, ils existent plusieurs autres définitions, par exemple la société Lingway[32] donne la définition suivante « *réseau sémantique dans lequel la seule relation est la seule hiérarchique* ». L'institut Montague[33] définit une taxonomie comme « *Un système pour nommer et organiser des objets en groupes qui partagent des caractéristiques similaires*». Quant au glossaire Dublin Core[34], il propose « *Classification systématique selon des principes ou lois généraux* ». Ce même glossaire affirme de plus que « *Un système de classification comme la Classification de la Bibliothèque du Congrès est un exemple de taxonomie*».

Nous sommes donc en présence de conceptions très hétérogènes : nous passons d'une définition restreinte aux systèmes classificatoires exclusivement hiérarchiques, à tout système

[32] Lingway vertical semantic solutions. URL : http://www.lingway.com/
[33] Montahue Institute Review. URL : http://www.montague.com/review/myths.html
[34] Dublin Core Metadata Initiative. URL : http://dublincore.org/documents/usageguide/glossary.shtml

de classification, qu'il soit hiérarchique ou non. La portée du terme est finalement étendue à tout langage documentaire doté, exclusivement ou non, d'une organisation hiérarchique.

Ontologie : Le terme « Ontologie » est issu du domaine de la philosophie, où il signifie «e*xplication systématique de l'existence* ». Dans le cadre de l'intelligence artificielle, Neches et ses collègues (Neches et al., 1991) étaient les premiers à en proposer une définition, à savoir : « *une ontologie définit les termes et les relations de base du vocabulaire d'un domaine ainsi que les règles qui indiquent comment combiner les termes et les relations de façon à pouvoir étendre le vocabulaire* ». Cette définition explicite comment élaborer une ontologie : repérer les termes de base et les relations entre les termes, identifier les règles servant à les combiner, fournir des définitions de ces termes et de ces relations. D'après cette définition, une ontologie inclut non seulement les termes qui y sont explicitement définis, mais aussi les termes qui peuvent être créés par déduction en utilisant les règles. En 1993, Gruber (Gruber, 1993) formule la définition suivante : « *une ontologie est une spécification explicite d'une conceptualisation* », qui deviendra célèbre et restera la définition la plus citée dans la littérature scientifique. En 1997, Borst (Borst, 1997) apporte une légère modification à la définition de Gruber en précisant que « *les ontologies se définissent comme une spécification formelle d'une conceptualisation commune* ». Studer et ses collègues (Studer et al., 1998) ont donné l'interprétation suivante de ces deux définitions : « *la conceptualisation renvoie à un modèle abstrait d'un quelconque phénomène après en avoir relevé les concepts significatifs* ». Ainsi, nous passons à une définition plus normative et abstraite, du fait qu'un terme est une désignation d'un concept. Par *explicite*, il faut entendre que le type de concepts utilisés, ainsi que leurs contraintes d'utilisation, sont définis de façon explicite. L'adjectif *formel* exprime le fait que l'ontologie doit être lisible par ordinateur. *Commun* renvoie à l'idée qu'une ontologie rend compte d'un savoir consensuel, c'est à- dire qu'elle n'est pas l'objet d'un individu, mais qu'elle est reconnue par un groupe.

Comme une petite synthèse de ces définitions, nous faisons référence aux travaux de (Lassila et al., 2001) qui ont situé ces différentes structures de données (terminologie, thésaurus, ontologie…) dans un continuum dont la dimension principale est le degré de formalisation (cf. Figure 3.1.1).

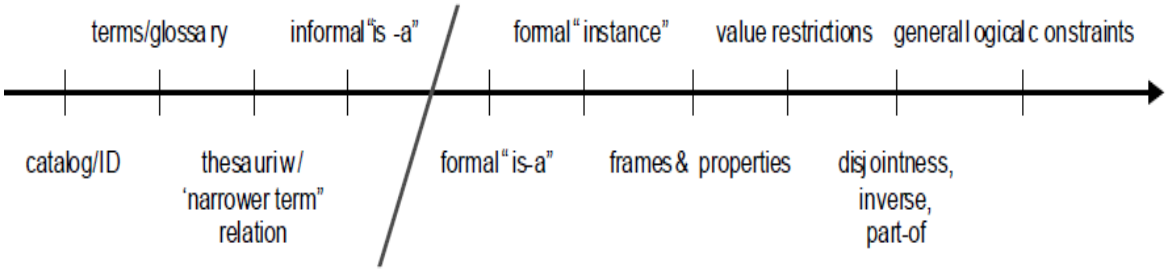


Figure 3.1.1. Différentes ressources terminologique et ontologie selon leur degré de formalisation

De plus en plus de travaux s'intéressent à la formalisation des terminologies et à la construction des ontologies. (Despres et al., 2008) proposent une méthode de construction d'ontologies mettant en relief la phase de conceptualisation : du plan linguistique, au plan

termino-ontologique pour avoir finalement l'ontologie. L'étude linguistique permet d'extraire les termes représentatifs du domaine. La construction du réseau termino-ontologique est faite à partir de l'interprétation des unités linguistiques (termes et relations lexicales les liant) constituant le réseau terminologique (concepts terminologiques et relations sémantiques les liant).

Enfin les concepts de l'ontologie et les relations conceptuelles les associant sont construits à partir des unités termino-ontologiques figurant dans les réseaux termino-ontologiques. Ces concepts ontologiques sont décrits dans un langage formel, organisés dans une structure hiérarchique, liés par des relations conceptuelles et contraints par des règles et des axiomes.

3.1.2 TERMINOLOGIES MEDICALES

Dans le cadre de cette thèse, nous avons utilisé la principale classification des médicaments utilisée en France et en Europe à savoir l'ATC, les différents codes des médicaments et des substances chimiques (CAS, CIP, CIS et UCD) ainsi que le thésaurus MeSH.

Dans la cadre du projet PSIP, six terminologies ont été utilisées : l'ATC, la CIM-10, l'IUPAC, la WHO-ICPS, la NCCMERP et la taxonomie du PSIP.

3.1.2.1 LA CLASSIFICATION ANATOMIQUE THERAPEUTIQUE ET CHIMIQUE

En 1969, le groupe de recherche pour l'utilisation des médicaments (Drug Utilisation Research Group : DURG) a été constitué d'un groupe d'experts conseillers de l'OMS, suite à un besoin ressenti pour un système de classification internationalement reconnu et qui pourrait être employé pour des études concernant des médicaments. En mettant à jour le système de classification European Pharmaceutical Market Research Association (EPhMRA) et en collaborant avec le Dépôt Médicinal Norvégien (NMD), les chercheurs norvégiens ont développé un système baptisé la classification ATC (Anatomique, Thérapeutique et Chimique). Une unité technique de mesure appelée la Defined Daily Dose (DDD) a été également développée. Le DDD est défini comme la « dose moyenne journalière pour un médicament, utilisée pour son indication principale pour les adultes». Le Conseil nordique sur les médicaments (Nordic Council on Medicines) en collaboration avec le NMD, a développé le système ATC/DDD. La méthodologie ATC/DDD a été employée pour la première fois en 1976 dans une publication «Nordic Statistics on Medicines» (WHO Collaborating Centre for Drug Statistics Methodology, 2009).

En 1981, le bureau régional Européen de l'OMS (Organisation Mondiale de la Santé) a recommandé le système ATC/DDD pour des études internationales concernant l'utilisation des médicaments. En 1982, un corps responsable de coordination de l'utilisation de la méthodologie, le Collaborating Centre for Drug Statistics Methodology de l'OMS, a été établi à Oslo, Norvège. Le centre est maintenant situé à l'institut norvégien de la santé publique et est financé par le gouvernement norvégien.

Depuis 1996, les sièges sociaux de l'OMS recommandent le système ATC pour des études portées sur l'utilisation des médicaments. L'OMS recommande le système ATC, qui est

également employé pour la détection des effets indésirables des médicaments, pour des comparaisons internationales. Depuis peu, on a décidé que la liste principale des médicaments de l'OMS devrait également être basée sur le système de classification ATC pour renforcer une utilisation plus répandue du système. Les centres de collaboration de l'OMS impliqués dans la surveillance des médicaments utilisent le système ATC pour la classification et les statistiques. Le centre de collaboration de l'OMS pour la surveillance internationale des médicaments (centre de surveillance d'Uppsala) en Suède maintient le dictionnaire des médicaments de l'OMS, une base de données pour la plupart des médicaments utilisés dans les pays participant au programme de l'OMS pour la surveillance internationale des médicaments (Family Medicine Research Center, 2010).

La structure de la classification

La classification ATC (Skrbo et al., 2004) classe des substances chimiques par catégorie à cinq niveaux différents selon l'organe ou le système sur lesquels elles agissent et selon leurs propriétés chimiques, pharmacologiques et thérapeutiques.

Le code ATC a la forme générale suivante : LCCLLCC (où L : lettre ; C : chiffre).
Dans ce système, les médicaments sont classés en groupes à cinq niveaux différents :

- 1er niveau : classe anatomique principale (1 caractère alphabétique).
- 2ème niveau : sous-classe thérapeutique (2 chiffres).
- 3ème niveau : sous-classe pharmacologique (1 caractère alphabétique).
- 4ème niveau : sous-classe chimique (1 caractère alphabétique).
- 5ème niveau : substance active (2 chiffres).

À chaque niveau de la classification correspond un code et un libellé ATC. Le libellé du cinquième niveau correspond à la DCI (Dénominations Communes Internationales)[35] de la substance, quand elle existe.

Le tableau ci-dessous illustre les 14 groupes principaux du premier niveau (Groupe anatomique) de la classification ATC.

A	Voies digestives et métabolisme
B	Sang et organes hématopoïétiques
C	Système cardiovasculaire
D	Médicaments dermatologiques
G	Système génito-urinaire et hormones sexuelles
H	Hormones systémiques, hormones sexuelles exclues
J	Antiinfectieux généraux à usage systémique
L	Antinéoplasiques et immunomodulateurs
M	Muscle et squelette
N	Système nerveux

[35] « Les DCI permettent d'identifier les substances pharmaceutiques ou leurs principes actifs ». Directives générales pour la formation de dénominations communes internationales applicables aux substances pharmaceutiques. URL : http://www.who.int/medicines/services/inn/GeneralprinciplesFr.pdf

P	Insecticides antiparasitaires
R	Système respiratoire
S	Organes sensoriels
V	Divers

Exemple de la hiérarchie de la substance « *Metformine* » :

Niveau	Code	libellé	Groupe
1	A	Voies digestives et métabolisme	Groupe anatomique principal
2	A10	Médicaments du diabète	Sous-groupe thérapeutique
3	A10B	Antidiabétiques sauf insulines	Sous-groupe pharmacologique
4	A10BA	Biguanides	Sous-groupe chimique
5	A10BA02	Metformine	Substance chimique

Le code ATC est attribué en fonction de son indication principale. Or, cette dernière peut varier d'un pays à l'autre, ce qui explique qu'il peut exister plusieurs codes ATC pour un même médicament en fonction du pays concerné. C'est le cas pour environ 10% des médicaments qui n'ont pas le même code ATC entre la France et le Danemark[36].

Ainsi, pour une même substance chimique, nous pouvons avoir plusieurs codes ATC différents, selon son effet thérapeutique, son effet pharmacologique ou encore son appartenance anatomique. Par exemple, la substance chimique « acide acétylsalicylique » est tantôt classée sous le groupe A « Voies digestives et métabolisme» ayant le code A01AD05 lorsqu'elle a un effet antalgique ou anti-inflammatoire et, tantôt sous le groupe B « Sang et organes hématopoïétiques » ayant le code B01AC06 lorsqu'elle a un effet antiagrégant plaquettaire.

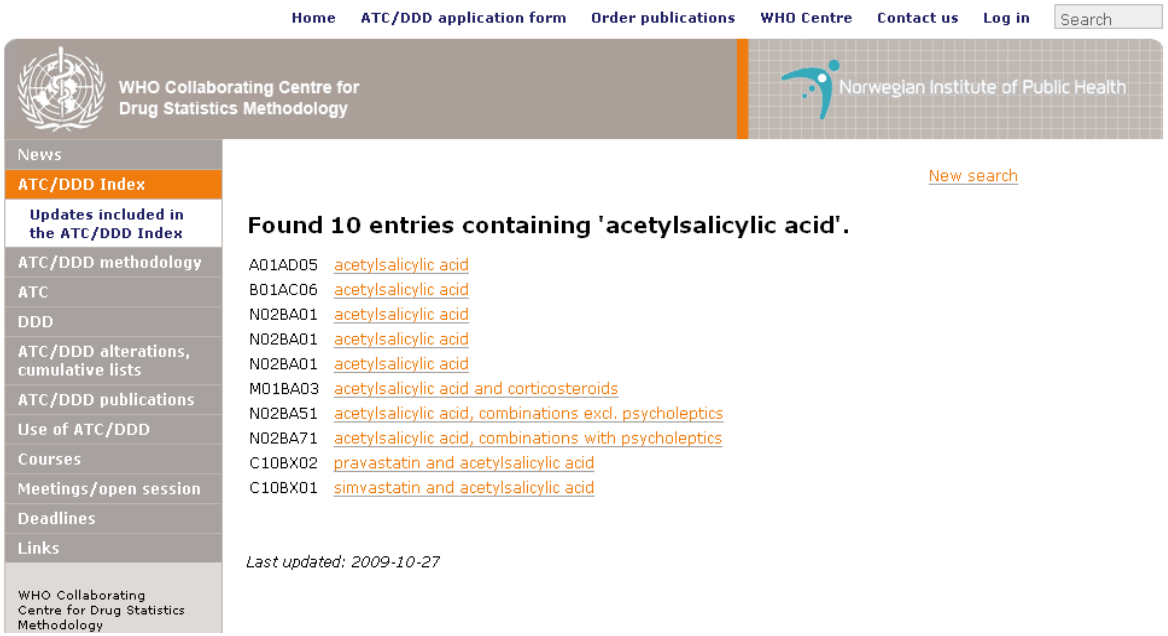


Figure 3.1.2.1. Les différents codes ATC pour la substance « acide acétylsalicylique » et ses dérivés

[36] Étude interne réalisée par la société Vidal dans le cadre du projet PSIP

La classification ATC est la plus utilisée en France et en Europe pour classer les médicaments. À ce titre, elle a été choisie dans le projet PSIP pour cet objectif[37]. Il faut signaler que la classification ATC est pratiquement inconnue aux États-Unis où RxNorm[38] est utilisée. Pour illustrer ce propos, remarquons que l'ATC n'est pas intégré dans l'UMLS[39].

3.1.2.2 CLASSIFICATIONS ET CODES UTILISES POUR LES MEDICAMENTS

- **La nomenclature CAS**

Le numéro CAS (Chemical Abstract Service) d'un produit chimique, polymère, séquence biologique ou d'un alliage est son numéro d'enregistrement unique auprès de la banque de données de Chemical Abstract Service (CAS), une division de l'American Chemical Society (ACS). De plus, CAS maintient et commercialise une base de données de ces substances, *CAS Registry*. Cette dernière contient plus de 55 millions de substances organiques et inorganiques et 62 millions de séquences[40]. Approximativement, 12.000 nouvelles substances sont ajoutées chaque jour[41]. Le but est de faciliter les recherches dans les bases de données, vu que les produits chimiques ont souvent différents noms. Presque toutes les bases de données de molécules actuelles permettent une recherche par numéro CAS. En effet, ce dernier est utilisé à l'échelle mondiale.

Le CAS assigne ces numéros, identifiables par un algorithme qui détermine les diagrammes structurels et alloue automatiquement un numéro C.A.S. unique, à chaque produit chimique (molécule, mélange d'isomères, produit industriel…) qui a été décrit dans la littérature. Compte tenu de la complexité de la nomenclature chimique et la possibilité de désigner une substance par plusieurs noms, le numéro CAS permet d'identifier les espèces chimiques sans aucune ambiguïté. Les numéros CAS sont attribués dans un ordre croissant et n'ont pas de signification particulière. Ce numéro se divise en trois parties, séparées par des tirets : YYYYYY-XX-X. La première partie peut contenir jusqu'à six chiffres, la deuxième contient deux chiffres, alors que la troisième contient un chiffre pour la somme de contrôle. La somme de contrôle se calcule en prenant le 1er chiffre fois 1, le 2eme fois 2, et ainsi de suite en partant de l'avant dernier (de gauche à droite). La somme de ces résultats intermédiaires est ensuite additionnée modulo 10. Par exemple, le numéro CAS de l'eau est 7732-18-5 : sa somme de contrôle vaut (8×1 + 1×2 + 2×3 + 3×4 + 7×5 + 7×6) mod 10 = 105 mod 10 = 5.

Via la NLM[42] (US National Library of Medicine), nous pouvons avoir toutes les informations concernant les médicaments et les substances chimiques. Par exemple, grâce à la page de recherche structurée *ChemIDplus Advanced* (Chemical Identification Plus Advanced), il est

[37] Dans le cadre du projet PSIP, l'équipe CISMeF a manuellement accentué la traduction française (2OOO-homme). Le travail a été délivré au centre Norvégien.
[38] National Library of Medicine. URL: http://www.nlm.nih.gov/research/umls/rxnorm/
[39] Se référer à la Section 3.2 pour plus de détails.
[40] Le Chemical Substance Index URL : http://www.cas.org/
[41] CAS: a division of the American Chemical Society. URL : http://www.cas.org/expertise/cascontent/registry/regsys.html#q1
[42] National Library of Medicine. URL : http://www.nlm.nih.gov/

possible d'accéder à plus de 260.000 substances chimiques[43]. Au sein du MeSH, le numéro CAS est indiqué comme un registry number[44].

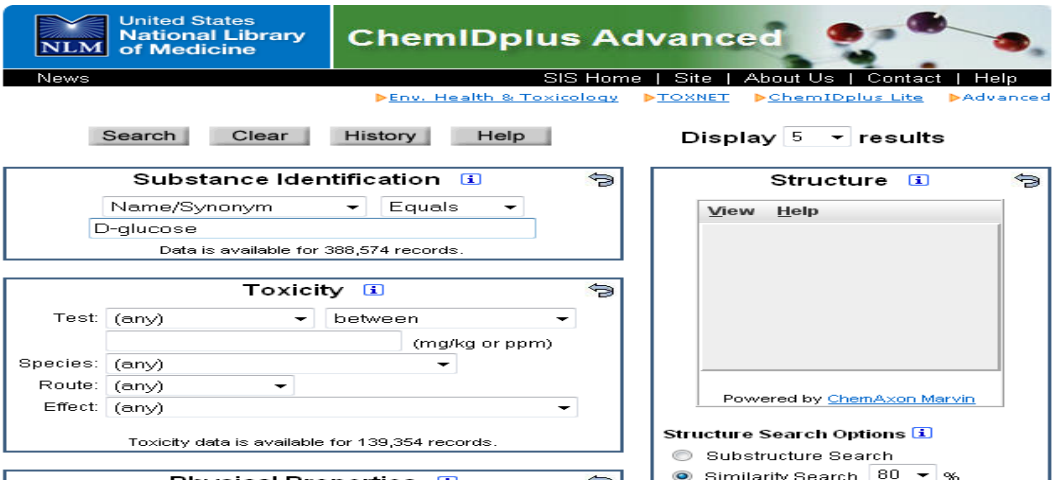


Figure 3.1.2.2. Exemple de recherche du code CAS pour la molécule D-glucose

Une particularité avec les numéros CAS est que chaque produit chimique possède un numéro CAS, permettant de l'identifier d'une manière unique au niveau de la base de données *CAS Registry* (Dittmar et al., 1976). Par exemple, les différents isomères d'une molécule ont des numéros CAS différents : le D-glucose admet comme numéro CAS, le 50-99-7, le L-glucose est identifié par 921-60-8, alors que le α-D-glucose est désigné par 26655-34-5, etc. À l'inverse parfois, une classe complète de molécules reçoit un seul numéro : le groupe des alcools déshydrogénases admet comme code CAS le 9031-72-5.

Lors de recherche par numéro CAS dans les bases de données, il est utile d'inclure le numéro de composés proches. Par exemple, pour chercher de l'information sur la cocaïne (CAS 50-36-2), il faut aussi chercher pour chlorhydrate de cocaïne (CAS 53-21-4), puisque c'est sous cette forme que la cocaïne est utilisée en tant que drogue.

- **Les codes CIS, CIP et UCD [45]**

Ces trois codes sont exclusivement utilisés en France.

Le code CIS (Code Identifiant de Spécialité) est un code numérique à 8 chiffres identifiant une spécialité pharmaceutique faisant ou ayant fait l'objet d'une Autorisation de Mise sur le Marché (AMM) en France.

Le code CIP (Code Identifiant de Présentation) est un identifiant composé de 13 chiffres (7 chiffres jusqu'à 2009[46]), correspondant à l'autorisation de mise sur le marché d'une

[43] Environmental Health & Toxicology; specialized information services. URL: http://sis.nlm.nih.gov/enviro/chemicaldruginformation.html
[44] Medical Subject Headings. URL : http://www.nlm.nih.gov/mesh/MBrowser.html
[45] Haute Autorité de Santé ; Glossaire Certification des LAP. URL : http://www.has-sante.fr/portail/jcms/c_671889/certification-des-lap?id=c_671889&#c_671927
[46] Agence française de sécurité sanitaire des produits de santé. URL : http://www.afssaps.fr/Activites/Autorisations-de-mise-sur-le-marche/Modification-des-codes-

présentation d'un médicament. La présentation d'un médicament est définie comme étant le conditionnement sous lequel une spécialité pharmaceutique est mise à disposition du public.

Une spécialité pharmaceutique peut être commercialisée sous différentes présentations : selon la taille ou la contenance du conditionnement. Se référant au tableau qui suit, nous constatons qu'un médicament peut être identifié par plusieurs numéros CIS, qui font référence à un dosage et/ou une forme galénique différents pour un médicament spécifique. Pour un même code CIS, nous pouvons avoir plusieurs codes CIP selon les différentes présentations existantes (la taille et/ou le conditionnement).

CIS	**Dénomination de la spécialité**	**Titulaire de l'AMM**	**CIP 7**	**CIP 13**	**Nom de la présentation**
61490049	HALDOL, 1 mg, comprimé	JANSSEN CILAG	3047143	3400930471432	plaquette(s) thermoformée(s) PVC-Aluminium de 40 comprimé(s)
61490049	HALDOL, 1 mg, comprimé	JANSSEN CILAG	5532977	3400955329770	plaquette(s) thermoformée(s) PVC-Aluminium de 400 comprimé(s)
62237643	HALDOL, 2 mg/ml, solution buvable en gouttes	JANSSEN CILAG	3047172	3400930471722	1 flacon(s) polyéthylène de 15 ml - avec compte-gouttes
62237643	HALDOL, 2 mg/ml, solution buvable en gouttes	JANSSEN CILAG	5533008	3400955330080	4 flacon(s) polyéthylène de 195 ml

Les deux codes (CIP, CIS) sont administrés par l'AFSSAPS (Agence Française de Sécurité Sanitaire des Produits de Santé)[47].

Pour informatiser et /ou automatiser les opérations à effectuer dans une pharmacie hospitalière, la codification des articles à gérer est indispensable, notamment pour :

identifiants-de-presentation-dans-les-AMM-de-specialites-pharmaceutiques/(offset)/3

[47] Haute Autorité de Santé. URL : http://www.has-sante.fr/portail/jcms/c_671889/certification-des-lap#c_671986

- ✓ l'approvisionnement et la gestion des stocks (médicaments, articles, accessoires et dispositifs médicaux) ;
- ✓ la dispensation des médicaments aux malades.

Parmi les codifications qui sont utilisées pour les médicaments, nous pouvons citer la série 900000 qui correspond à l'Unité Commune de Dispensation et/ou de Distribution (UCD).

En effet, à la demande des pharmaciens hospitaliers et en accord avec la DHOS (Direction de l'Hospitalisation et de l'Organisation des Soins[48]), la CNAM (Caisse Nationale d'Assurance Maladie), l'AFSSAPS et le LEEM (Les Entreprises du médicament), le CIP (Club Inter Pharmaceutique) a développé et a pris en charge de gérer une codification des Unités Communes de Dispensation (UCD).

Le code UCD caractérise la plus petite unité intègre utilisée pour la dispensation des médicaments dans les établissements de soins. Le code UCD a été retenu comme norme d'échange par le Ministère de la Santé dans le cadre de la tarification à l'activité (T2A) et de la rétrocession. L'arrêté du 2 août 2004 publié au Journal Officiel du 22 août 2004) identifie les médicaments onéreux par leur code UCD. Il s'agit de la première publication de l'UCD au Journal Officiel qui devient la référence pour les échanges économiques et la gestion interne des établissements de soins[49].

Les fichiers des médicaments codés en UCD délivrés en établissements de santé, sont désormais disponibles sur la Base des médicaments à code UCD[50]. Ils sont mis à jour en fonction des publications au Journal officiel[51].

3.1.2.3 LE THÉSAURUS MESH : MEDICAL SUBJECT HEADINGS[52]

Le thésaurus MeSH est un vocabulaire contrôlé crée par la NLM et est essentiellement utilisé pour indexer les articles scientifiques de la base de données bibliographiques MEDLINE[53]. Il est employé aussi pour cataloguer et rechercher l'information biomédicale et les documents relatifs à la santé.

De la première édition du MeSH en 1960, à la deuxième édition en 1963, plusieurs améliorations ont été faites. En effet, une arborescence des termes du MeSH est établie pour

[48] Par décret et arrêté du 15 mars 2010 publiés au journal officiel le 16 mars, la direction générale de l'offre de soins (DGOS) est créée au sein du ministère chargé de la santé, en lieu et place de la direction de l'hospitalisation et de l'organisation des soins (DHOS).

[49] Site hospitalier Club Inter Pharmaceutique. URL : http://www.ucdcip.org/#menu5

[50] Base des médicaments et informations tarifaires. URL : http://www.codage.ext.cnamts.fr/codif/bdm_it/index_tele_ucd.php?p_site=AMELI

[51] Base des médicaments à code UCD. URL: http://www.ameli.fr/professionnels-de-sante/directeurs-d-etablissements-de-sante/codage/medicaments/base-des-medicaments-a-code-ucd.php. Article mis à jour le 31 juillet 2008.

[52] Introduction to MeSH-2010. URL: http://www.nlm.nih.gov/mesh/introduction.html

[53] MEDLINE est une base de données bibliographique qui couvre tous les domaines médicaux de l'année 1966 à nos jours : plus de 11 millions de références issues de 4 300 périodiques, principalement en langue anglaise.

la première fois qui contient 13 principales hiérarchies et un total de 58 groupes répertoriés en sous-catégories et en catégories principales.

Ces listes classées par catégorie ont pour but de faciliter la tâche de l'utilisateur à trouver plus de termes connexes que dans l'ancienne structure de référence. En 1963, le MeSH disposait de 5.700 descripteurs, comparés à 4.400 dans l'édition 1960. En revanche, l'édition 2010 du MeSH contient 25.588 descripteurs.

Les principales composantes du thésaurus MeSH sont : les descripteurs, les qualificatifs, et les concepts chimiques supplémentaires (CCSs). Les types de publication, alors que techniquement considérés comme des descripteurs depuis quelques années, sont également inclus, puisqu'ils sont employés différemment dans certains cas.

1- Les descripteurs : ils sont employés pour décrire des publications et indexer des citations dans la base de données MEDLINE de NLM et dans d'autres bases de données. Les descripteurs sont généralement mis à jour dans une base annuelle mais peuvent, occasionnellement, être mis à jour plus fréquemment.

   Les descripteurs MeSH sont organisés en 16 catégories : par exemple la catégorie A pour des termes anatomiques, la catégorie B pour des organismes, la catégorie C pour les maladies, et la catégorie D pour des médicaments et des substances chimiques, etc.[54] Chaque catégorie est divisée en sous-catégories. Dans chaque sous-catégorie, des descripteurs sont rangés hiérarchiquement du plus général au plus spécifique dans jusqu'à onze niveaux hiérarchiques. Cette structure ne représente pas un système de classification bien fondé, mais plutôt une organisation utile des descripteurs pour les indexeurs des articles scientifiques ou les utilisateurs qui emploient le MeSH lors de leur recherche dans la littérature. La structure représente fréquemment un compromis entre les besoins des disciplines particulières et des utilisateurs. Chaque descripteur a un (ou des) numéro d'arborescence qui le positionne dans la hiérarchie et qui le relie à la catégorie de départ. Un descripteur peut avoir plusieurs numéros d'arborescence du fait qu'il appartienne à plusieurs catégories (cf. Figure 3.1.2.3). Ces numéros ne servent qu'à localiser les descripteurs, ils n'ont aucune signification intrinsèque. Par exemple, le fait que D12.776.641 et D12.644.641 ont, tous les deux, le groupe de trois chiffres 641 n'implique aucune caractéristique commune. Les nombres peuvent changer quand de nouveaux descripteurs sont ajoutés ou la structure hiérarchique est mise à jour pour refléter des changements du vocabulaire.

   La tâche des indexeurs est de distinguer le descripteur MeSH le plus spécifique et le plus approprié pour illustrer chaque concept représentatif de l'article.

[54] MeSH tree structures. URL: http://www.nlm.nih.gov/mesh/intro_trees.html

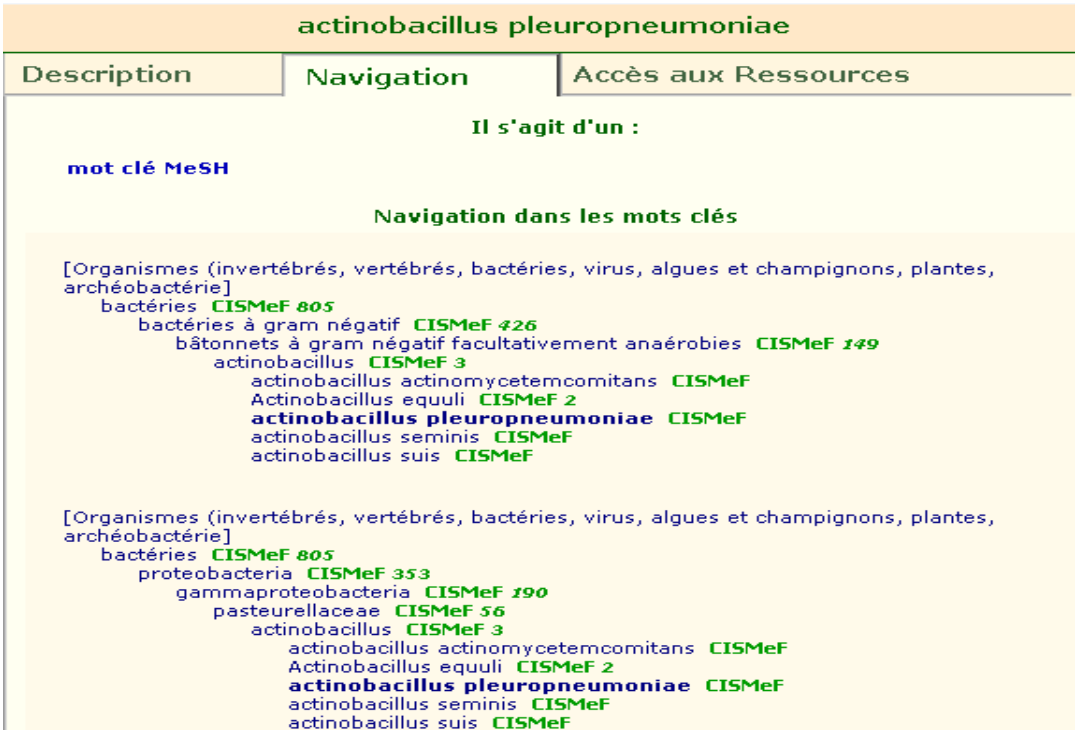


Figure 3.1.2.3. Exemple illustré par le catalogue CISMeF de deux hiérarchies différentes pour le terme « *actinobacillus pleuropneumoniae*»

Chaque descripteur MeSH est identifié par un identifiant unique et peut posséder des synonymes et être affecté à un ensemble de qualificatifs qui lui donneront un sens particulier.

2- Les qualificatifs : il y a 83 qualificatifs[55] utilisés pour indexer et cataloguer les articles, en conjonction avec les descripteurs. Les qualificatifs précisent le sens d'un descripteur et permettent de regrouper ensemble les citations qui se rapportent à un thème particulier. Par exemple, une indexation du type (descripteur/qualificatif) « *foie/action des médicaments et substances chimiques »* indique que la ressource fait référence, plus précisément, aux effets des médicaments et des substances chimiques sur le foie. Le nombre de qualificatifs est plutôt stable et rares sont les modifications les concernant.

Chaque descripteur du MeSH a une liste contextuelle de qualificatifs à affilier. Par exemple, il n'est pas possible d'affilier le qualificatif « diagnostic » au descripteur « bibliothèque médicale ».

3- Les concepts chimiques supplémentaires (CCSs) : ils sont employés pour indexer des produits chimiques, des médicaments, et d'autres concepts pour MEDLINE.

À la différence des descripteurs, les CCSs ne sont pas hiérarchisés. Cependant, chaque concept chimique supplémentaire est lié à un ou plusieurs descripteurs. Ils possèdent des relations sémantiques avec ces derniers. Pour chaque CCS, le MeSH recommande une projection vers des descripteurs, mais aussi mentionnerait le ou les descripteur(s)

[55] Dans la version 2010 du MeSH

correspondant à (aux) l'action(s) pharmacologique(s) de la substance décrite. Par exemple, suite à cette règle définie par le MeSH, une indexation avec le CCS «cétuximab », est complétée par une indexation avec le descripteur « anticorps monoclonal » et « antinéoplasiques » qui représente l'action pharmacologique correspondante.

Les CCSs sont mis à jour chaque semaine. Il y a actuellement plus de 186.000 enregistrements de CCSs avec plus de 465.000 termes de CCSs[56].

4- Les types de publications : ils sont considérés comme étant des descripteurs MeSH et ont pour but d'indiquer le type de l'article indexé, en d'autres termes, son contenant plutôt que son contenu, par exemple, « *article historique* ». Ils peuvent inclure des composantes d'une publication, tel que *graphiques* ; Formats de publication, tel que *éditorial* ; et caractéristiques d'une étude, tel que *essai clinique*. Ces données peuvent être considérées comme des métadonnées de contenant, plutôt que des informations décrivant le contenu des articles.

Les types de publications du MeSH sont organisés hiérarchiquement que depuis 1997.

3.1.2.4 LA TERMINOLOGIE CISMEF : UNE TERMINOLOGIE FONDEE SUR LE MESH

La terminologie CISMeF encapsule la version française du thésaurus MeSH (Douyère et al., 2004) dans la mesure où elle représente une extension des concepts déjà existants, d'une part, et elle emploie de nouveaux concepts, d'autre part. En effet, dans le but d'adapter le MeSH et faire face à la problématique de l'indexation des ressources de santé disponibles sur l'Internet, plusieurs améliorations ont été réalisées depuis la création du catalogue en 1995.

En plus des descripteurs MeSH (termes qui permettent l'indexation des ressources), des qualificatifs MeSH (qui permettent de préciser le sens d'un descripteur et d'en souligner un aspect particulier) et des concepts chimiques supplémentaires, les notions de métatermes, de types de ressources et de stratégies de recherche (requêtes préconstruites) ont été ajoutées.

1. Les types de ressources (TR) (N=300) sont une extension des types de publication de MEDLINE. Comme l'a défini Dublin Core Metadata Initiative, « *les types de ressources sont utilisés afin de catégoriser la nature du contenu de la ressource* ». Par exemple, dans le cas d'une ressource constituant un guide de bonnes pratiques concernant l'intoxication au monoxyde de carbone, le descripteur MeSH « intoxication au monoxyde de carbone » pour le contenu et, le type de ressource « recommandation pour la pratique clinique» pour le contenant, sont deux termes d'indexation (parmi d'autres) pour la ressource.

   Les types de ressources de la terminologie CISMeF sont organisés, pareillement que les descripteurs et les qualificatifs MeSH, hiérarchiquement avec des relations de subsomption[57].

[56] Medical Subject Heading; MeSH Record Types. URL: http://www.nlm.nih.gov/mesh/intro_record_types.html (2009)

2. Les métatermes (MT) (N=126) sont des super-concepts, qui ont été conçus pour représenter, généralement, une spécialité médicale ou une science biologique. Les métatermes ont été sélectionnés manuellement par le conservateur des bibliothèques de l'équipe CISMeF (Benoit Thirion) et ont des liens sémantiques avec un ou plusieurs descripteurs MeSH, qualificatifs, types de ressources et stratégies de recherche (cf. Figure 3.1.2.4). Par exemple, le métaterme « oncologie » a des liens sémantiques avec le type de ressource « service oncologie hôpital », le qualificatif « radiothérapie » et les descripteurs « cancérogènes », « gènes tumoraux », « cellules souches tumorales »…

   Une étude de l'équipe CISMeF a montré que l'utilisation des MT améliore la recherche d'information (Gehanno et al., 2007). En effet, les résultats ont montré que l'utilisation des requêtes avec seulement des descripteurs MeSH ont un rappel de 0,44 par rapport à 1 en cas d'utilisation des métatermes. Par exemple, introduire le terme « psychiatrie » comme métaterme constitue une stratégie plus efficace pour avoir plus de résultats : au lieu d'exploser une seule hiérarchie MeSH correspondant au descripteur MeSH « psychiatrie », une expansion automatique de la requête est réalisée en explosant les hiérarchies de tous les termes sémantiquement liés à ce MT « psychiatrie », comme le descripteur MeSH « hôpital psychiatrique » ou encore le type de ressource « centre public santé mentale ».

3. Les stratégies de recherche (N=316) sont des requêtes préparées et préconstruites par l'équipe CISMeF afin de faire face aux ambiguïtés des requêtes des utilisateurs et pour améliorer la recherche d'informations sur des notions qui n'ont pas d'équivalent dans le reste de la terminologie CISMeF.

   En effet, ils existent des concepts médicaux importants qui ne sont pas représentés par la terminologie (et notamment par le thésaurus MeSH) tel que « *chirurgie du côlon* ». Dans ce cas, la requête de l'utilisateur sera réécrite sous la forme « *maladies du côlon/chirurgie.mc ou colon/chirurgie.mc ou colectomie.mc ou colostomie.mc ou coloscopie.mc ou coloscope.mc* » afin de rechercher, par exemple, les ressources indexées par le descripteur MeSH « *maladies du côlon* » et se rapportant au thème de la chirurgie.

[57] La subsomption désigne une relation hiérarchique entre des concepts, dans les logiques de description. Cette notion est proche de la relation « est impliqué par » en logique classique, ou encore « contient » en logique ensembliste.

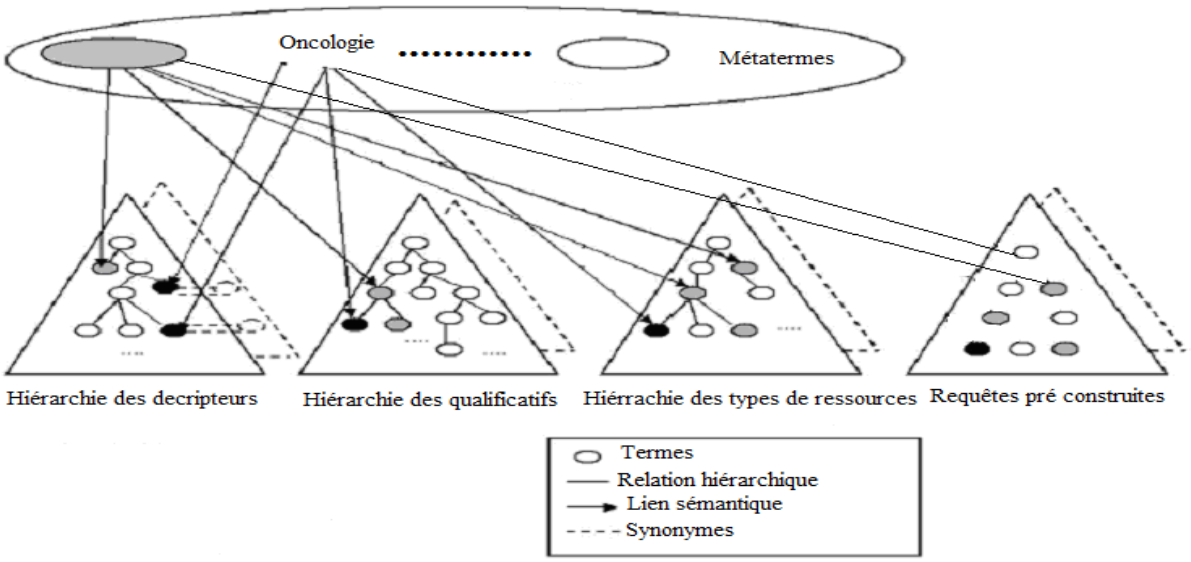


Figure 3.1.2.4. La terminologie CISMeF : lien sémantique entre les métatermes et les descripteurs, qualificatifs MeSH, les types de ressources et les requêtes préconstruites.

En outre, plusieurs autres améliorations autour du MeSH ont été mises en application :

1- dans MEDLINE, la paire (descripteur/qualificatif) décrit le centre d'intérêt de la ressource. Les qualificatifs MeSH sont associés aux descripteurs pour mieux décrire et spécifier le centre d'intérêt de la ressource ou un aspect particulier de cette dernière. Par exemple, une ressource concernant le traitement médicamenteux de l'asthme est indexée, entre autres, par la paire *asthme/traitement médicamenteux*.

   Dans la même perspective, et pour mieux spécifier l'indexation des ressources et améliorer par la suite la recherche d'information, l'équipe CISMeF a proposé une combinaison de triplet « *(descripteur/qualificatif)\type de ressource* » (Darmoni et al. 2007). Par exemple, si une ressource est indexée par le terme « *asthme/thérapie\recommandation* », cela signifie qu'il s'agit d'une recommandation des méthodes thérapeutiques pour l'asthme ;

2- l'indexation par majeur/mineur pour les types de ressources et les métatermes. En effet, cette spécificité concerne non seulement les descripteurs MeSH mais aussi les types de ressources et les métatermes. Une notion peut être majeure lorsqu'elle est traitée de façon prépondérante. On parle également de pondération, c'est la mesure de l'importance d'une notion dans un document. Les termes en majeur décrivent les principales idées abordées dans la ressource, alors que les termes en mineur représentent les concepts marginaux ;

3- l'enrichissement des concepts chimiques supplémentaires. Depuis leur intégration dans le serveur de terminologie CISMeF, plus de 8.576 ont été traduits en français et plus de 10.000 synonymes français ont été créés et intégrés[58].

[58] Statistiques datant du Septembre 2010.

4- Combinaison (« *concept chimique supplémentaire/qualificatif* ») des concepts chimiques supplémentaires (CCSs) avec certains qualificatifs (tels que : administration et posologie, pharmacologie, intoxication…). Dans MEDLINE, cette association n'est pas possible.

   L'objectif de cette nouvelle fonctionnalité est l'amélioration de l'indexation et de la recherche d'information au sein du catalogue CISMeF.

Toutes ces améliorations sont disponibles sur le portail terminologique CISMeF (URL : http://terminologiecismef.chu-rouen.fr/) bientôt remplacé par le portail multi-terminologique de santé (URL : http://pts.chu-rouen.fr/).

3.1.2.5 QUELQUES EXEMPLES D'AUTRES TERMINOLOGIES MEDICALES

Dans le cadre du projet PSIP, six terminologies ont été utilisées pour l'extraction des données à partir des différentes bases de données ainsi pour l'indexation des documents non structurés : la CIM-10 a été utilisée pour les diagnostics ; la classification ATC pour les médicaments ; la nomenclature IUPAC pour les tests cliniques ; la WHO-ICPS pour la sécurité des patients ; la taxonomie NCCMERP pour la description des effets indésirables des médicaments et la taxonomie de PSIP pour la description des éventuelles situations dangereuses de la médication (Darmoni et al., 2010).

- **La classification CIM-10[59]**

L'origine de la CIM (Classification Internationale des Maladies) remonte aux années 1850, avec *the International List of Causes of Death* de William Farr, qui reprenait, entre autres, les travaux de John Graunt datant de 1700 (Greenwood, 1948). Elle était adoptée par *the International Statistical Institute* en 1893, grâce aux travaux de Jacques Bertillon qui publie la Nomenclature Internationale des Causes de Décès (Bertillon, 1912). C'est à partir de cette classification que nait la première révision en 1900 avec comme principe d'une mise à jour décennale.

La 6ième révision de cette classification est adoptée par l'Organisation Mondiale de la Santé (OMS) en 1948 ((Organisation Mondiale de la Santé, 1950a), (Organisation Mondiale de la Santé, 1950b)). Ensuite jusqu'en 1996, la 9ème révision (CIM-9) « *Classification Internationale des Maladies, Traumatismes et Causes de Décès »* a été utilisée dans le cadre du PMSI (Programme de Médicalisation des Systèmes d'Information). En 1993, la 10ème révision (CIM-10) *« Classification statistique internationale des maladies et des problèmes de santé connexes »* a vu le jour (Organisation Mondiale de la Santé, 1993), alors que, la CIM-9 est encore utilisée dans certains pays tels que les États-Unis, l'Espagne….

En 1994, la CIM-10 analytique (Vol.1 ; V.F) *Table analytique* a été réalisée. Il s'agit de la classification elle-même, la classification de la morphologie des tumeurs, les listes pour les

[59] Une version 11 de la CIM est encore d'élaboration. Notre portail multi-terminologique pourrait être utilisé pour gérer la version française de cette future CIM-11.

mises en tableaux, les définitions, le règlement. En 1995, le manuel d'utilisation (Vol.2 ; V.F) de la CIM-10 *Mode d'utilisation* a été établi. Il s'agit des indications et des instructions pour l'utilisation du volume 1. En 1996, le manuel d'utilisation (Vol.3 ; V.F) de la CIM-10 *Index alphabétique* a été mis en place.

La classification a comme but de permettre l'analyse systématique, l'interprétation et la comparaison des données de mortalité et de morbidité recueillies dans différents pays ou régions et à des époques différentes[60]. (World Health Organizations, 2010)

Ainsi, elle représente l'unique classification diagnostique internationale pour :

- ✓ l'épidémiologie, et la description des problèmes de prise en charge sanitaire ;
- ✓ l'étude des problèmes financiers (recouvrement des coûts, allocation de fond).

La classification est mono-axiale et faite soit par :

- ✓ systèmes : par exemple, maladies cardio-vasculaires, Digestives ;
- ✓ étiologies (causes des maladies) : maladies infectieuses, tumeurs.

La CIM-10 est ordonnée en une hiérarchie à héritage simple. Chaque terme possède un ascendant unique. La hiérarchie de la CIM-10 est organisée jusqu'à 6 niveaux et elle est partitionnée en 21 chapitres classés par appareil fonctionnel et représentés par une lettre (exemple : la lettre E est associée au chapitre « Maladies endocriniennes, nutritionnelles et métaboliques »).

Les chapitres sont divisés en groupes, eux-mêmes divisés en catégories à 3 caractères qui sont répertoriés en sous-catégories à 4 caractères. Les catégories à 3 caractères représentent l'unité diagnostique signifiante de base ; c'est-à-dire le niveau minimum de codification dans la plupart des pays. Enfin, des subdivisions peuvent apparaitre de manière facultative dans certains chapitres. A chaque niveau (chapitre, catégorie, sous-catégories), la CIM-10 peut indiquer des inclusions ou des exclusions permettant d'orienter vers une autre partie de la classification.

❖ La nomenclature IUPAC[61]

La nomenclature IUPAC est un système pour nommer les composés chimiques et pour décrire la science de la chimie en général. Elle est développée et mise à jour sous les auspices de l'organisme international IUPAC (International Union of Pure and Applied Chemistry).

IUPAC est la nomenclature officielle en chimie organique. La nomenclature en chimie est l'ensemble des règles, symboles, vocables, destinés à représenter et à prononcer les noms des corps étudiés.

L'objectif essentiel d'une nomenclature est d'aboutir à des noms de composés chimiques sans ambiguïté, à savoir qu'un même nom ne doit jamais servir à désigner deux composés

[60] Classification statistique internationale des maladies et des problèmes de santé connexes- CIM-10. URL : http://www.spieao.uhp-nancy.fr/~kohler/CIM10/CIM10.HTM

[61] Home page of International Union of Pure And Applied Chemistry; URL: http://www.chem.qmul.ac.uk/iupac/

chimiques différents. Par contre, un même composé chimique suffisamment complexe peut recevoir plusieurs noms différents provenant de différentes nomenclatures, ou même parfois provenant de la même nomenclature.

Il est préférable que le nom de la substance chimique contienne quelques informations au sujet de la structure ou de la composition chimique du composé. Les numéros CAS sont un exemple extrême de noms qui ne remplissent pas cette fonction : chaque numéro réfère à un unique composé mais aucun ne contient d'information au sujet de la structure. Par exemple, nous pouvons être tentés d'ajouter du [7647-14-5] dans son assiette, mais pas du [133-43-9] : le premier est du chlorure de sodium, le second du cyanure de sodium.

❖ La WHO-International Patient Safety Classification (ICPS)[62]

Le but de la ICPS (traduction en français : Classification Internationale pour la sécurité des patients) est de permettre la catégorisation de l'information sur la sécurité des patients en utilisant un ensemble normalisé de concepts avec des définitions prédéfinies, des termes préférés et des relations entre ces derniers en se basant sur une ontologie explicite de domaine. Elle est conçue pour faciliter la description, la comparaison, la surveillance, l'analyse et l'interprétation de l''information afin d'améliorer le soin des patients, et pour des fins de planification épidémiologique et sanitaire (World Alliance & WHO Health Information Systems Department, 2009).

L'ICPS n'est pas encore une classification. C'est un projet conceptuel pour une classification internationale qui fournit une compréhension raisonnable de la sécurité des patients à laquelle il existe des classifications nationales qui peuvent y faire référence. L'ICPS est multiaxiale et hiérarchique selon dix classes.

Parmi les utilisations prévues de l'ICPS, nous mentionnons :

- ✓ comparer les données des incidents de sécurité des patients pour les différentes disciplines et entre les organismes locaux, nationaux et internationaux ;
- ✓ développer les connaissances concernant les incidents de sécurité des patients ;
- ✓ déterminer les problèmes liés à la sécurité des patients dans les différents secteurs de soin ;
- ✓ examiner le rôle des facteurs humains et le rôle des systèmes pour la sécurité des patients ;
- ✓ déterminer les applications et les limitations des stratégies existantes pour réduire le facteur risque ;
- ✓ identifier les éventuels problèmes liés de la sécurité des patients à travers les recherches basées sur l'évidence ;
- ✓ développer des solutions de priorités et de sécurité.

[62] World Health Organization; International Classification for Patient Safety (ICPS). URL : http://www.who.int/patientsafety/implementation/taxonomy/en/index.html

❖ La taxonomie des erreurs médicamenteuses : National Coordinating Council for Medication Error Reporting and Prevention (NCCMERP)[63]

Le but de cette taxonomie est de fournir un langage et une structure standards des données liées aux erreurs médicamenteuses pour le développement des bases de données analysant les rapports d'erreurs médicamenteuses.

La taxonomie NCCMERP est la classification de référence des conséquences cliniques d'erreurs médicamenteuses par niveau de gravité et par importance de préjudice (National Coordinating Council for Medication Errors Reporting and Prevention NCCMERP, 2002).

Cette classification est indispensable à l'analyse approfondie des erreurs médicamenteuses et conditionne la qualité des échanges entre les programmes de recueil et de prévention des erreurs médicamenteuses.

Il est recommandé d'utiliser la taxonomie du NCCMERP, classification des causes d'erreur médicamenteuse employée par la plupart des programmes de recueil et de prévention d'erreurs médicamenteuses, notamment par le Réseau REEM (Schmitt et al. 2006).

Les causes d'erreur médicamenteuse peuvent être définies comme facteurs (situation, événement) antérieurs à l'erreur médicamenteuse et peuvent être reconnus comme étant à l'origine de la survenue d'une erreur médicamenteuse. Chercher la ou les causes d'une erreur médicamenteuse, c'est répondre à la question : « Pourquoi l'erreur médicamenteuse s'est-elle produite ? ». Dans le cas d'une cascade d'erreurs médicamenteuses, la cause directe de l'erreur médicamenteuse est la conséquence d'une erreur primitive.

La taxonomie NCCMERP est multiaxiale et dispose d'une hiérarchie de vingt cinq classes.

❖ La taxonomie du PSIP

La sureté des médicaments est une composante essentielle de la sécurité des patients. À l'échelle mondiale, elle dépend de la puissance des systèmes nationaux qui contrôlent la mise au point et la qualité des médicaments, notifient leurs effets nocifs et fournissent des informations exactes pour les utiliser sans danger[64].

Les réactions nocives et inattendues aux médicaments qui se produisent aux posologies thérapeutiques habituelles sont appelées effets indésirables des médicaments. Ceux-ci font partie des principales causes de mortalité dans de nombreux pays.

On appelle pharmacovigilance, la prévention et la détection des effets indésirables des médicaments. L'évaluation attentive des risques et des bienfaits des médicaments s'applique tout au long de leur cycle de vie, depuis la phase précédant l'homologation jusqu'à leur utilisation.

---

[63] National Coordinating for Medication Error Reporting and Prevention. URL: http://www.nccmerp.org/

[64] Organisation Mondiale de la Santé ; Médicament : sécurité et effets indésirables. URL : http://www.who.int/mediacentre/factsheets/fs293/fr/index.html

La circulation des informations à l'échelle mondiale sur les effets indésirables renforce la sécurité des médicaments dans les pays et peut se traduire par des décisions politiques prises en temps voulu pour préserver la sécurité des patients lorsqu'un problème surgit.

Pour cet effet, certains des systèmes de détection des incidents, des effets indésirables des médicaments ou des erreurs médicales étaient spécifiquement conçus pour la détection des effets indésirables des médicaments et d'identifier les facteurs qui les causent. Tous ces systèmes de détection sont soutenus, explicitement ou implicitement, par des taxonomies décrivant une description structurée des effets détectés.

Dans le cadre du projet PSIP, un grand ensemble de données médicales générées par le modèle de données (Darmoni et al., 2010) est déjà disponible pour décrire les cas susceptibles d'être des effets indésirables des médicaments. Cependant, le modèle a été conçu pour l'exploitation et l'extraction des données et, par conséquent, il a besoin d'adaptation pour être employé comme base pour la conception des différents modules des systèmes d'aide à la décision clinique.

Par ailleurs, les taxonomies employées pour la détection des effets indésirables des médicaments expriment moins d'information que le modèle de données de PSIP déjà établi. Il est alors souhaitable de fusionner ce modèle avec une taxonomie existante pour fournir une description structurée plus riche et plus complète pour détecter les effets indésirables des médicaments. Dans un but de trouver la meilleure combinaison, sept taxonomies existantes concernant les effets indésirables médicaux ont été analysées et employées avec le modèle de données de PSIP :

- ✓ **NCC-MERP** : National Coordinating Council for Medication Error Reporting and Prevention: http://www.nccmerp.org
- ✓ **AAQTE** (Bates et al. 2003) : Association for Quality Assurance in Therapeutics and Evaluation: http://adiph.org/aaqte/index.html
- ✓ **USP-ISMP** (Morimoto et al. 2004) : U.S. Pharmacopeia (USP) - Institute for Safe Medication Practices (ISMP). URL: https://www.ismp.org/orderForms/reporterrortoISMP.asp
- ✓ **MedWatch** : US Food and Drug Administration (FDA). Cela concerne les réactions indésirables, les problèmes de qualité des produits et les erreurs d'utilisation. URL: http://www.fda.gov/medwatch
- ✓ **ICPS** : International Classification for Patient Safety (ICPS). URL: http://www.who-icps.org/
- ✓ **DPSD** : Danish Patient Safety Database Danish National Board of Health. URL: www.dpsd.dk.
- ✓ **JCAHO** (Beuscart-Zephir et al. 2009) : Joint Commission on Accreditation of Healthcare Organizations (US)**.**

La combinaison des taxonomies avec le modèle de données de PSIP a permis d'identifier 16 catégories tels que données patientes, données de séjour, diagnostics, procédures, type d'erreur, cause de l'erreur…

La taxonomie de PSIP a une structure multiaxiale et hiérarchisée.

3.2 PASSAGE DU MONDE MONO-TERMINOLOGIQUE VERS UN UNIVERS MULTI-TERMINOLOGIQUE

Le besoin de passage d'un monde mono-terminologique (limité au thésaurus MeSH, pour l'indexation et la recherche) à un univers multi-terminologique (fondée sur plusieurs terminologies médicales) est ressenti de plus en plus par le fait que chaque terminologie a des objectifs et des contextes d'utilisation différents, d'une part, et pour pallier les éventuelles imperfections du thésaurus MeSH en termes d'indexation et de recherche d'information, d'autre part.

En effet, selon le contexte d'utilisation certaines terminologies peuvent s'avérées plus appropriées que d'autres. Par exemple, un pharmacien ou un chimiste utilise mieux la classification ATC ou le code CAS pour rechercher un document spécifique à ses attentes plutôt qu'une autre terminologie. À l'inverse, un étudiant en médecine pourrait préférer employer le thésaurus MeSH pour rechercher ses documents bibliographiques.

Ainsi, dans ce cadre multi-contextes et avec un souci d'améliorer le système actuel afin d'avoir une recherche d'information plus exhaustive et plus efficace, l'équipe CISMeF a pris la décision stratégique de passer d'un monde mono-terminologique à un univers multi-terminologique.

Cet objectif doit prendre en compte la disponibilité de plusieurs terminologies, classifications, thésaurus et nomenclatures médicaux disponibles en français[65] et, la mise en pratique des interactions existantes entre ces derniers.

Dans le domaine médical, UMLS (Unified Medical Language System) est le programme de recherche lancé par la NLM pour établir des sources de connaissance afin de faciliter le développement des systèmes qui aident les professionnels de santé à rechercher une information biomédicale.

Les sources de connaissance peuvent être employées pour lier les systèmes d'information hétérogènes et pallier les problèmes d'intégration de plusieurs terminologies à cause de leurs différences. Les trois sources de connaissance de l'UMLS sont le Métathesaurus, le réseau sémantique, et un lexique médical Specialist Lexicon[66].

Ainsi, l'un des objectifs de l'UMLS est de fournir une plateforme permettant de regrouper tous les thésaurus, nomenclatures et classifications existants dans le domaine médical (Bodenreider, 2004).

[65] Il y a 130 terminologies de santé en anglais contre une dizaine en français

[66] National Library of Medicine; Unified Medical Language System (UMLS). URL : http://www.nlm.nih.gov/research/umls/about_umls.html

Dans le cadre de la recherche médicale, plusieurs liaisons et interactions entre terminologies ont été réalisées (Merabti, 2010). Parmi lesquelles, nous évoquons :

- ✓ l'alignement conceptuel en passant par l'UMLS. Par exemple, la mise en correspondance des terminologies disponibles dans le méta thésaurus UMLS ;
- ✓ l'alignement manuel : exemple MeSH-CCAM ; MeSH-ATC ;
- ✓ l'alignement automatique avec les outils du TAL : exemple Orphanet-CIM-10.

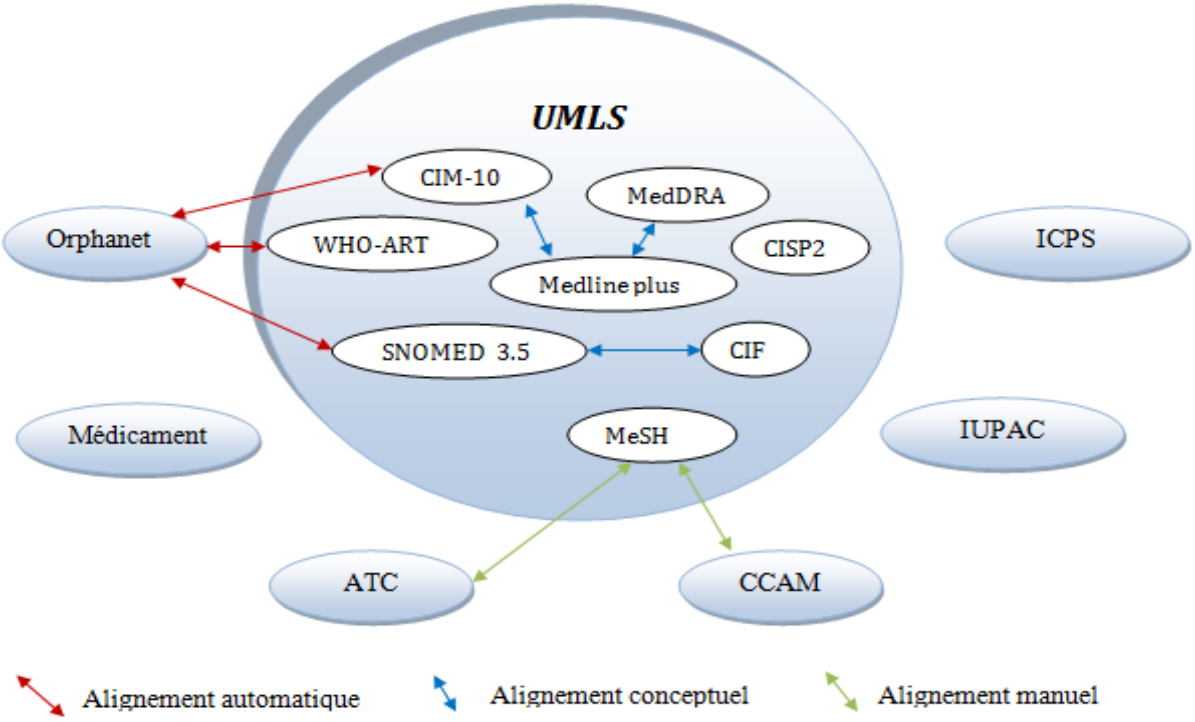


Figure 3.2.1. Relations existantes entre les terminologies médicales

En effet, grâce aux différentes relations terminologiques nous pouvons améliorer la recherche d'information et mieux répondre à la requête de l'utilisateur et ce via l'expansion ou la reformulation de la requête. Les relations inter et intra terminologiques permettent d'assurer la navigation entre les terminologies. Nous pouvons chercher toutes les liaisons possibles entre les termes de la requête appartenant à une terminologie donnée et tous les termes des autres terminologies qui ont une correspondance avec les termes en question. Cette procédure permet d'élargir le champ de la recherche de l'utilisateur selon son contexte, sans néanmoins mettre en cause la pertinence thématique de l'information ni le degré de précision du système. Par exemple, grâce à la correspondance entre le terme MeSH « *appareil correction auditive* » et le terme SNOMED « *prothèse auditive* », nous pouvons enrichir notre résultat et retrouver toutes les ressources indexées par l'un ou l'autre de ces termes.

Le passage à un univers multi-terminologique[67] se traduit par l'intégration, dans le back-office de CISMeF, des terminologies principales de santé disponibles en français (cf. Figure 3.2.2) :

[67] Ce passage a nécessité l'intervention de trois ingénieurs de l'équipe CISMeF assisté par deux équipes de huit ingénieurs de l'INSA de Rouen.

- ✓ le thésaurus MeSH ;
- ✓ la SNOMED 3.5 ((Coté, 1986) ; (Cornet et al., 2008) ; (Lussier et al., 1998)) ;
- ✓ la CIF (Classification Internationale du Fonctionnement, du handicap et de la santé)[68] (Baron, 2008) ;
- ✓ la CIM-10 (World Health Organizations, 2010) ;
- ✓ la CCAM (Classification Commune des Actes Médicaux) ((Hanser et al., 2006) ; (Zaiss et al., 2007)) ;
- ✓ la CISP2 (Classification Internationale des Soins Primaires, deuxième édition) (Soler et al., 2008) ;
- ✓ le DRC (Dictionnaire des Résultats de Consultation) (Morel, 1996) ;
- ✓ la classification ATC (Anatomical Therapeutic Chemical) ;
- ✓ le MedDRA (Medical Dictionary for Regulatory Activities) ((Bousquet et al,. 2004) ; (Santé Canade, 2010)) ;
- ✓ Medline plus (Miller et al., 2000) ;
- ✓ la WHO-ART (World Health Organisation – Adverse Reaction Terminology)[69] (Brown, 2002) ;
- ✓ la WHO-ICPS (International Classification for Patient Safety) ;
- ✓ le thésaurus Orphanet pour décrire les maladies rares (Aymé et al., 1998).

[68] World Health Organization International Classification of Functioning, Disability and Health. URL: http://www.who.int/classifications/icf/en/

[69] World Health Organization Adverse Reactions Terminology. URL : http://www.umc-products.com/DynPage.aspx?id=4918

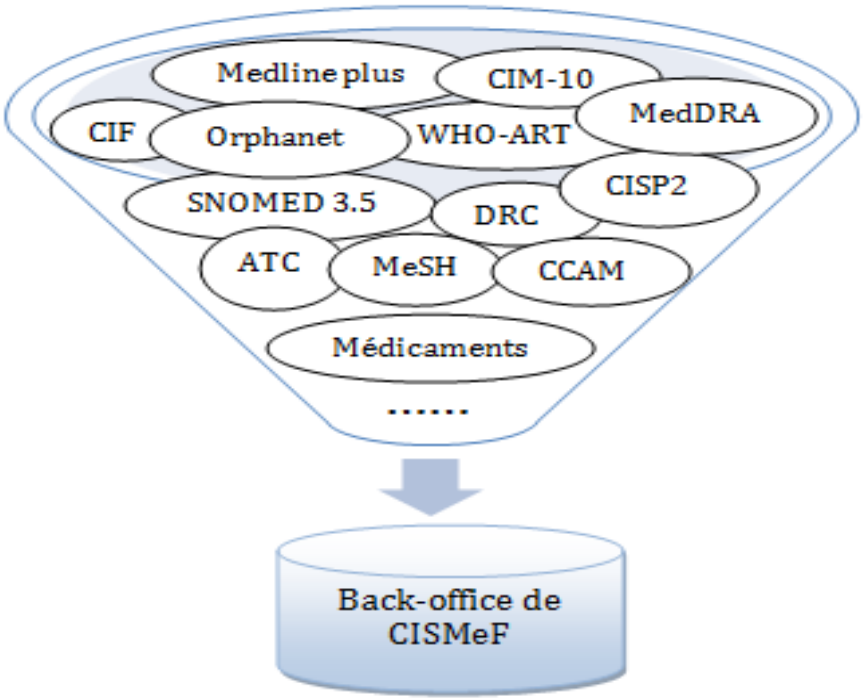


Figure 3.2.2. Intégration des terminologies médicales dans le back-office de CISMeF

Lors de cette intégration, il a fallu prendre en compte les formats d'origine (forme SQL, fichiers texte, bases de données,…) de toutes ces terminologies, afin de générer un modèle générique et uniforme. Cette tâche est réalisée en développant des parseurs générant le format RDF[70].

Pour réunir dans une même structure des terminologies, des thésaurus, des nomenclatures et des classifications de natures différentes ayant chacun des spécificités particulières, il a fallu, d'abord, modéliser chacune de ces terminologies[71]. Ensuite, nous avons créé un modèle générique qui tient compte de toutes ces particularités, pour avoir la possibilité d'intégrer d'autres terminologies en cas de besoin.

Le modèle générique obtenu est décrit ci-dessous (cf. Figure 3.2.3) :

[70] Voir Annexe A pour le parseur ATC
[71] Voir Annexe A pour quelques exemples de modélisations de quelques terminologies médicales.

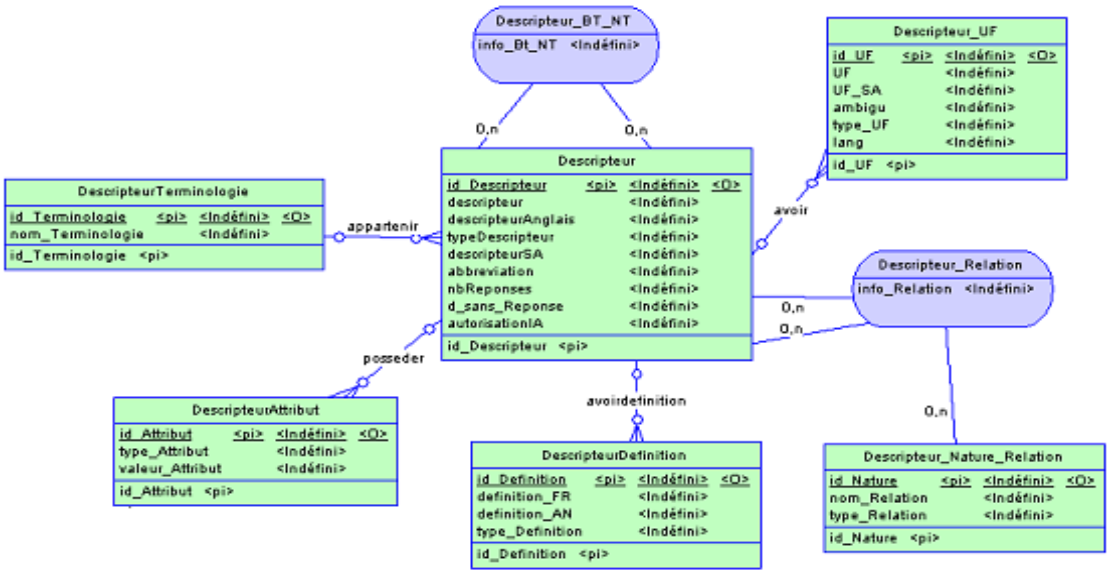


Figure 3.2.3. Le modèle générique dans le cadre de la recherche d'information multi-terminologique

Ce modèle est centré sur l'entité **Descripteur**. Celle-ci inclut tous les termes, les mots clés, les qualificatifs, les types de ressources, les métatermes, les éléments, les noms de catégories, les groupes, les blocs, les chapitres qui peuvent exister au niveau des terminologies. Cette classe définit les attributs communs aux différentes terminologies. Les attributs spécifiques sont représentés par l'entité **DescripteurAttribut**, ce qui permet d'être plus générique et plus souple, lors de la mise à jour du modèle. Les définitions des descripteurs sont multilingues et de types différents (DRC, MeSH, Vidal…) et sont décrites par l'entité **Descripteur_Definition**. Les synonymes sont illustrés par l'entité **Descripteur_UF**. N'étant pas considérés comme des termes d'indexation, les synonymes permettent de représenter un terme de plusieurs façons et, par la suite, un enrichissement de la requête de l'utilisateur lors du processus de recherche d'information.

Un descripteur peut avoir plusieurs synonymes. Chaque descripteur appartient à une des terminologies intégrées. Ces dernières sont représentées par l'entité **DescripteurTerminologie**, dont la structure est définie par l'identifiant de la terminologie (id_Terminologie) et le nom de la terminologie (nom_Terminologie). La relation **Descripteur_BT_NT** est définie entre deux descripteurs et réservée pour les relations hiérarchiques au sein d'une même terminologie, à l'inverse de l'association Descripteur_Relation qui décrit les relations non hiérarchiques intra-terminologiques (si les deux descripteurs appartiennent à la même terminologie) et les relations inter-terminologiques (si les deux descripteurs appartiennent à des terminologies distinctes). Ces derniers types de relations permettent de relier des terminologies entre elles, inspiration faite du réseau sémantique et du méta-thésaurus d'UMLS. L'entité **Descripteur_Nature_Relation** a pour but de définir les types de relations non hiérarchiques, elle contient les noms et les types des toutes les relations existantes entre deux descripteurs quelconques appartenant à la même terminologie ou non, telles que *ne pas confondre*, *voir aussi*, *inclusion*, *exclusion*, ...

La validation de notre modèle générique se traduit par l'implantation d'une recherche d'information multi-terminologique dans notre système d'information CISMeF et la mise en place du Portail Terminologique de Santé (PTS). Des études seront menées (suite à l'utilisation de CISMeF et du PTS) permettront de vérifier et discuter le modèle.

Ainsi, l'application du modèle nous a permis d'enrichir et d'améliorer la recherche d'information. Par exemple, en mono-terminologie avec le thésaurus MeSH une recherche concernant la requête « *appareil locomoteur* » fournit 1.013 ressources tandis que cette même requête en multi-terminologie fournit 1.505 ressources (cf. Figure 3.2.4 et Figure 3.2.5).

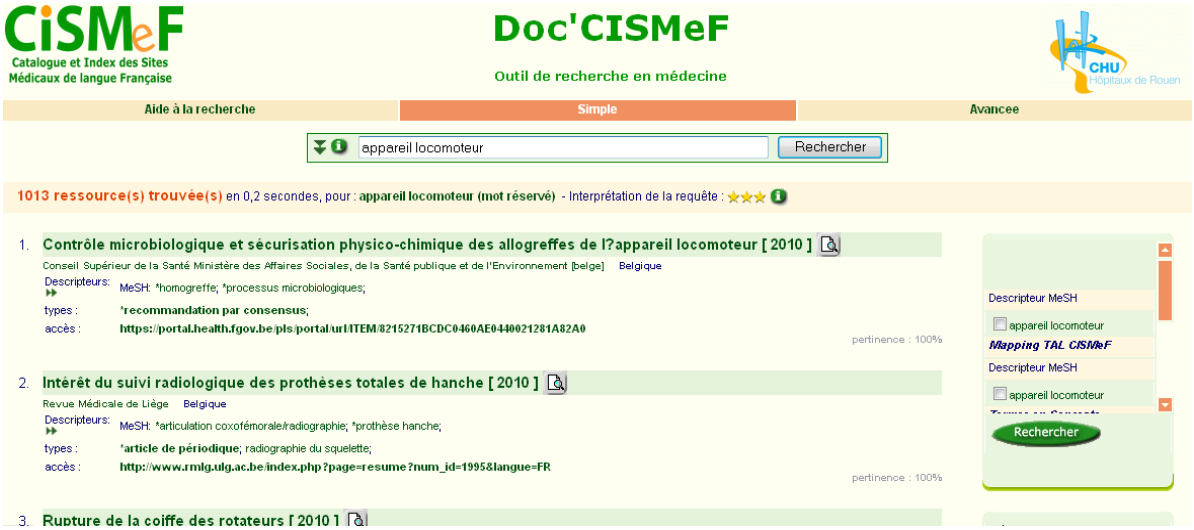


Figure 3.2.4. Résultat de la recherche d'information mono terminologique pour la requête « appareil locomoteur »

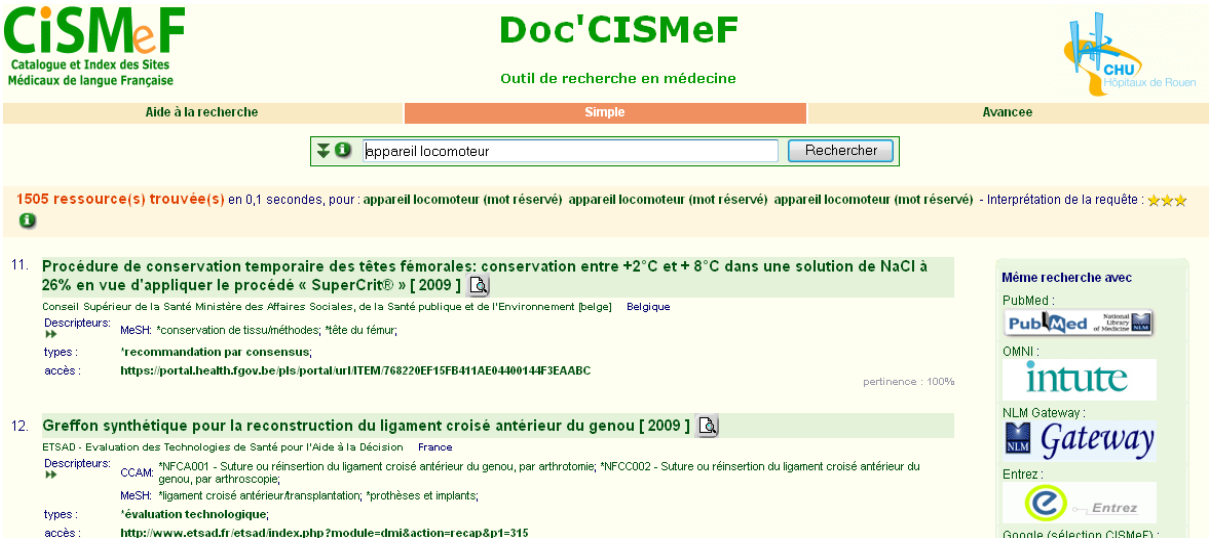


Figure 3.2.5. Résultat de la recherche d'information multi-terminologique pour la requête « appareil locomoteur »

Par ailleurs, grâce au modèle générique de l'univers multi-terminologique, nous avons mis en œuvre le PTS qui permet un accès groupé aux principales terminologies de santé disponibles

en français sans se soucier, ni de leur gestion, ni de leur mise à jour.

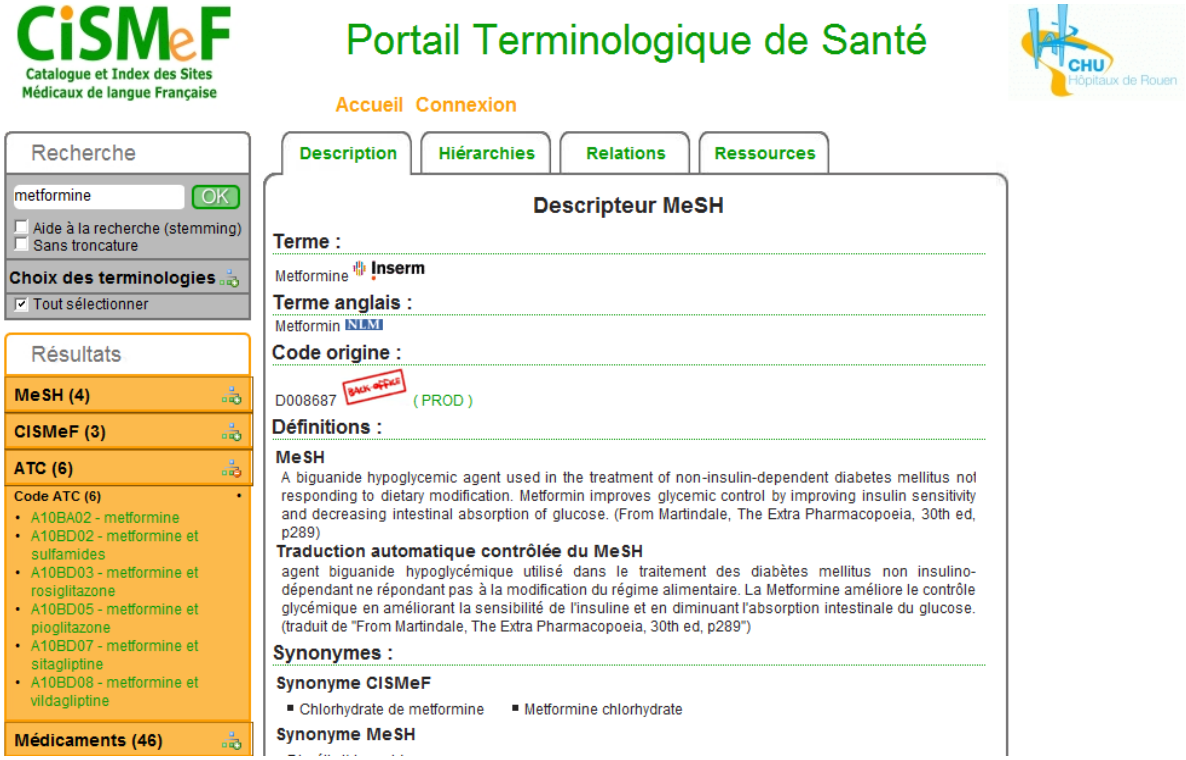


Figure 3.2.6. Page de recherche multi-terminologique au sein du Portail de Terminologies de Santé (PTS)

À partir de cette page de recherche multi-terminologique au sein du PTS, nous pouvons mettre en relief quatre onglets :

- *description* permettant de définir le terme recherché ;
- *hiérarchie* permettant d'accéder aux hiérarchies de toutes les terminologies ;
- *relation* permettant de connaître toutes les relations intra et inter-terminologies favorisant la navigation entre ces dernières ;
- *ressources* donnant accès contextuel à 50 sites de bases d'information en français (CISMeF) et en anglais (Pub Med).

CONCLUSION

A travers ce chapitre, nous avons donné un aperçu sur les définitions et les caractéristiques des terminologies médicales que nous avons manipulé au cours de la thèse et, celles utilisées dans le cadre du projet PSIP.

Le passage du monde mono-terminologique fondé essentiellement sur le thésaurus MeSH vers l'univers multi-terminologique enrichi par les différentes classifications, nomenclatures et thésaurus nous a permis d'améliorer la recherche d'information et d'avoir une information médicale plus appropriée pour l'utilisateur du catalogue CISMeF. Nous expliquons dans le cinquième chapitre, plus en détails, l'algorithme de recherche d'information multi-terminologique ainsi que l'apport présumé de cet univers.

CHAPITRE 4

APPROCHE DE L'INDEXATION AUTOMATIQUE POUR LES MEDICAMENTS

INTRODUCTION

Dans ce chapitre, nous présentons la première réalisation faite autour de l'univers multi-terminologique ; à savoir la création d'un Portail d'Information bilingue sur les Médicaments (PIM). Cette réalisation nous a permis, par la suite, une exploitation plus analytique des informations concernant les médicaments, en mettant en place une approche d'indexation automatique par la classification ATC. Enfin, nous concluons ce chapitre par l'exposition des résultats de l'étude réalisée, mettant en avant les avantages de la correspondance entre le thésaurus MeSH et la classification ATC pour améliorer la recherche d'information.

4.1 CREATION DU PORTAIL D'INFORMATION SUR LES MEDICAMENTS

4.1.1 ÉTUDE DE L'EXISTANT

D'après le dictionnaire Larousse, en informatique, un portail est « *un site conçu pour être le point d'entrée sur Internet et proposant aux utilisateurs des services thématiques et personnalisés* ». Se focalisant sur le domaine de la santé, un portail doit avoir certaines propriétés spécifiques et respecter des standards de qualité (Koch, 2000). En France, la référence utilisée pour certifier les sites de e-santé est le Health On the Net code (Boyer, 2007) qui a été sélectionné par la Haute Autorité de Santé en 2007. Ainsi, pour avoir une information précise concernant le domaine de la santé et en particulier pour les médicaments, un utilisateur pourrait bien avoir recours à ce type de mode d'accès.

En 2008, la NLM a mis en place le « Drug Information Portal»[72]. Ce portail représente une passerelle pour les utilisateurs afin d'avoir les informations concernant les médicaments de la NLM et d'autres agences gouvernementales. Il permet d'accéder aux informations concernant plus de 12.000 médicaments. La recherche peut s'y effectuer à partir du nom générique ou du nom commercial (ex. « phénol ») ou par catégorie (ex. « analgésiques » ou « anti-infectieux »). À notre connaissance, la recherche par les codes relatifs aux médicaments, tel que le code ATC, n'est pas possible.

Depuis 2009, en France, le portail public du médicament du gouvernement français *MedicFrance* est accessible en ligne sur « *http://www.portailmedicaments.sante.gouv.fr* ».

Ce portail devrait permettre au grand public de retrouver une information fiable, objective et récente sur les médicaments. La page de navigation présente les attributions des instances publiques nationales responsables des décisions en matière de médicaments ainsi que des liens vers ces sites et oriente l'internaute vers les informations pouvant être trouvées sur chacun de ces sites. Nous pouvons ainsi accéder au site de l'Agence française de sécurité sanitaire des produits de santé (AFSSAPS) qui évalue les bénéfices et les risques des médicaments, ou encore au site de la Haute Autorité de santé (HAS) qui évalue les médicaments en vue de leur remboursement… Enfin, à partir de cette même page de navigation, il est encore possible de consulter en ligne la base de données sur les médicaments de l'Assurance maladie qui porte sur les produits commercialisés en France.

Un moteur de recherche, ciblant les liens vers les sites institutionnels à partir des recherches formulées sur le portail, devrait progressivement être développé pour permettre des recherches de plus en plus précises, sans toutefois modifier les informations sources établies par les instances responsables des décisions.

[72] Drug Information Portal. URL : http://druginfo.nlm.nih.gov/drugportal/drugportal.jsp

4.1.2 LE PORTAIL D'INFORMATION SUR LES MEDICAMENTS DE L'EQUIPE CISMEF

Pendant la période (2007-2009) et dans le cadre du projet européen Patient Safety Through Intelligent Procedures in Medication (PSIP) (voir chapitre 1), nous nous sommes intéressés à mettre au point un Portail d'Information sur les Médicaments (PIM) (cf. Figure 4.1.2.1), permettant de faciliter l'accès aux principales ressources francophones concernant les médicaments (Letord et al., 2008). Le PIM est un portail d'information bilingue (français/anglais) sur les médicaments, dans un contexte multi-terminologique dans la mesure où les recherches peuvent s'effectuer grâce à plusieurs terminologies et/ou différents codes relatifs aux médicaments précédemment décrits. Le PIM se restreint, par un choix éditorial de l'équipe CISMeF aux informations médicamenteuses qui émanent d'institutions ou de sociétés savantes. Il s'est largement inspiré du back office de CISMeF et du moteur de recherche Doc'CISMeF. Le PIM est le résultat d'une collaboration entre l'équipe CISMeF et la société privée Vidal[73], spécialiste de l'information sur les médicaments.

Pour s'adapter à l'information sur les médicaments, l'équipe CISMeF a amélioré son serveur de terminologie, de façon à ce que les utilisateurs du PIM puissent accéder à toutes les substances chimiques (y compris médicamenteuses), aux actions pharmacologiques, ainsi qu'aux types de ressources liés aux médicaments.

En effet, au sein du thésaurus MeSH[74], les noms des substances chimiques (y compris les substances médicamenteuses) peuvent correspondre soit à des descripteurs hiérarchisés, soit à des concepts chimiques supplémentaires non hiérarchisés, soit à des synonymes de ces termes. Si l'on considère l'« information médicamenteuse », le plus important désormais est de retenir la notion de substance et non plus la notion du concept chimique supplémentaire ou descripteur MeSH. C'est pourquoi, pour les besoins du PIM, nous avons créé le concept « Substance » qui permet de regrouper l'ensemble des substances chimiques.

Au sein du thésaurus MeSH, comme au sein du serveur de terminologie CISMeF, la plupart des termes correspondant à des substances sont reliés à des actions pharmacologiques. Selon la 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* ». Ainsi, une action pharmacologique peut correspondre à un concept particulier qui permet de regrouper l'ensemble des substances (qu'il s'agisse de descripteurs (Des) ou de concepts chimiques supplémentaires (CCS)) ayant une action pharmacologique commune. Par exemple, l'action pharmacologique « antianémiques » permet de regrouper les substances suivantes : acide folique (Des), composés de fer III (Des), darbépoétine alfa (CCS), dextriferron (Des), époétine alfa (Des), extraits hépatiques (Des), ferric oxide, saccharated (CCS), gluconate ferreux (CCS), gluconate ferrique (CCS), hexaméthylène bisacétamide (CCS), hydroxocobalamine (Des), *iron protein succinylate* (CCS) et le téferrol (CCS).

[73] **VIDAL** | L'information de référence sur les produits de santé. URL : http://www.vidal.fr/
[74] Se référer au Chapitre 3, pour plus de détails concernant le thésaurus MeSH

À ce jour[75], 374 actions pharmacologiques provenant du MeSH ont été intégrées au serveur de terminologie CISMeF.

La terminologie CISMeF a dû aussi être adaptée à la nature des informations sur les médicaments et ce, grâce à l'ajout de types de ressources spécifiques du médicament. Une définition de chacun de ces types de ressources a été fournie soit par l'équipe CISMeF, soit par une institution (le plus souvent l'Agence française de sécurité sanitaire des produits de santé (AFSSAPS)). Ainsi, une arborescence spécifique sur les médicaments a été créée, avec en tête d'arborescence, le type de ressource le plus général « information sur le médicament ».

L'arborescence spécifique des types de ressources sur les médicaments :

Information sur le médicament
- Avis de vigilance sanitaire
- Évaluation médicament
  - Avis de la commission de transparence
- Formulaire pharmaceutique
- Monographie pharmacie
- Notice médicamenteuse
- Recommandation de bon usage du médicament
- Résumé des caractéristiques du produit

La mise en place d'une terminologie adaptée aux médicaments se fait aussi par l'intégration des noms commerciaux, des Dénominations Communes Internationales (DCI) et des différents codes nationaux et internationaux, liés aux médicaments et aux substances chimiques tels que le Code Identifiant de Présentation (CIP), le Code Identifiant de Spécialité (CIS) et l'Unité Commune de Dispensation (UCD) pour les codes nationaux et les codes de la classification Anatomique, Thérapeutique et Chimique (ATC), Chemical Abstract Service (CAS)[76], European Inventory of Existing Commercial Substances ou encore Inventaire Européen des Substances Commerciales Existantes (EINECS/ELINCS)[77] pour les codes internationaux. Ces fichiers nous ont été fournis, en partie, par la société Vidal.

Une fois la phase de prétraitement achevée, grâce à l'adaptation du serveur terminologique de CISMeF (intégration des actions pharmacologiques et les codes spécifiques aux médicaments, adaptation des types de ressources…), la construction du PIM (http://doccismef.chu-rouen.fr/servlets/PIM) a pu être réalisée. Le PIM s'inspire largement du portail CISMeF, bien qu'il ait des fonctionnalités spécifiques et plus orientées médicaments. Il a été développé en quatre étapes :

[75] Statistiques datant de Janvier 2010

[76] Se référer au chapitre 4

[77] Les codes EINECS/ELINCS sont représentés par un inventaire qui définit la liste définitive de toutes les substances chimiques censées se trouver sur le marché communautaire entre le 1er janvier 1971 et le 18 septembre 1981. Quant aux numéros European LIst of Notified Chemical Substances (ELINCS), ils sont décrits par une liste qui complète la liste EINECS et qui attribue un numéro aux nouvelles substances mises sur le marché européen après le 18 septembre 1981. Les nouvelles substances sont incluses au fur et à mesure de leur notification et paraissent lors des mises à jour de l'ELINCS.

- Étape 1

La première étape a été de créer le métaterme[78] « Médicaments » permettant de regrouper les descripteurs, les qualificatifs et les types de ressources qui correspondent à la thématique du médicament. Pour ce faire, nous avons rattaché manuellement au métaterme « médicaments » tous les descripteurs MeSH en rapport avec le médicament, tels que « *actions pharmacologique »s,* « *agrément de médicament* », « *contamination de médicaments* »[79] . . . Ensuite, nous avons sélectionné les qualificatifs qui sont utilisés pour l'indexation des documents relatifs à des médicaments, à savoir : *action des médicaments et substances chimiques*, *pharmacocinétique*, *traitement médicamenteux* et *administration et posologie*. Enfin, nous avons relié, à ce métaterme, les types de ressource concernant l'«information sur le médicament ».

De plus, de fait de l'organisation hiérarchique de la terminologie CISMeF (descripteurs, qualificatifs et types de ressources), tous les termes, hiérarchiquement inférieurs à l'ensemble de ces termes précédemment rattachés manuellement, sont ainsi annexés implicitement au métaterme.

Le regroupement de l'ensemble de ces termes au niveau du métaterme « Médicaments » permet d'élargir le champ de recherche de la requête des utilisateurs, dans la mesure où nous aurons tous les documents indexés par tous ces concepts relatifs au médicament. La création du métaterme « Médicaments » a permis de regrouper plus de 14.000 ressources[80].

Exemple d'expansion de requête concernant le métaterme « Médicaments » :

Requête : médicaments.mt
Reformulation de la requête : traitement médicamenteux.**mc** *ou* coût médicament.**mc** *ou* médicament orphelin.**mc** *ou* pharmacologie.**mc** *ou* toxicité des médicaments.**mc** *ou* utilisation médicament.**mc** *ou* phénomènes chimiques et pharmacologiques.**mc** *ou* préparations pharmaceutiques.**mc** *ou* évaluation préclinique médicament.**mc** *ou* voies d'administrations des médicaments.**mc** *ou* actions pharmacologiques.**mc** *ou* malformations dues aux médicaments et aux drogues.**mc** *ou* agrément de médicament.**mc** *ou* évaluation médicament.**mc** *ou* hypersensibilité médicamenteuse.**mc** *ou* produits biopharmaceutiques.**mc** *ou* système distribution médicaments.**mc** *ou* technologie pharmaceutique.**mc** *ou* contamination de médicaments.**mc** *ou* rythme administration médicament.**mc** *ou* surveillance médicament.**mc** *ou* stents à élution de médicament.**mc** *ou* biomarqueurs pharmacologiques.**mc** *ou* information sur le médicament.**tr** *ou* action des médicaments et substances chimiques.**qu** *ou* traitement médicamenteux.**qu** *ou* administration et posologie.**qu** *ou* pharmacocinétique.**qu**

***mc** : code booléen correspondant à un descripteur MeSH; **qu** : code booléen correspondant à un qualificatif ; **tr** : code booléen correspondant à un type de ressource ; les caractères gras pour mettre en relief les codes booléens ; le caractère italique (ou) pour signaler l'opérateur booléen permettant l'expansion de requête.*

[78] Se référer au Chapitre 3 ; Section 3.1.2.4 pour plus d'information sur la notion *métaterme.*
[79] Se référer à l'Annexe B pour la liste exhaustive des descripteurs.
[80] Statistiques datant de l'année 2008, aujourd'hui (2010) à peu près 25.000 ressources.

- Étape 2

La deuxième étape a été de créer le site portail et d'y associer des formulaires de recherche simple et avancée bilingues.

Respectant la définition et les caractéristiques d'un portail, le PIM contient un moteur de recherche qui est inspiré largement du celui de CISMeF « Doc'CISMeF », mais avec quelques spécificités centrées sur le médicament. L'outil de recherche de PIM contient une recherche simple et une recherche avancée, les deux sous forme bilingue (français et anglais). Le choix d'avoir ces deux modes de recherche s'est fondé sur le fait de vouloir avoir un portail quasi international, d'une part, et de s'adapter à certaines spécificités des codes, notamment les codes ATC, d'autre part. En effet, dans certains cas, nous pouvons avoir des codes ATC variant d'un pays à l'autre pour un même médicament[81].

La recherche simple peut se faire sur le nom commercial ou la DCI, ou sur n'importe quel code relatif aux médicaments et aux substances chimiques (code ATC, code CAS, code CIP, code CIS…) ou encore sur un terme MeSH.

La recherche avancée, quant à elle, permet une recherche spécifique grâce à une combinaison de ces codes : nous pouvons affiner notre champ de recherche en spécifiant à la fois, par exemple la Dénomination Commune Internationale (DCI) et l'action pharmacologique.

- Étape 3

La troisième étape permet la mise en place des liens contextuels vers des banques de données médicamenteuses anglophones, en particulier Drug Information Portal de la NLM[82] et, Entrez, outil de recherche du NCBI (National Center for Biotechnology Information) dans les sciences de la santé qui englobe, notamment, *PubMed*[83] et *PubChem Substance* [84].

- Étape 4

La quatrième étape s'est achevée par la mise au point d'un « Google sélection PIM » permettant d'effectuer une recherche Google limitée à une sélection de sites éditeurs de qualité concernant les médicaments, déjà recensés par les documentalistes de l'équipe CISMeF. Nous avons utilisé « Google TM Custom Search Engine » (Google CSE), en utilisant la plateforme « Google Co-opTM »[85].

Du fait que le moteur de recherche de Google récupère au moins toutes les pages statiques d'un site, le corpus de « Google sélection PIM » inclut toutes les ressources de PIM, mais aussi d'autres pages qui n'ont pas été sélectionnées manuellement par l'équipe CISMeF.

---

[81] Se référer au Chapitre 3; Section 3.1.2.1 pour plus de détails

[82] US NLM Drug Information Portal. URL: http://druginfo.nlm.nih.gov/drugportal/drugportal.jsp

[83] PubMed est un service de la Bibliothèque nationale de la médicine des USA qui inclut plus de 19 millions de citations de MEDLINE et d'autres journaux des sciences de la vie. URL : http://www.ncbi.nlm.nih.gov/pubmed/

[84] Entrez, The Life Sciences Search Engine. URL: http://www.ncbi.nlm.nih.gov/sites/gquery

[85] Google Coop. URL: http://www.google.com/coop/

Ainsi, grâce à ces quatre étapes décrites ci-dessus, nous avons mis en place le Portail d'Information sur les Médicaments permettant de recenser et d'organiser les ressources web de qualité dédiées aux médicaments, afin d'en faciliter l'accès.

Figure 4.1.2.1. Page d'accueil du Portail d'Information sur les Médicaments

Les résultats d'une requête dans le PIM sont présentés sous forme de notices descriptives, inspirées du catalogue CISMeF[86]. Au sein de chaque notice, un champ nommé « substance », (équivalent au « Substance Name » de PubMed), a été mis en place permettant de regrouper et de repérer les substances impliquées dans l'indexation des ressources, ainsi que leurs actions pharmacologiques contextuelles.

Ce portail est, actuellement, d'accès restreint (identification = CISMeF ; mot de passe=demoweb) jusqu'à la fin du projet PSIP (juin 2011). Ensuite, plusieurs scénarios sont possibles :

(a) accès libre sur le portail CISMeF ;
(b) commercialisation par la société VIDAL.

4.2 CONCEPTION DE L'APPROCHE DE L'INDEXATION AUTOMATIQUE PAR LA CLASSIFICATION ATC

S'intéressant du plus près au domaine médicamenteux (Sakji et al., 2009b) et se souciant d'améliorer la recherche d'information concernant les médicaments, nous avons conçu une approche permettant l'indexation des ressources CISMeF avec la classification ATC. Nous avons appliqué cette indexation sur le corpus restreint aux médicaments de CISMeF, à savoir celui du PIM.

[86] Se référer au Chapitre 1; Section 1.1.2 pour plus de détails concernant les notices descriptives.

Le choix de la classification ATC pour l'indexation est justifié par le fait qu'elle représente le système le plus utilisé en France et en Europe pour classer les médicaments et, par le fait qu'elle soit contrôlée et actualisée sous la responsabilité de l'OMS (http://www.whocc.no/atcddd).

Par ailleurs, grâce à l'indexation par la classification ATC, nous pouvons pallier le manque du thésaurus MeSH en fournissant une information complémentaire aux utilisateurs, en terme d'indexation des ressources, d'une part et concernant la substance chimique elle-même, d'autre part. À travers la figure 4.2.1, nous pouvons remarquer que la ressource restituée est indexée par le descripteur MeSH « *acide acétylsalicylique* », ayant comme action pharmacologique « *anti-inflammatoires non stéroïdiens* ». En plus, moyennant la classification ATC, nous avons une information complémentaire concernant l'indexation et la substance chimique « *acide acétylsalicylique* » dans la mesure où cette dernière appartient au groupe thérapeutique « *analgésiques* » et au groupe pharmacologique « *autres analgésiques et antipyrétiques* » et qu'elle agit sur le système nerveux. En effet, l'acide acétylsalicylique peut avoir plusieurs effets thérapeutiques et agit sur différents organes anatomiques. Par exemple, il agit sur le système nerveux lorsqu'il a un effet thérapeutique analgésique ou encore sur les voies digestives et métabolisme lorsqu'il a un effet thérapeutique des préparations stomatologiques. De ce fait, la substance chimique peut avoir plusieurs codes ATC.

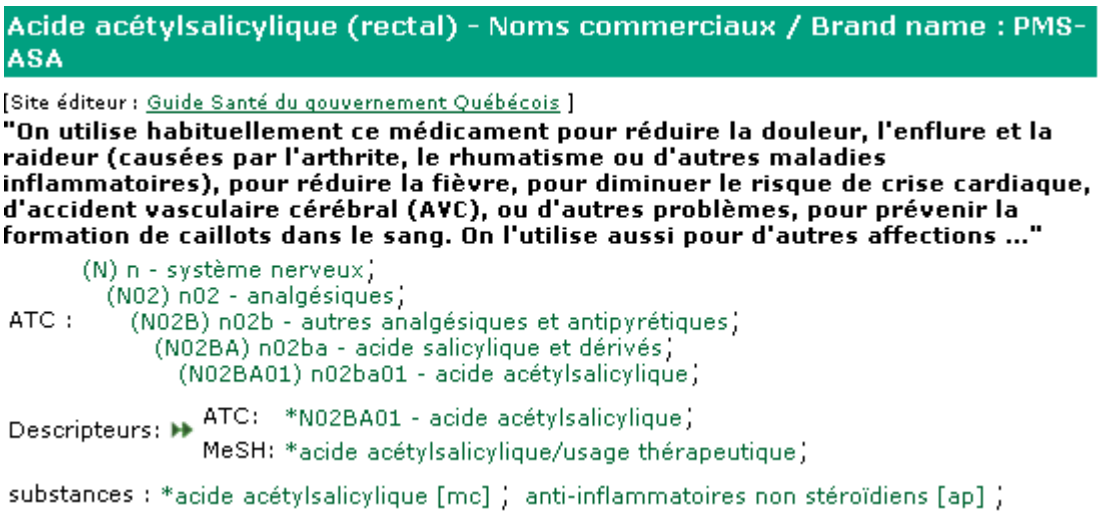


Figure 4.2.1. Indexation bi-terminologique (thésaurus MeSH et classification ATC) d'une ressource : des informations complémentaires concernant les substances chimiques

Afin de fournir plus de connaissances concernant le système de la classification ATC, d'une part, et de contextualiser autant que possible l'information sur les médicaments, nous avons choisi d'afficher la hiérarchie complète de la substance chimique qui indexe les ressources (cf. Figure 4.2.2). Ainsi, d'un point de vue pédagogique, les utilisateurs et en particulier les étudiants en médecine peuvent obtenir des informations plus exhaustives sur les médicaments, leurs caractéristiques thérapeutiques et chimiques et les organes ou les systèmes sur lesquels ils agissent.

Dans le même contexte et pour faciliter l'accès aux connaissances sur le médicament, des travaux similaires ont été réalisés par (Lamy et al., 2009) afin de mieux détecter les contre-indications et les effets indésirables des médicaments, ainsi, que les interactions médicamenteuses par les professionnels de santé. Pour ce faire, les auteurs ont conçu une interface graphique s'appuyant sur un langage iconique et repose sur des techniques de visualisation d'information.

Figure 4.2.2. Résultat de la recherche d'information dans le PIM mettant en relief les différents champs permettant de décrire une ressource ainsi que la hiérarchie de la classification ATC

4.2.1 PRINCIPE DE FONCTIONNEMENT : TROIS ETAPES SEQUENTIELLES

L'approche de l'indexation automatique par la classification ATC, étant appliquée au PIM, peut être résumée en trois étapes séquentielles (voir la section 4.2.1.2) :

- ✓ méthode par titre : la recherche du code ATC au niveau du titre de la ressource ;
- ✓ méthode par nom commercial : la recherche du nom commercial (NC) de la substance au niveau du titre de la ressource. Ensuite, l'attribution du code ATC correspondant ;
- ✓ méthode par indexation : la recherche du code ATC selon l'indexation de la ressource (indexation par les descripteurs et/ou les concepts chimiques supplémentaires du thésaurus MeSH).

4.2.1.1. LA MISE AU POINT DES PRETRAITEMENTS

Pour mettre au point cette stratégie, des prétraitements ont été réalisés. En effet, les libellés du cinquième niveau de la classification ATC ont été automatiquement appariés avec les descripteurs MeSH, d'une part, et les concepts chimiques supplémentaires (CCSs) d'autre part. Ce prétraitement permet de mettre en corrélation le système de la classification ATC et le thésaurus MeSH.

Pour ce faire, nous avons procédé à un traitement automatique basé essentiellement sur les techniques de traitement de langage naturel (TAL). Le TAL est une discipline qui a été développée depuis plusieurs années et classée depuis les années 60 comme un domaine de l'intelligence artificielle et de la linguistique dans le but de mieux cerner les problèmes de la compréhension du langage naturel (Vallez et al., 2007). Suite à ce traitement réalisé entre les termes de la classification ATC et les concepts qui représentent les substances chimiques du MeSH, les seuls résultats obtenus correspondaient à une correspondance exacte entre les termes

Une évaluation de cette étape a été nécessaire afin de valider le traitement automatique d'appariement. La documentaliste-pharmacienne de l'équipe CISMeF, notre expert humain (considérée comme notre gold standard pour cette approche), a validé cet appariement ce qui nous a permis de le mettre au point, le compléter ou corriger des erreurs non distinguables automatiquement.

Cette validation nous a permis de déceler quelques anomalies du processus. Les modifications majeures qui ont été implémentées concernaient essentiellement les « associations médicamenteuses » et les substances chimiques inorganiques comme les dérivés du « Potassium », par exemple. En effet, le traitement automatique

- n'associait, dans la plupart des cas, les termes ATC contenant le mot « *potassium* » qu'au descripteur MeSH « *potassium* ».

  Le complément de cet appariement était d'ajouter le descripteur MeSH « *composés du potassium* » à ce type d'association.

- ne prenait pas en compte les associations médicamenteuses dans la mesure où ces dernières ne sont pas facilement détectables.

  Pour pallier ce manque, il était nécessaire d'ajouter le descripteur MeSH « *association médicamenteuse* » si le terme *en association* était présent dans le libellé ATC, ou si le libellé ATC représente une association de plusieurs substances chimiques. Ce dernier cas a posé un problème puisque les libellés des associations de substances chimiques n'ont pas une représentation standard (par exemple : « *fludrocortisone et antiinfectueux* »; « *sulfate ferreux-glycine* »; « *fer, vitamine B12 et acide folique* »). Par exemple, avec la première version du traitement automatique d'appariement, nous n'avons pu relier le terme ATC « *ENALAPRIL ET DIURETIQUES* » qu'avec les descripteurs MeSH « *diurétiques* » et « *énalapril* ».

- ne permettait de réaliser, d'une façon générale, qu'un appariement syntaxique. Or dans certains cas, ce type de correspondance peut s'avérer faux.

  En effet, un appariement entre un code ATC et un concept chimique supplémentaire (CCS) est favorisé à un appariement entre un code ATC et un descripteur MeSH puisqu'il est considéré plus précis. Comme exemple d'ambiguïté, nous pouvons citer l'association (1..n) du code ATC *J07CA02* ayant le libellé « *Diphtérie-coqueluche-poliomyélite-tétanos* » avec les descripteurs MeSH « *coqueluche* », « *diphtérie* », « *poliomyélite* » *et* « *tétanos* » au lieu du CCS *« vaccin DTCP »*, un vaccin contre la diphtérie, la tétanos, la poliomyélite et le coqueluche.

  D'un point de vue syntaxique et traitement du langage naturel, un tel appariement est parfaitement correct. Cependant, d'un point de vue sémantique et dans un cadre spécifique au domaine médical et médicamenteux, une telle association conduit à une erreur.

La validation et les corrections d'erreurs et/ou d'ambiguïtés ont été réalisées par la documentaliste-pharmacienne de l'équipe CISMeF.

De plus, nous avons étendu le traitement automatique d'appariement en ajoutant le principe de l'explosion[87] notamment pour les descripteurs MeSH qui correspondent aux actions pharmacologiques. Cette amélioration est faite grâce à l'expertise de notre pharmacienne. L'automatisation d'une telle procédure semble être impossible du fait que nous n'appliquons pas l'explosion à tous les descripteurs, ce qui justifie le travail fastidieux de cette étape. Par exemple, le terme ATC « *dexaméthasone et anti-infectieux* » ayant le code S01CA01 doit être apparié, soit avec les descripteurs MeSH « *anti-infectieux* », « *association médicamenteuse* » et « *dexaméthasone* », soit avec les descripteurs MeSH « *association médicamenteuse* », « *dexaméthasone* » et tous les descripteurs MeSH qui subsument « *anti-infectieux* ». Par conséquent, grâce à ce principe, le terme ATC « *dexaméthasone et anti-infectieux* » sera aussi apparié avec les descripteurs MeSH « *association médicamenteuse* », « *dexaméthasone* » et « *anti-infectieux urinaires* » puisque ce dernier est hiérarchiquement inférieur à « *anti-infectieux* » (cf. Figure 4.2.1.1).

[87] L'explosion des descripteurs se traduit par la recherche de tous les termes qui subsument [*[philo.] Fait de considérer une chose comme faisant partie d'un tout*] le descripteur le plus haut de la hiérarchie.

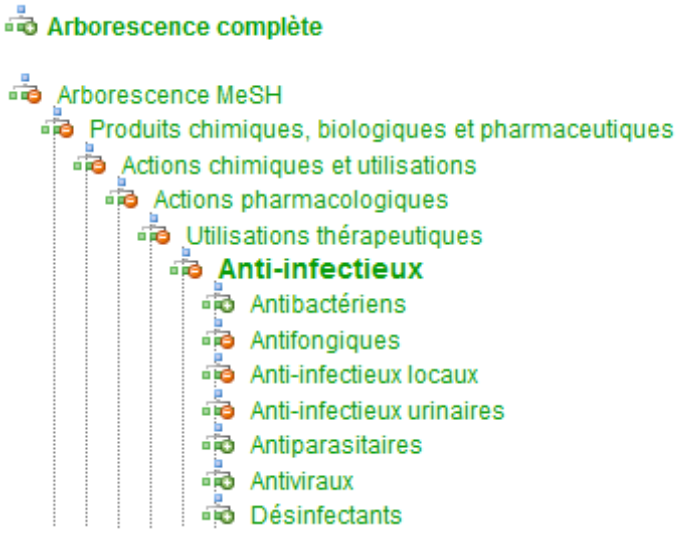


Figure 4.2.1.1. Arborescence MeSH du descripteur « Anti-infectieux »

4.2.1.2. CONCEPTION DE L'APPROCHE

L'algorithme se résume en trois étapes séquentielles :

1ère étape : détection automatique du code ATC du cinquième niveau (à 7 caractères) au niveau du titre de la ressource. Si c'est le cas, la ressource est indexée avec ce dernier ;

Autrement, 2ème étape : détection automatique du nom commercial du médicament au niveau du titre de la ressource. Si c'est le cas, le code ATC associé au nom commercial est assigné à la ressource.

Pour ce faire, nous disposions d'une table reliant le nom commercial des médicaments et le code ATC. Elle est partiellement fournie par le système d'information de l'hôpital de Rouen, complétée par des données en provenance du Vidal.

À cette étape, une amélioration a dû être réalisée puisque certains noms commerciaux sont reliés à plusieurs codes ATC, ce qui a entravé l'attribution du code ATC le plus adéquat à la ressource. Ces noms commerciaux correspondent généralement aux médicaments génériques et peuvent se différencier par leur forme galénique. Exemple : DICLOFENAC TEVA 1 % gel a pour code ATC M02AA15 et le DICLOFENAC TEVA 25 mg cp enr gastrorésis a pour code ATC M01AB05 ;

Sinon, 3ème étape : indexation automatique par le code ATC grâce à l'indexation MeSH (descripteurs et/ou CCSs) de la ressource. Un code ATC est attribué à la ressource s'il est en correspondance avec les termes d'indexation MeSH de cette dernière.

Cette étape se base sur le prétraitement réalisé précédemment mettant en relation les descripteurs MeSH et les CCSs avec les termes ATC. Cette étape est largement perfectionnée par les règles de post-coordination (cf. section 4.2.1.3).

En revanche et étant donné qu'une ressource peut être indexée par plusieurs descripteurs et/ou CCSs MeSH, plusieurs codes ATC peuvent être candidats. Pour résoudre ce problème, un calcul de score, basé sur la fréquence de chaque code ATC en liaison avec l'indexation,

permet de départager ces candidats. En effet, plus un code ATC admet une pondération élevée, plus il est susceptible d'être le bon candidat pour l'indexation ATC.

4.2.1.3. REGLES DE POST COORDINATION

Afin d'affiner le résultat de la méthode de l'indexation automatique, nous avons eu recours à des règles de post-coordination. En effet,

- pour améliorer l'indexation automatique des associations des médicaments par la classification ATC, la pharmacienne nous a recommandé des mesures à respecter qui nous ont permis de construire la règle suivante : une ressource est indexée par un code ATC si, et seulement si, tous les termes MeSH (descripteurs et CCSs) associés au terme ATC correspondant à ce code, sont également des termes d'indexation de la ressource. Par exemple, une ressource devrait être indexée par le code ATC S01CA01 si, et seulement si, la ressource est indexée par tous les descripteurs MeSH « *association médicamenteuses* », « *dexaméthasone* » et « *anti-infectieux* ».
- comme nous l'avons mentionné auparavant, pour chaque concept chimique supplémentaire (CCS), le MeSH recommande une projection vers des descripteurs MeSH. Pour notre approche, l'application de cette règle est occultée dans la mesure où la ressource ne devrait être indexée que par le code ATC qui est apparié avec le CCS, sans prendre en compte la projection vers le(s) descripteur(s) MeSH. Par exemple, le MeSH recommande d'utiliser le descripteur « *aciclovir* » pour le CCS « *Valaciclovir* ». Cependant, si une ressource est indexée avec le CCS « *Valaciclovir* », elle devrait être indexée seulement par le terme ATC « *Valaciclovir* » ayant comme code J05AB11 et non pas aussi avec les codes ATC D06BB03, J05AB01, J05AB01 et S01AD03 correspondants à l' « *aciclovir* ». Toutefois, cette règle ne s'applique pas aux termes ATC qui sont des associations de substances.

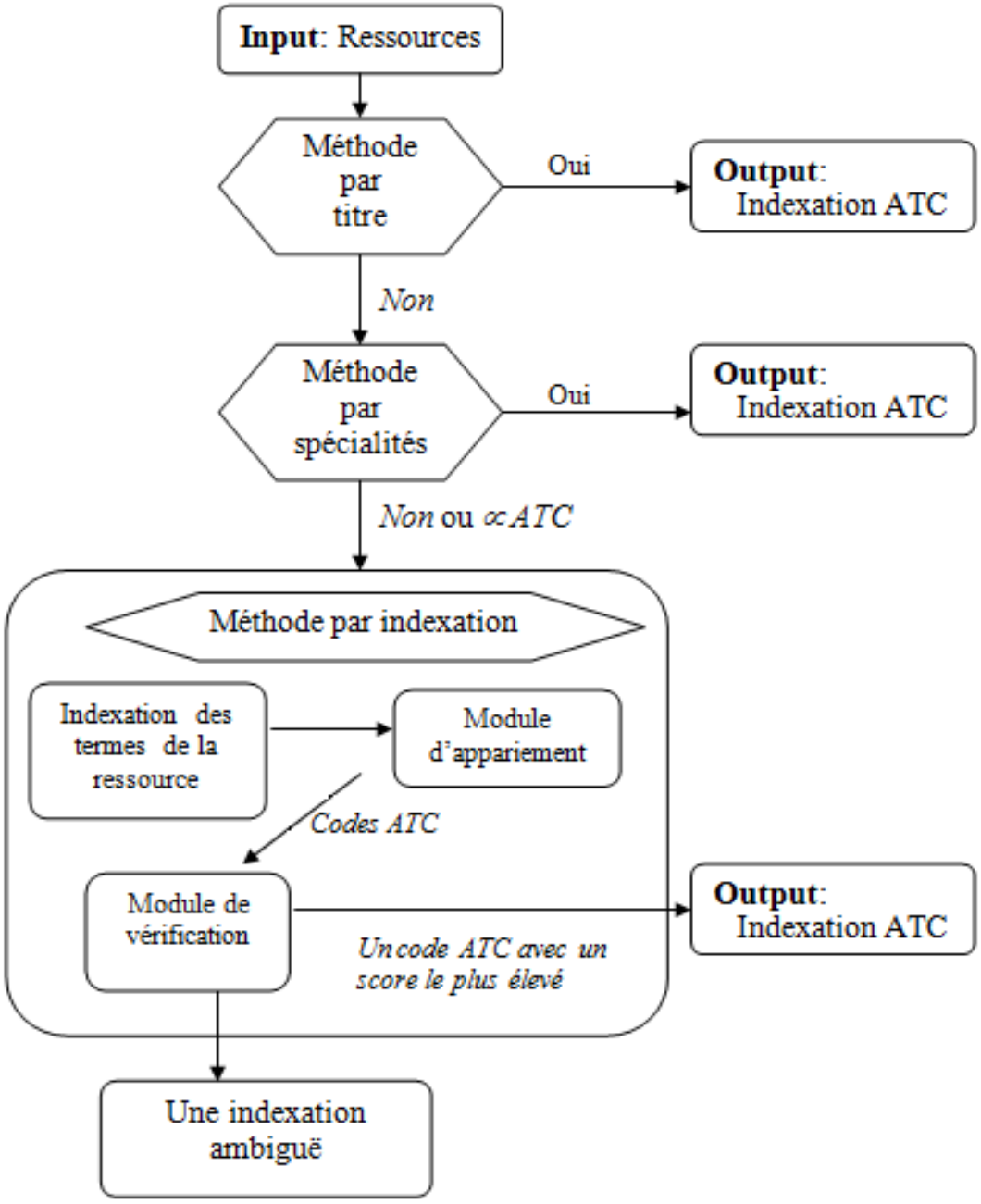


Figure 4.2. Résumé de l'approche de l'indexation automatique par la classification ATC

4.2.1.4. LE CORPUS D'APPLICATION

Il faut noter qu'au moment de la réalisation de cette approche d'indexation automatique par la classification ATC, nous étions encore dans un monde mono-terminologique dans la mesure où les ressources n'étaient indexées (manuellement ou automatiquement) que par le thésaurus MeSH. Dès lors, cette méthode permettait de compléter l'indexation pour être bi-terminologique.

L'approche de l'indexation automatique par la classification ATC a été réalisée sur le corpus du Portail d'Information sur les Médicaments (PIM) constitué, alors, de 10.250 ressources :

5.177 ressources sont manuellement indexées à l'aide du thésaurus MeSH et 5.073 l'étaient automatiquement.

En fait, au moment de l'implémentation de notre approche, nous avons eu le choix de l'appliquer, soit sur le corpus du catalogue CISMeF, soit sur celui du PIM. Cependant, se focalisant sur les médicaments, il était plus judicieux de se concentrer sur le deuxième vu qu'il a été conçu à cet effet.

4.2.1.5 IMPLEMENTATION DE L'APPROCHE

Étant donné que nous disposons d'un système de gestion de base de données (SGBD) Oracle (actuellement en version 11g) et que l'équipe CISMeF utilise entre autres le PL/SQL (Procedural Language/SQL) pour les procédures stockées[88], nous avons choisi d'implanter notre approche avec le même langage.

PL/SQL est un langage procédural de quatrième génération d'Oracle corporation étendant SQL. Il permet de combiner les avantages d'un langage de programmation classique, avec les possibilités de manipulation de données offertes par SQL. Parmi ses avantages, nous pouvons noter les instructions procédurales et la gestion des erreurs. Le langage PL/SQL intègre parfaitement le langage SQL en lui apportant une dimension procédurale. Certes, SQL permet d'exprimer des requêtes dans un langage relativement simple, mais il n'intègre aucune structure de contrôle permettant, par exemple, d'exécuter une boucle itérative. PL/SQL autorise la manipulation complexe des données contenues dans une base Oracle en transmettant un bloc de programmation au SGBD au lieu d'envoyer une requête SQL. De cette façon, les traitements sont directement réalisés par le système de gestion de bases de données. Cela a pour effet, notamment, de réduire le nombre d'échanges à travers le réseau et donc d'optimiser les performances des applications. Les structures de PL/SQL sont similaires à celles des langages évolués et fournissent une méthode souple, pour manipuler l'information d'une base de données.

Le langage PL/SQL définit, aussi, en standard un grand nombre d'exceptions (ou d'erreurs), il offre un moyen de les identifier et de les traiter à l'aide du mécanisme des exceptions. De plus, l'utilisateur peut définir ses propres exceptions, ce qui offre de nombreuses possibilités.

4.2.2 RESULTAT : ÉVALUATION DE L'APPROCHE

Pour mesurer le degré de pertinence et l'apport de notre approche, nous avons été amenés à effectuer une double évaluation : la première se porte sur l'appariement entre la classification ATC et le thésaurus MeSH, puisque ce prétraitement intervient dans le processus d'indexation et peut, par conséquent, influencer le résultat. La deuxième évaluation est faite sur le résultat de la méthode elle-même : la pertinence ou non de l'indexation des ressources par la classification ATC.

[88] Une procédure stockée (ou *stored procedure* en anglais) est un ensemble d'instructions SQL pré-compilées, stockées sur le serveur, directement dans la base de données. Elles peuvent être exécutées sur demande : lancées par un utilisateur, un administrateur DBA ou encore de façon automatisée par un événement déclencheur.

4.2.2.1 EVALUATION DE L'APPARIEMENT DU PRETRAITEMENT

Pendant le module d'appariement, la correspondance entre les termes ATC et les termes MeSH (descripteurs et CCSs) nous a permis de réaliser la troisième étape de l'approche (méthode par indexation).

Au cours du processus de l'appariement, nous n'avons pas réussi à avoir une correspondance parfaite de tous les termes du cinquième niveau de la classification ATC (correspondants aux substances chimiques). Les cas de non-correspondance détectés sont principalement dus au fait que :

- ✓ le thésaurus MeSH ne couvre pas forcément toutes les substances chimiques. De ce fait, certains termes de la classification ATC n'ont pas de correspondance avec des termes du thésaurus MeSH ;
- ✓ au moment de la réalisation de cette approche, certains descripteurs ou CCSs MeSH n'étaient pas été encore créés (passage de la version 2008 à celle de 2009) ;
- ✓ plusieurs concepts chimiques supplémentaires (CCSs) n'étaient pas encore traduits en français.

A part ces quelques lacunes, la performance du module d'appariement entre la classification ATC et le thésaurus MeSH (descripteurs et CCSs), en termes de précision et rappel, est jugée bonne avec 90% de précision et 87% de rappel.

4.2.2.2 EVALUATION DES RESULTATS DE L'APPROCHE D'INDEXATION

Pour l'évaluation de notre approche, nous avons mesuré, en premier temps, le nombre de ressources (à partir du corpus de PIM) qui ont pu être indexées par la classification ATC. Ensuite, nous avons la qualité des ressources en terme de besoin informationnel.

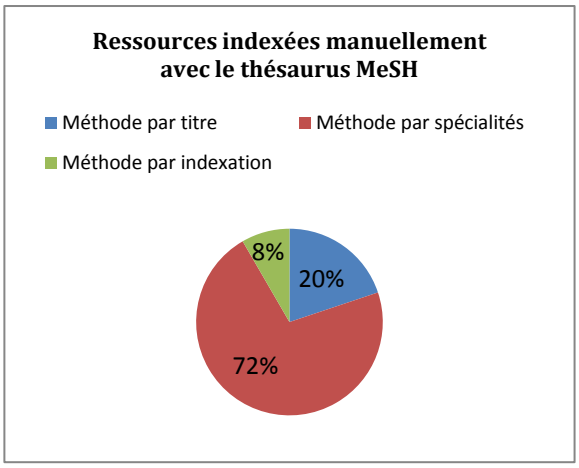
Sur les 5.177 ressources indexées manuellement par le thésaurus MeSH, 3.634 ressources ont été indexées automatiquement par la classification ATC. Sur ces 3.634 ressources, nous avons 2.608 ressources indexées grâce à la méthode par spécialités ; soit 71,76%, 722 indexées par la méthode par titre ; soit 19,86% et finalement 304 ressources indexées par la méthode par indexation ; soit 8,26%.

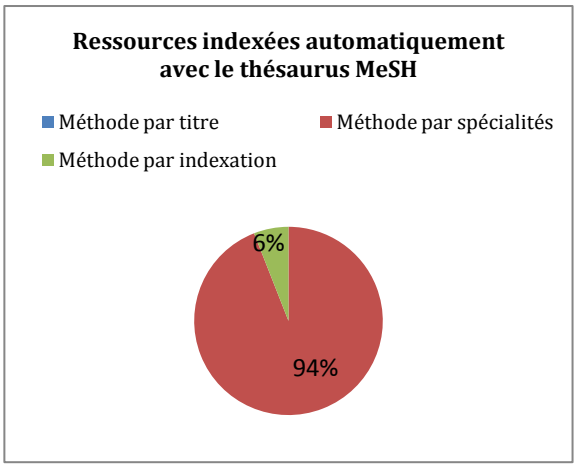
Sur les 5.073 ressources indexées automatiquement par le thésaurus MeSH, 1.341 ont été indexées automatiquement par la classification ATC. Sur ces 1.341 ressources, nous avons 1.261 ressources indexées par la méthode par spécialités ; soit 94,03%, suivi par la méthode par indexation (5,96%) avec 80 ressources et ensuite la méthode par titre.

	Ressources indexées manuellement avec le thésaurus MeSH	**Ressources indexées automatiquement avec le thésaurus MeSH**
Méthode par titre	722 (19,86%)	0 (0%)
Méthode par spécialités	2.608 (71,76%)	1.261 (94,03%)
Méthode par indexation	304 (8,36%)	80 (5,96%)

Total	*3.634 sur 5.177 (70,2%)*	*1.341 sur 5.073 (26,4%)*

Tableau 4.2.2.2.1. Résultat de l'indexation automatique par la classification ATC selon les trois méthodes de l'approche





L'indexation automatique par la classification ATC est principalement effectuée par la méthode par spécialités. Cela est dû au fait que l'indexation automatique par l'ATC est appliquée sur le corpus du PIM. Au sein de ce dernier, les types de ressources sont relatifs à l'« information sur les médicaments » et les titres des ressources mentionnent souvent les noms commerciaux des médicaments.

Le résultat « nul » de l'indexation automatique par la classification ATC, par la méthode par titre, est dû à l'absence des codes ATC dans les titres des ressources indexées automatiquement.

Le bon score, pour les ressources indexées manuellement par le thésaurus MeSH concernant la méthode par titre, s'explique par l'ajout manuel par les indexeurs des codes ATC au niveau des titres des ressources.

L'évaluation du résultat a été réalisée par notre gold standard (documentaliste-pharmacienne experte du domaine) sur 200 ressources choisies aléatoirement. Ces dernières sont répertoriées en ressources manuellement et automatiquement indexées avec le thésaurus MeSH. La pertinence globale était estimée à 76%, alors que la non pertinence à seulement 20,5% (cf. Tableau 4.2.2.2.2)

L'indexation des ressources du PIM par la classification ATC est jugée selon quatre critères qui nous ont permis d'établir une évaluation qualitative :

- ✓ *pertinent* quand le code ATC est correct ;
- ✓ *non pertinent* quand le code ATC est complètement erroné ;
- ✓ *partiel* quand il y a potentiellement plusieurs codes ATC et la fonction nous retourne seulement un seul code ATC ;

- ✓ *incomplet* quand le code ATC affiché est relatif à un code de niveaux supérieurs de la substance chimique (le cinquième niveau, composé de 7 caractères), en d'autres termes, si l'indexation de la ressource est faite par un code relatif aux actions thérapeutiques/pharmacologiques de la substance chimique.

	Ressources indexées manuellement avec le thésaurus MeSH	**Ressources indexées automatiquement avec le thésaurus MeSH**	**Total**
Pertinent	91 (91%)	61 (61%)	152 (76%)
Non pertinent	5 (5%)	36 (36%)	41 (20,5%)
Partiel	3 (3%)	0 (0%)	3 (1,5%)
Incomplet	1 (1%)	3 (3%)	4 (2%)

Tableau 4.2.2.2.2. L'évaluation de l'indexation automatique par la classification ATC

4.2.3 DISCUSSION

De façon générale, pour les 200 ressources évaluées, la pertinence a été estimée à 76%, alors que la non pertinence était à 20,5%. Pour illustrer les résultats non pertinents, prenons par exemple, une ressource qui a été indexée par la « méthode par indexation » avec le code ATC *G04BE03* qui correspond au « *sildénafil* » (administré comme médicament utilisé pour les troubles de l'érection). Cependant, même si la ressource a été indexée par le terme MeSH « *sildénafil* », dans ce document, le médicament est administré pour le traitement de l'hypertension artérielle pulmonaire et, non pas pour le dysfonctionnement érectile. Ainsi l'indexation ATC devrait être avec le terme « *autres vaso-dilatateurs périphériques* » ayant le code C04AX.

Cet exemple d'erreurs illustre, bel et bien, l'intérêt d'une évaluation manuelle par un expert afin de se focaliser sur le contexte et d'améliorer l'algorithme d'indexation automatique par la classification ATC.

Les résultats partiels sont détectés lorsqu'il y a plusieurs codes ATC candidats et, en résultat, nous avons eu seulement un code ATC. Par exemple, le nom commercial *thiovalone* admet deux codes ATC à savoir le R02AA05 et le R01AD07. Une indexation partielle est illustrée quand une ressource est indexée par l'un ou l'autre.

Les résultats incomplets sont détectés quand le code ATC d'indexation n'est pas le code du cinquième niveau qui représente le principe actif du médicament. Par exemple, une ressource était indexée par « *vaccins contre les diarrhées à rotavirus* » ayant le code ATC *J07BH* au lieu de « *rotavirus, pentavalent, virus vivant* » ayant le code ATC *J07BH02*.

Parmi les autres types d'erreurs auxquelles nous avons dû faire face, nous trouvons celles liées à la représentation des codes. Par exemple, dans la figure 4.4.1, même si le code ATC est présent au niveau du titre, l'indexation de la ressource n'était faite qu'à la troisième étape de

l'algorithme « méthode par indexation ». Le code ATC n'a pas été détecté au premier niveau de l'approche due à sa mauvaise représentation.

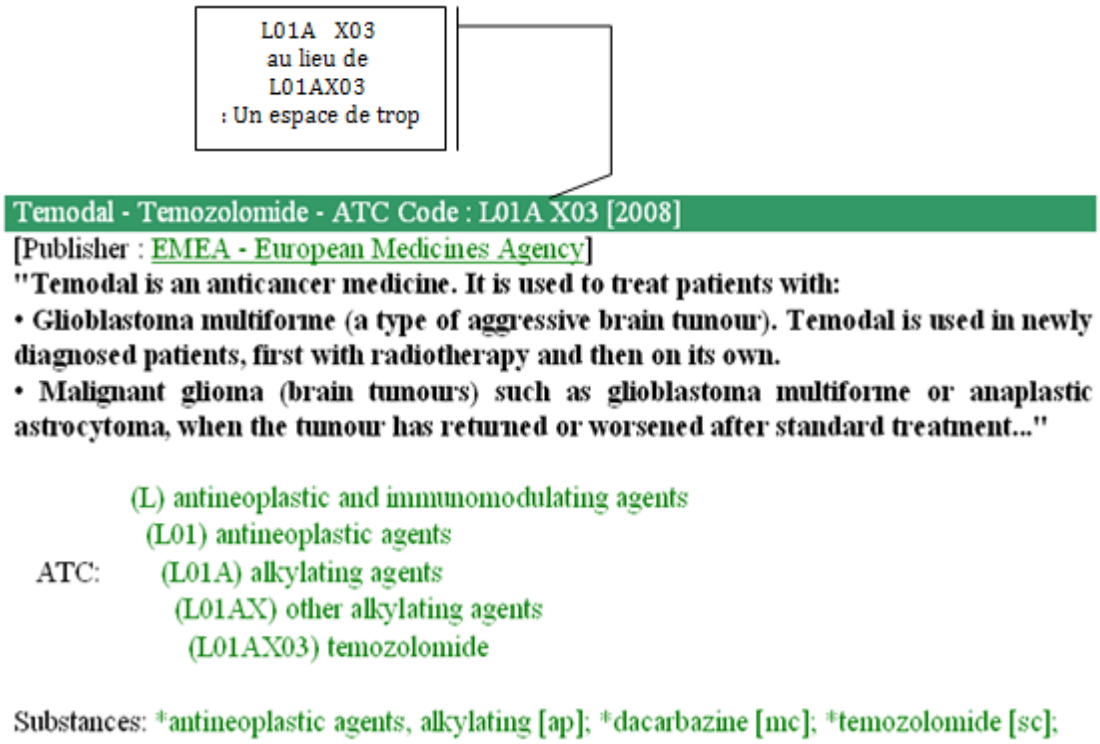


Figure 4.2.3. Résultat de l'indexation automatique par la classification ATC

4.3 AMELIORATION DE LA RECHERCHE D'INFORMATION PAR EXTENSION MESH-ATC

4.3.1 ENONCE DE L'ETUDE

Dans le but d'améliorer la recherche d'information concernant les médicaments, nous avons mis au point une approche permettant la construction de requêtes basées sur un alignement de la classification ATC avec le thésaurus MeSH. Ceci a pour but de placer les substances chimiques dans leurs contextes et, par-là même, de minimiser les erreurs qu'on a eues lors de l'évaluation de nos résultats antérieurs.

L'étude se base sur l'appariement des différents codes et libellés des différents niveaux de la classification ATC avec le thésaurus MeSH. Pour cela, nous avons complété l'appariement fait précédemment avec une mise en correspondance des libellés de premier, deuxième, troisième et quatrième niveaux de la classification ATC avec des stratégies de recherche CISMeF[89]. L'appariement est fait exclusivement d'une manière manuelle par la documentaliste-pharmacienne de l'équipe CISMeF. Le choix d'un tel appariement se base sur le fait que les stratégies de recherche sont plus concises ce qui peut contextualiser les substances chimiques (un problème déjà rencontré et expliqué à la section précédente).

[89] Se référer au Chapitre 3, Section 3.1.2.4 pour plus d'information sur les stratégies de recherche

Exemple d'appariement manuel :

Code ATC	**Libellé ATC**	**Appariement MeSH**
A	voies digestives et métabolisme	(maladie de l'appareil digestif/traitement médicamenteux.mc sauf tumeurs.mc) ou agents gastro-intestinaux.mc ou maladies métaboliques et nutritionnelles/traitement médicamenteux.mc ou maladies du système stomatognathique/traitement médicamenteux.mc
A01	préparations stomatologiques	maladies du système stomatognathique/traitement médicamenteux.mc ou préparations pharmaceutiques en odontologie.mc ou bains de bouche.mc
A01A	préparations stomatologiques	maladies du système stomatognathique/traitement médicamenteux.mc ou préparations pharmaceutiques en odontologie.mc ou bains de bouche.mc
A01AA	médicaments prophylactiques anti-caries	cariostatiques.mc ou caries dentaires/prévention et contrôle.mc
A01AB	anti-infectieux pour traitement oral local	(maladies du système stomatognathique/traitement médicamenteux.mc ou préparations pharmaceutiques en odontologie.mc ou bains de bouche.mc) et anti-infectieux.mc et administration topique.mc
A01AC	corticoïdes pour traitement oral local	((stéroïdes.mc et anti*inflammatoires.mc) ou hormones corticosurrénaliennes.mc ou glucocorticoïdes.mc ou minéralocorticoïdes.mc) et (maladies du système stomatognathique/traitement médicamenteux.mc ou préparations pharmaceutiques en odontologie.mc ou bains de bouche.mc)
A01AD	autres médicaments pour traitement oral local	((maladies du système stomatognathique/traitement médicamenteux.mc et administration topique.mc) ou préparations pharmaceutiques en odontologie.mc ou bains de bouche.mc) sauf (cariostatiques.mc ou caries dentaires/prévention et contrôle.mc ou anti-infectieux.mc ou (stéroïdes.mc et anti-inflammatoires.ap) ou hormones corticosurrénaliennes.mc ou glucocorticoïdes.ap ou minéralocorticoïdes.ap)

Exemple d'appariement automatique du 5ème niveau :

Code ATC	**Libellé ATC**	**Appariement MeSH**	**Type terme MeSH**
A01AD01	EPINEPHRINE	épinéphrine	descripteur
A01AD05	ACETYLSALICYLIQUE ACIDE	acide acétylsalicylique	descripteur
A01AD06	ADRENALONE	adrénalone	CCS
A01AD07	AMLEXANOX	amlexanox	CCS

Après une étude des correspondances mises en place (tableau ci-dessus), nous avons choisi celles des trois derniers niveaux de la classification ATC à savoir : le cinquième niveau correspondant à la substance chimique, le quatrième niveau correspondant au sous-groupe

chimique et le troisième niveau correspondant au sous-groupe pharmacologique. Nous avons exclu le premier niveau (groupe anatomique principal) et le deuxième niveau (sous-groupe thérapeutique) car ils représenteraient un appariement trop général pour notre étude.

Afin d'évaluer la valeur ajoutée d'un tel appariement pour la recherche d'information, nous avons eu recours aux différents types de requêtes, selon les différents niveaux de la classification ATC.

Pour le troisième et le quatrième niveaux, nous avons trois types de requêtes :

- requête1 : le code ATC ;
- requête2 : l'appariement MeSH du code ATC ;
- requête3 : le code ATC **OU** l'appariement MeSH (requête1 ou requête2).

Pour le cinquième niveau, nous avons ajouté à ces trois types de requêtes deux autres permettant de contextualiser la substance chimique. À ce niveau, il faut remarquer, comme nous l'avons mentionné dans le chapitre précédent, qu'ils existent des substances chimiques ayant plusieurs codes ATC selon leurs caractéristiques chimiques, thérapeutiques et selon l'organe sur lequel elles agissent (nous les avons nommés ATC multiples) et d'autres substances chimiques ayant un seul code ATC (nous les avons nommés ATC unique). Ainsi, nous avons :

- requête4 : l'appariement MeSH du libelle ATC du 5ème niveau **ET** l'appariement MeSH du libelle ATC du 1er niveau ;
- requête5 : (l'appariement MeSH du libelle ATC du 5ème niveau **ET** l'appariement MeSH du libelle ATC du 1er niveau) **OU** code ATC du 5ème niveau. (requête1 ou requête4).

Exemples de requêtes pour le 3ème niveau : *P01A - médicaments contre l'amibiase et autres protozooses*

Requête1: code ATC	P01A.ca
Requête2: appariement MeSH	antitrichomonas.mc ou antiamibiens.mc ou coccidiostatiques.mc
Requête3: requête1 ou requête2	(P01A.ca) OU (antitrichomonas.mc ou antiamibiens.mc ou coccidiostatiques.mc)

Exemples de requêtes pour le 4ème niveau : *G04BD - antispasmodiques urinaires*

Requête1: code ATC	G04BD.ca
Requête2: appariement MeSH	parasympatholytiques.mc et maladies urologiques/traitement médicamenteux.mc
Requête3: requête1 ou requête2	(G04BD.ca) OU (parasympatholytiques.mc et maladies urologiques/traitement médicamenteux.mc)

Exemples de requêtes pour le 5ème niveau (ATC multiple) : *M01AB05 - diclofénac*

Requête1: code ATC	M01AB05.ca
Requête2: appariement MeSH	diclofénac.mc
Requête3: requête1 ou requête2	(M01AB05.ca) OU (diclofénac.mc)
Requête4: appariement MeSH du libellé ATC du 5ème niveau **ET** appariement MeSH du libellé ATC du 1er niveau.	(diclofénac.mc) ET (maladies ostéomusculaires/traitement médicamenteux.mc ou antirhumatismaux.mc ou agents de maintien de la densité osseuse.mc ou agents neuromusculaires.mc)
Requête5: requête1 ou requête4	(M01AB05.ca) OU ((maladies ostéomusculaires/traitement médicamenteux.mc ou antirhumatismaux.mc ou agents de maintien de la densité osseuse.mc ou agents neuromusculaires.mc) et diclofénac.mc)

Exemples de requêtes pour le 5ème niveau (ATC unique) : *R06AD08 - oxomémazine*

Requête1: code ATC	R06AD08.ca
Requête2: appariement MeSH	oxomémazine.sc
Requête3: requête1 ou requête2	(R06AD08.ca) OU (oxomémazine.sc)
Requête4: appariement MeSH du libellé ATC du 5ème niveau **ET** appariement MeSH du libellé ATC du 1er niveau.	(oxomémazine.sc) ET (agents de l'appareil respiratoire.mc ou maladies de l'appareil respiratoire/traitement médicamenteux.mc)
Requête5: requête1 ou requête4	(R06AD08.ca) OU ((oxomémazine.sc) ET (agents de l'appareil respiratoire.mc ou maladies de l'appareil respiratoire/traitement médicamenteux.mc))

4.3.2 RESULTATS

Lors de l'évaluation des résultats de la recherche d'information grâce à ces différents types de requêtes, nous avons distingué les ressources indexées manuellement (avec le thésaurus MeSH) et celles indexées automatiquement. Cette distinction se base sur le fait que les premières sont plus pertinentes en termes d'indexation[90].

[90] Se référer au Chapitre 1 pour plus de détails

Ainsi, pour un nombre de réponses global concernant une requête, nous obtenons le nombre de réponses issues de l'indexation manuelle, le nombre de réponses issues de l'indexation automatique, ainsi que le nombre de réponses correctes respectives. Ce qui nous a conduits au calcul d'une précision moyenne.

Étant donné que l'évaluation du résultat a été faite exclusivement manuellement, nous avons dû choisir un échantillon. Ce dernier est pris au hasard et est constitué de 6 codes ATC du 4ème niveau, 6 codes ATC du 3ème niveau, 10 codes ATC uniques du 5ème niveau et 10 codes ATC multiples du 5ème niveau).

- Sur les six codes ATC du 4ème niveau, nous avons obtenu le résultat suivant:

	Précision moyenne
Requête1	0,61
Requête2	1,00
Requête3	0,73

Tableau 4.3.2.1. Précision moyenne des ressources indexées manuellement par les codes ATC du 4ème niveau

	Précision moyenne
Requête1	0,70
Requête2	1,00
Requête3	0,70

Tableau 4.3.2.2. Précision moyenne des ressources indexées automatiquement par les codes ATC du 4ème niveau

- Sur les six codes ATC du 3ème niveau, nous avons obtenu le résultat suivant :

	Précision moyenne
Requête1	0,82
Requête2	0,74
Requête3	0,72

Tableau 4.3.2.3. Précision moyenne des ressources indexées manuellement par les codes ATC du 3ème niveau

	Précision moyenne
Requête1	0,78
Requête2	0,76
Requête3	0,77

Tableau 4.3.2.4. Précision moyenne des ressources indexées automatiquement par les codes ATC du 3ème niveau

- Sur les dix codes ATC uniques du 5ème niveau, nous avons obtenu le résultat suivant:

	Précision moyenne
Requête1	1,00
Requête2	0,99
Requête3	0,99
Requête4	1,00
Requête5	1,00

Tableau 4.3.2.5. Précision moyenne des ressources indexées manuellement par les codes ATC uniques du 5ème niveau

	Précision moyenne
Requête1	1,00
Requête2	0,90
Requête3	0,98
Requête4	1,00
Requête5	1,00

Tableau 4.3.2.6. Précision moyenne des ressources indexées automatiquement par les codes ATC uniques du 5ème niveau

➢ Sur les dix codes ATC multiples du 5ème niveau, nous avons obtenu le résultat suivant:

	Précision moyenne
Requête1	0,34
Requête2	0,25
Requête3	0,25
Requête4	0,73
Requête5	0,29

Tableau 4.3.2.7. Précision moyenne des ressources indexées manuellement par les codes ATC multiples du 5ème niveau

	Précision moyenne
Requête1	0,37
Requête2	0,21
Requête3	0,24
Requête4	0,84
Requête5	0,33

Tableau 4.3.2.8. Précision moyenne des ressources indexées automatiquement par les codes ATC multiples du 5ème niveau

Pour mettre en relief un exemple de corrélation entre la précision et le rappel du système de recherche d'information, notamment celui du PIM, nous avons mesuré le rappel pour les dix requêtes de type « code ATC » (pour les codes ATC multiples du 5ème niveau).

Pour ce faire, pour chaque code ATC, la pharmacienne-documentaliste a comptabilisé le nombre de ressources (indexées manuellement et automatiquement) qui auraient dû être retrouvées ; en d'autres termes, le total de documents pertinents dans le corpus.

Les résultats obtenus sont illustrés dans les figures ci-dessous :

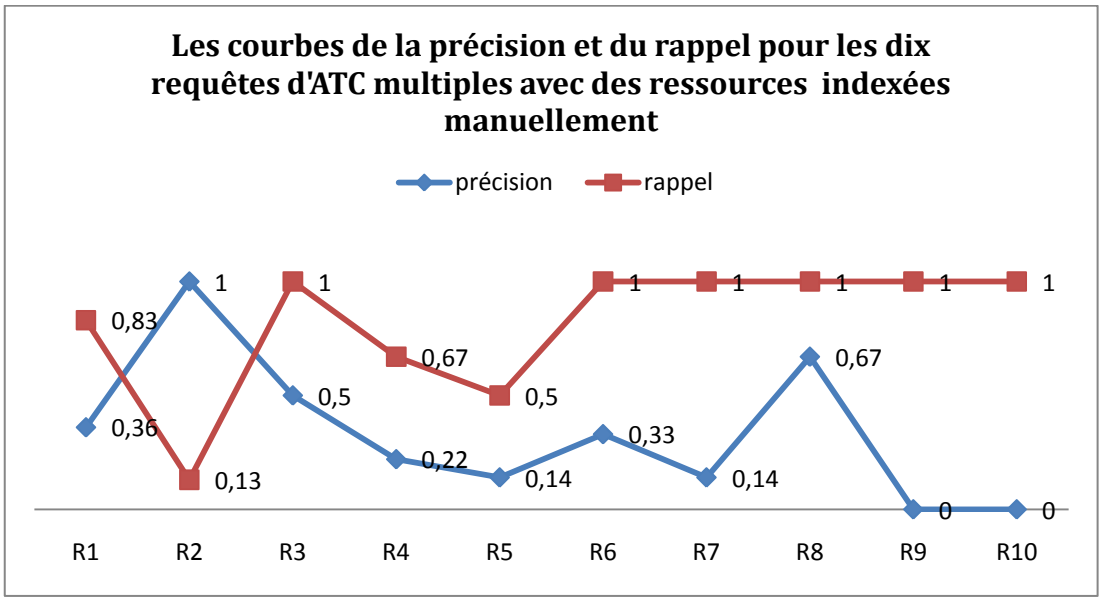


Figure 4.3.2.1. Illustration de la corrélation entre la précision et le rappel pour les requêtes ayant code ATC multiple sur un corpus indexé manuellement

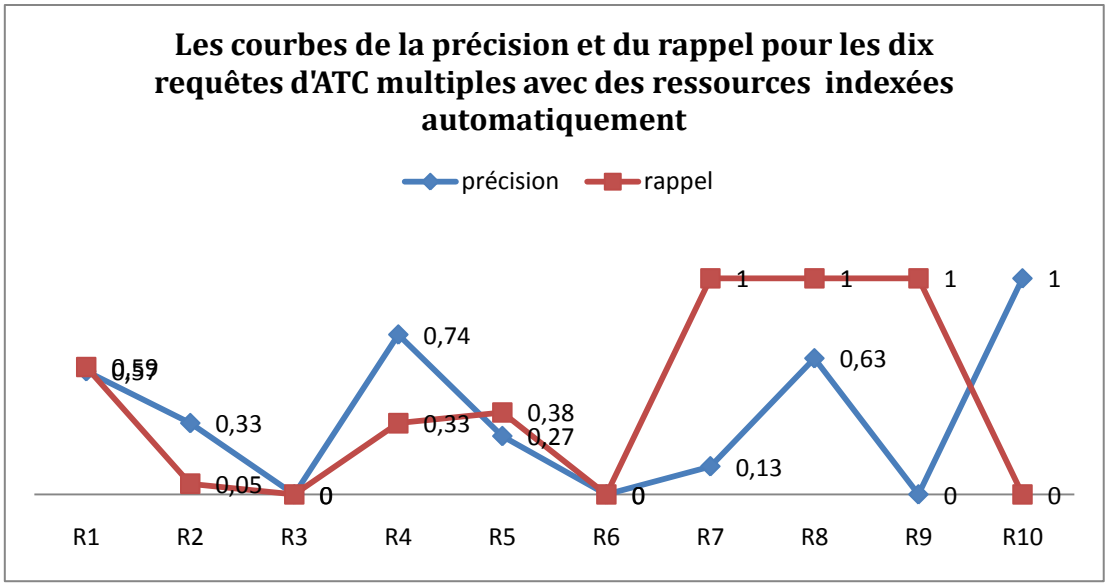


Figure 4.3.2.2. Illustration de la corrélation entre la précision et le rappel pour les requêtes ayant code ATC multiple sur un corpus indexé automatiquement

4.3.3 DISCUSSION

Les résultats, concernant les codes ATC du 4ème niveau, mettent en relief l'avantage de l'appariement de la classification ATC avec le thésaurus MeSH. Dans ce cas, une recherche d'information sur les médicaments grâce au descripteur (ou au concept chimique supplémentaire) MeSH parait plus pertinente.

Contrairement au premier cas, les résultats des codes ATC du 3ème niveau montrent qu'une recherche par code ATC pour les actions pharmacologiques des médicaments (3ème niveau de la classification ATC) donne un meilleur résultat.

Pour les codes ATC uniques, les résultats ne suggèrent pas une grande différence entre les différents modes de recherche. Ceci peut être expliqué par le fait que les substances chimiques, ayant un code ATC unique, ne représentent pas des cas de confusions lors de l'indexation et la recherche d'information.

Pour les codes ATC multiples, une jointure, entre le terme MeSH correspondant au code ATC du 5ème niveau et celui correspondant au code ATC du 1er niveau, révèle un résultat bien meilleur que les autres modes de recherche. Effectivement, une telle requête se focalise sur le contexte de la substance chimique, en d'autres termes, sur quel organe elle agit. Ce cas, est très intéressant pour notre approche d'indexation automatique des ressources du PIM par la classification ATC, dans la mesure où nous pouvons nous inspirer d'un tel appariement pour résoudre l'ambiguïté engendrée par les codes ATC multiples.

Les figures résumant les courbes de la précision et du rappel pour les codes ATC multiples en utilisant les requêtes de type « code ATC » montrent bien que, d'une manière générale pour les ressources indexées manuellement, nous avons un bon rappel mais cela reste au détriment de la précision. Néanmoins, cette différence est moins importante pour les ressources indexées automatiquement.

CONCLUSION

Le Portail d'Information sur les Médicaments (PIM), conçu et mis en œuvre pendant cette thèse, respecte la définition et les caractéristiques de base d'un portail informatique qui peut être défini comme étant « un site Web qui catalogue les principales ressources disponibles pour un domaine particulier, qui comporte généralement un moteur de recherche et offre des services thématiques et personnalisés ».

La construction de ce portail nous a permis, par la suite, d'effectuer nos travaux de recherche sur les médicaments, afin d'améliorer leur exploitation.

A notre connaissance, l'indexation par la classification ATC est une réalisation innovante appliquée à un site web sur les médicaments. Notre approche nous a permis une meilleure indexation des ressources du PIM par une terminologie autre que le thésaurus MeSH, ce qui lui a donné son aspect multi-terminologique (bi-terminologie).

Les résultats prometteurs de l'étude sur l'extension de la classification ATC par le thésaurus MeSH nous laissent optimistes pour consolider notre approche et améliorer l'indexation des ressources par la classification ATC notamment dans le PIM.

CHAPITRE 5
RECHERCHE D'INFORMATION MULTI-TERMINOLOGIQUE APPLIQUEE AU DOMAINE MEDICAL

INTRODUCTION

Ce chapitre décrit notre algorithme de recherche d'information multi-terminologique. Nous présentons, tout d'abord, une panoplie de travaux et de systèmes de recherche d'information du domaine de la santé basés sur l'expansion de requêtes et de la sémantique. Nous détaillons par la suite les algorithmes de recherche d'information mono-terminologique et multi-terminologique, appliqués au sein du catalogue CISMeF. Ensuite, nous exposons l'évaluation qui a été faite, afin de mettre en relief la valeur ajoutée de notre approche. Enfin, nous décrivons le classement du résultat de la recherche d'information.

5.1 LA RECHERCHE D'INFORMATION DE L'EQUIPE CISMEF

5.1.1 ETUDE DE L'EXISTANT

S'appliquant, particulièrement au domaine de la santé, plusieurs approches et systèmes d'information et de recherche ont été mis en place permettant d'améliorer et/ou d'assister les utilisateurs au moment de la recherche d'information. (McCray et al., 2004) ont développé un moteur de recherche qui utilise un serveur terminologique. Les requêtes des utilisateurs sont

analysées et étendues avec des variantes orthographiques et des synonymes, et des suggestions sont offertes à l'utilisateur pour modifier sa requête.

Le système HIQuA (Zeng et al., 2006) propose des termes aux utilisateurs pour affiner leurs requêtes. Pour cela, les requêtes sont appariées avec les concepts de l'UMLS, puis, grâce aux relations sémantiques existantes dans l'UMLS et aux cooccurrences entre concepts dans la littérature médicale, les concepts les plus proches sont identifiés. L'inconvénient de ce système est que l'appariement entre les concepts des utilisateurs et les concepts d'UMLS n'était pas toujours faisable ce qui a limité la performance du processus d'enrichissement des requêtes.

WRAPIN, un moteur de recherche en santé proposé par (Gaudinat et al. 2006), permet de mettre à disposition des citoyens, des sites Internet de santé de qualité accrédité. WRAPIN propose un processus de recherche d'information fondé sur une reformulation de requêtes. Ces dernières sont enrichies par une liste pertinente de termes du thésaurus MeSH et du domaine médical. Ceci permet d'obtenir une requête plus précise. Par exemple, une requête concernant une « maladie » aboutit à une recherche se rapportant à ces catégories : « complications », « traitement », « prévention »…Par rapport à notre approche de recherche multi-terminologique, nous pouvons considérer une limite de ce travail est l'utilisation d'une seule terminologie médicale (notamment le thésaurus MeSH) pour l'expansion de requêtes.

(Bratsas et al. 2007) ont mis au point une méthodologie et une procédure pour définir une expansion, fondée sur la logique floue, du modèle d'ontologie et des requêtes. De plus, ils ont construit un modèle d'espace vectoriel fondé sur les ontologies permettant un appariement pertinent entre les critères de recherche, prédéfinis par l'utilisateur, et les connaissances, déjà acquises, concernant un problème de santé. L'expansion de requête se fait en ajoutant les concepts ayant même CUI (Concept Unique Identifier), les synonymes, les types sémantiques, les relations de subsomption de l'UMLS.

En 2008, (Abdou et al., 2008) proposent, pour une recherche dans la base de données Medline, un modèle d'expansion de requêtes basé sur le modèle d'espace vectoriel *tf-idf*[91]. Pour cela, ils construisent un premier ensemble de recherche formé par tous les termes de la requête initiale de l'utilisateur et tous les termes d'indexation appartement aux premiers documents les mieux classés. Par la suite, à chaque terme est associé un poids qui reflète son degré d'importance. À la deuxième étape, le nouvel ensemble est formé par les termes ayant les poids les plus élevés. Et ainsi de suite… Les expérimentions réalisées sur une collection de Medline, mettent en relief la performance du modèle probabiliste utilisé par rapport aux modèles d'espace vectoriel.

En 2009, une étude a été réalisée par (Lu et al., 2009) permettant d'évaluer l'expansion des requêtes en utilisant le MeSH lors de la recherche d'information dans PubMed/Medline. En effet, les auteurs ont essayé d'étudier l'efficacité d'employer le MeSH dans PubMed grâce à son processus d'expansion automatique de requête : appariement automatique des termes (ATM). Pour cela, ils ont construit automatiquement, en premier lieu, une requête en

[91] Se référer au Chapitre 2 pour plus de détails sur le modèle vectoriel

choisissant des mots-clés à partir de la requête initiale. Après, chaque requête est étendue par l'ATM. Les résultats expérimentaux suggèrent que l'expansion des requêtes en utilisant le MeSH dans PubMed peut généralement améliorer la performance des résultats. Dans la même année, l'équipe CISMeF a proposé une optimisation de l'algorithme ATM de PubMed pour améliorer la recherche d'information dans Medline (Thirion et al., 2009). Les nouvelles requêtes construites pour cet effet sont plus précises que les requêtes PubMed actuelles (54.5% vs. 27%)[92]. En effet, la nouvelle approche permet de restituer de nouveaux documents pertinents grâce à la manière d'introduire les synonymes des descripteurs MeSH dans les requêtes.

- **Représentation des textes en sac de mots**

Le sac des mots est la représentation de textes la plus simple qui a été introduite dans le cadre du modèle vectoriel. Il s'agit de transformer les textes des documents en vecteurs dont chaque composante représente un mot. Les mots ont l'avantage de posséder un sens explicite. Nous pouvons le considérer comme étant une suite de caractères appartenant à un dictionnaire, ou bien, de façon plus pratique, comme étant une séquence de caractères non délimiteurs encadrés par des caractères délimiteurs (la ponctuation). Pour cela, il faut alors gérer les sigles, ainsi que les mots composés, ce qui nécessite un prétraitement linguistique. Par exemple, nous pourrons conserver les majuscules pour aider à la reconnaissance de noms propres, mais dans ce cas il faut résoudre le problème des débuts de phrases.

La notion de sac de mots fait référence au fait que la représentation des textes exclut toute analyse grammaticale et toute notion de distance entre les mots. En effet, les composantes du vecteur sont en fonction de l'occurrence des mots dans le texte. D'autres auteurs parlent d'«ensemble de mots » lorsque les poids associés sont binaires.

Dans ce modèle, chaque flexion[93] d'un mot est considérée comme un descripteur différent et donc, une dimension de plus. Ainsi, les différentes formes d'un verbe constituent autant de mots. Par exemple, les mots « déménageur, déménageurs, déménagement, déménagements, déménager, déménage, déménagera, etc. » sont considérés comme des descripteurs différents alors qu'il s'agit de la même racine «déménage ».

Pour faire face à ce problème, les techniques de désuffixation (troncature ; stemming) et de lemmatisation peuvent être appliquées[94].

Lors de la représentation en sac de mots, chacun des mots d'un corpus est un descripteur potentiel. Ceci peut poser le problème de la grande dimension de l'espace de représentation. Or, pour un corpus de taille raisonnable, ce nombre peut être de plusieurs dizaines de milliers.

Pour beaucoup d'algorithmes d'apprentissage, la sélection d'un sous-ensemble de descripteurs peut être indispensable afin de faire face :

[92] De nouvelles optimisations ont été développées en 2010. Je participerai après ma thèse à les évaluer dans le cadre d'amélioration de la recherche d'information.

[93] Variation de la forme des unités lexicales en fonction de facteurs grammaticaux ; nous distinguons traditionnellement la déclinaison (nom, adjectif, pronom) et la conjugaison (verbe).

[94] Se référer au Chapitre 2 ; Section 2.3.3.2 pour plus de détails de la désuffixation et de la lemmatisation.

- ✓ au coût du traitement car le nombre des termes intervient dans l'expression de la complexité de l'algorithme ; plus ce nombre est élevé, plus le volume de calcul est important ;
- ✓ à la faible fréquence de certains termes : nous ne pouvons pas construire des règles fiables à partir de quelques occurrences dans l'ensemble d'apprentissage.

Pour réduire la dimension de l'espace de représentation, nous pourrons supprimer les mots les plus fréquents, puisqu'ils n'apporteraient pas d'information sur la catégorie d'un document. De même, les mots très rares, qui n'apparaissent qu'une ou deux fois dans un corpus, sont supprimés, car leurs faibles fréquences ne permettent pas de construire des règles stables. Cependant, même après la suppression de ces deux catégories de mots, le nombre de candidats peut rester élevé. Dans ce cas, nous pouvons utiliser une méthode permettant de choisir les mots ayant un sens sémantique (appartenant à des terminologies) pour représenter les documents.

Dans le cadre de cette approche, nous présentons, dans les paragraphes qui suivent, l'interprétation de la requête de l'utilisateur en utilisant les concepts des terminologies médicales intégrées dans notre système d'information CISMeF.

Par ailleurs, nous décrivons la migration de notre modèle de recherche du monde mono-terminologique vers l'univers multi-terminologique. Notre modèle est *inspiré* du modèle basé-concepts (cf. Chapitre2).

En effet, une des définitions des ontologies, nous pouvons citer celle de (Zweigenbaum, 1999) qui présente l'ontologie comme *« l'aboutissement formel de la définition d'une terminologie* ». Les principales caractéristiques de la terminologie CISMeF par rapport à une ontologie sont :

- ✓ le vocabulaire est bien connu des documentalistes et des professionnels de la santé et il correspond à celui du domaine médical ;
- ✓ chaque concept a un terme préférentiel (descripteur) pour l'exprimer en langage naturel, un ensemble de propriétés, une définition, un ensemble de synonymes, un ensemble de règles et de contraintes ;
- ✓ les concepts sont organisés selon une relation de subsumption allant du concept le plus général au plus spécifique.

Cependant, ce qui manque à la terminologie CISMeF c'est la dimension formelle qui caractérise plus spécifiquement les ontologies.

5.1.2 STRATEGIE DE RECHERCHE D'INFORMATION MONO TERMINOLOGIQUE DE L'EQUIPE CISMEF

Le but de la recherche d'information est d'apparier la requête de l'utilisateur avec les ressources du catalogue CISMeF les plus représentatives du besoin informationnel de l'utilisateur.

Pour cela, les ressources sont indexées en amont d'une manière manuelle, supervisée ou automatique permettant d'avoir une représentation dans un espace conceptuel.

Étant donné une requête de l'utilisateur exprimée en langage naturel, trois étapes essentielles sont appliquées permettant d'obtenir sa représentation dans un espace conceptuel :

- 1ère étape : normalisation et découpage en mots

Cette étape consiste à analyser la requête initiale de l'utilisateur, la normaliser (enlever la ponctuation, rendre les termes minuscules et sans accents), ensuite la découper en mots et enfin, enlever les mots vides et ranger les mots, ayant un sens sémantique, par ordre alphabétique. Les mots vides sont des termes non significatifs qui peuvent générer du bruit lors du processus de la recherche d'information. Par exemple, dans la requête « le rouge et le noir », nous devrons éliminer les termes « le» et «et » si nous ne souhaitons pas avoir les documents indexés par « **le** corbeau **et le** renard » à cause de la présence de ces termes vides.

Les termes rares dans un document ne peuvent pas le représenter et les termes qui apparaissent fréquemment dans tous les documents ne peuvent pas être utilisés pour les différencier. Ainsi, ils sont assimilés à des mots vides. L'ensemble des termes à retenir sont les termes qui ont un poids entre le seuil des termes rares et le seuil des mots vides (Luhn, 1958). Une autre hypothèse, qui peut être prise en considération, suggère que les poids des termes dans les documents sont définis en appliquant des méthodes statistiques. Selon ces poids, les termes les plus descriptifs sont retenus (Zipf, 1949).

Pour notre traitement des requêtes, nous avons recours à une liste des mots vides obtenue à partir de Lexique[95]. Cette liste est régulièrement maintenue par notre équipe afin d'ajouter et de mettre à jour les termes qui peuvent être reconsidérés (ou non) comme pertinents pour la recherche d'information. S'ajoute à cet ensemble, une liste d'expressions vides (tel que « tout d'abord ») développée au fur et à mesure, dans le but de diminuer le bruit tant que possible.

À la fin de cette étape, nous avons donc tous les termes significatifs de la requête rangés par ordre alphabétique.

- 2ème étape : désuffixation des termes

Disposant du sac de mots constitué de l'ensemble des mots les plus significatifs de la requête, rangés par ordre alphabétique, un deuxième traitement est appliqué afin de supprimer la trop grande variabilité des mots. En effet, les variabilités flexionnelles (pluriel, conjugaison) et les variabilités dérivationnelles (passage d'une catégorie morphosyntaxique à une autre) introduisent un grand nombre de termes différents raccordés à une même racine et donc, dans la plupart des cas, à un même sens.

Dans l'équipe CISMeF, nous avons utilisé une technique qui repose sur une liste de suffixes et un ensemble de règles de désuffixation construites à priori et qui permettent de retrouver

[95] **Lexique**. Lexique 3 est une base de données qui fournit pour 135 000 mots du français: les représentations orthographiques et phonémiques, la syllabation, la catégorie grammaticale, le genre et le nombre, les fréquences, les lemmes associés… URL : http://www.lexique.org/

les radicaux des termes. L'idée générale est d'éliminer ou remplacer, au fur et à mesure, les suffixes rencontrés selon des règles de désuffixation dépendant de la taille du mot, du suffixe et du mot. L'ordre de traitement des suffixes dépend de leurs tailles en favorisant les plus longs en premier. Par exemple, pour le mot « fonctionnelles », nous obtenons le radical « fonc » suite à trois passages par la liste des suffixes et en respectant les règles de désuffixation.

- 3ème étape : identification des descripteurs MeSH

La troisième étape consiste à identifier les potentiels descripteurs MeSH à partir du sac des mots de la requête de l'utilisateur.

En effet, nous cherchons la combinaison des termes (Nb = nombre de mots non vides de la requête) la plus longue qui pourrait être identifiée comme un descripteur. Ensuite, nous examinons la combinaison de longueur (Nb-1) et, ainsi de suite jusqu'à l'épuisement de toutes les combinaisons possibles.

Exemple : Soit la requête de l'utilisateur « *maladie de l'appareil digestif* ». Après l'élimination des mots vides et la désuffixation des termes les plus significatifs (1ère et 2ème étapes), le sac de mots est constitué de cet ensemble {appareil ; digestif ; maladi}. La combinaison la plus longue, à trois termes nous permet d'identifier le descripteur « *maladie de l'appareil digestif* » du thésaurus MeSH. En conséquence, le processus s'arrête à ce stade sans chercher d'autres termes d'indexation de longueur inférieure.

Une fois les descripteurs du thésaurus MeSH identifiés, cette 3ème étape pourrait être enrichie par l'identification des termes devant être indexés avec les descripteurs ainsi identifiés, tels que l'affiliation des qualificatifs, l'application des règles d'indexation du MeSH et des actions pharmacologiques (cf. Chapitre2). Par exemple, si la requête est indexée par le concept chimique supplémentaire « *Racécadotril* », elle doit aussi être indexée par le descripteur « *Antidiarrhéiques* » représentant son action pharmacologique.

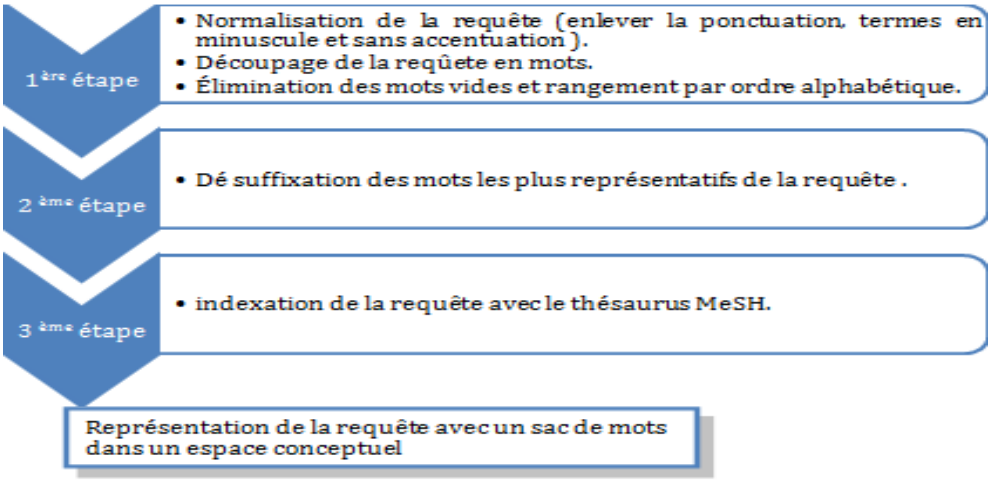


Figure 5.1.2. Résumé du traitement pour représenter la requête de l'utilisateur dans un monde mono-terminologique

Ainsi, à ce stade, nous avons la représentation de la requête de l'utilisateur d'une part, et celle des ressources du catalogue CISMeF (déjà indexées), d'autre part, dans un même espace conceptuel.

Afin de trouver la meilleure correspondance entre les ressources du catalogue CISMeF et la requête de l'utilisateur, l'algorithme de recherche de l'équipe CISMeF (Soualmia et al. 2006) était basé sur trois étapes largement inspirées des heuristiques de PubMed permettant l'accès à la base de données bibliographique MEDLINE (PubMed help, 2005):

- 1ère phase : *la recherche au niveau des termes d'indexation ou au niveau des titres des ressources*. Si les termes représentatifs de la requête de l'utilisateur correspondent à des termes au niveau du titre de la ressource ou aux termes d'indexation de cette dernière, le processus s'arrête. Nous aurons, comme résultat, non seulement les ressources indexées par les descripteurs ainsi identifiés, mais aussi les ressources indexées par les descripteurs qui les subsument directement ou indirectement et au niveau de toutes les hiérarchies possibles[96] ;
- 2ème phase : *la recherche dans les métadonnées des ressources*. Si la première étape ne donne aucun résultat, la recherche s'effectue au niveau des métadonnées (résumé, auteurs, éditeur…) des ressources avec une mesure d'adjacence fixée empiriquement à n ; n étant 5*(nombre des mots de la requête -1) ;
- 3ème phase : *la recherche en plein texte.* Si la deuxième étape ne donne pas de résultat, la recherche s'effectue en plein texte avec une mesure d'adjacence fixée empiriquement égale à n termes ; n étant 10*(nombre des mots de la requête-1). Pour cette phase, nous utilisons l'outil d'oracle *Oracle text*[97] qui permet l'indexation, l'interrogation et la présentation des documents.

En cas d'échec de ces trois étapes de recherche de l'algorithme CISMeF, nous avons mis au point une recherche complémentaire :

- 4ème phase : *la recherche d'information étendue à Google-CISMeF*. Google-CISMeF[98] consiste à indexer les pages de Google en se restreignant aux sites éditeurs de CISMeF (Gehanno et al. 2009). Le corpus de CISMeF est d'environ 10^5 pages (pour 73.800 ressources ; plusieurs ressources ont plusieurs URL) alors que le corpus de Google est d'environ 10^6 pages. Nous avons utilisé le moteur de recherche personnalisé de Google[99] qui a permis d'inclure plusieurs sites et pages web et d'effectuer des recherches automatiques rapides dans les liens. Le résultat de la recherche est affiché selon l'algorithme PageRank (Brin et al., 1998). L'évaluation de la pertinence des

[96] En effet, un descripteur MeSH peut appartenir à plusieurs hiérarchies tel que « Fluorure de phosphate acidulé » appartenant tantôt à l'arborescence D « Produits chimiques, biologiques et pharmaceutiques » tantôt à l'arborescence J « Technologie aliments et boissons ».
[97] Introduction to Oracle text. URL : http://download.oracle.com/docs/cd/B10500_01/text.920/a96517/cdefault.htm
[98] CISMeF, Outils de recherche personnalisés. URL : http://www.chu-rouen.fr/documed/cismefgoogle.htm
[99] Google recherche personnalisée. URL : http://www.google.com/cse/

ressources restituées avec Google-CISMeF et notre moteur de recherche nous a permis d'enregistrer une meilleure couverture en faveur de Google-CISMeF (100% vs. 96%).

5.1.3 STRATEGIE DE RECHERCHE D'INFORMATION MULTI-TERMINOLOGIQUE DE L'EQUIPE CISMEF

À la différence du précédent algorithme mono-terminologique, notre nouvelle stratégie de recherche au sein du catalogue CISMeF est basée, non seulement sur une expansion de requêtes s'appuyant sur l'enrichissement par synonymie et hiérarchisation, mais aussi par appariement entre les différentes terminologies présentes dans notre base de données (back office CISMeF)[100].

En effet, grâce au passage du monde mono-terminologique vers l'univers multi-terminologique, nous avons pu réaliser une recherche d'information médicale multi-terminologique qui a permis l'enrichissement de l'information fournie à l'utilisateur selon ses propres connaissances terminologiques.

Notre nouvel algorithme se différencie principalement, de ce qui précède au niveau de l'indexation de la requête de l'utilisateur. Le sac de mots, déjà employé par (Soualmia, 2004) (Pereira, 2008), a été modifié selon nos propres besoins. S'ajoute à ceci, la mise à jour de la méthode de désuffixation. Une étude, permettant de comparer différents algorithmes de désuffixation a été faite par (Pereira, 2008) et a permis de mettre en relief les avantages de l'algorithme de Lucene (Hatcher et al., 2004) qui s'inspire des travaux de Porter (Porter, 1980). L'évaluation a été réalisée avec trois algorithmes, à savoir celui de l'équipe CISMeF que nous avons utilisé jusqu'à maintenant, celui de Carry (Paternostre et al. 2002) et celui de Lucene. Bien que, d'une manière générale, le principe de désuffixation est à peu près le même, la différence observée entre les trois algorithmes est due aux règles appliquées. Le résultat de l'évaluation enregistre une F-mesure à 77,9% pour l'algorithme de Lucene, 70,4% pour celui de CISMeF et enfin 66,7% pour celui de Carry.

5.1.3.1 ALGORITHMIQUE

- 1ère étape : normalisation et découpage en mots

Les mêmes traitements de base sont réalisés pour la requête de l'utilisateur. À la fin de cette étape, nous avons donc tous les termes significatifs de la requête rangés par ordre alphabétique

- 2ème étape : désuffixation des termes

L'algorithme de Lucene se déroule en 6 étapes permettant l'élimination des suffixes standards, le traitement des suffixes verbaux, des formes particulières et des caractères

[100] Se référer au Chapitre 3 pour plus de détails sur le back office CISMeF

doubles… Pour chaque étape, une liste de règles est appliquée, dépendant d'une ou de plusieurs conditions[101].

- 3ème étape : identification des descripteurs des terminologies

La troisième étape consiste à identifier les potentiels descripteurs à partir du sac des mots de la requête de l'utilisateur. Ces descripteurs appartiennent aux différentes terminologies intégrées dans la base de données CISMeF.

En effet, pour chaque terminologie, nous cherchons la combinaison des termes (Nb = nombre de mots non vides de la requête) la plus longue qui pourrait être identifiée comme un descripteur. Si c'est le cas, le processus d'identification de descripteurs s'arrête, pour cette terminologie.

Si pour une terminologie donnée, un descripteur de longueur (Nb) n'a pas pu être identifié, nous examinons la combinaison de longueur (Nb-1) ; et, ainsi de suite, jusqu'à l'épuisement de toutes les combinaisons possibles.

Exemple : Soit la même requête de l'utilisateur « *maladie de l'appareil digestif* » et le sac de mots correspondant {appareil ; digestif ; maladi} ; Nb=3.

En procédant aux trois étapes de l'algorithme, pour les différentes terminologies disponibles (une à une), nous obtenons les descripteurs avec la combinaison des termes la plus longue {appareil ; digestif ; maladi} « *maladie de l'appareil digestif* » du thésaurus MeSH, « *maladies de l'appareil digestif* » de la classification MedlinePlus, « *maladies de l'appareil digestif* » de la CIM-10 et « *maladies de l'appareil digestif* » de la nomenclature SNOMED. S'ajoute à cet ensemble, le descripteur « *appareil digestif* » du dictionnaire DRC et de la CCAM obtenu avec la combinaison des termes de longueur (Nb-1=2), {appareil ; digestif}, {appareil ; maladi}, {digestif ; maladi}, soit deux termes.

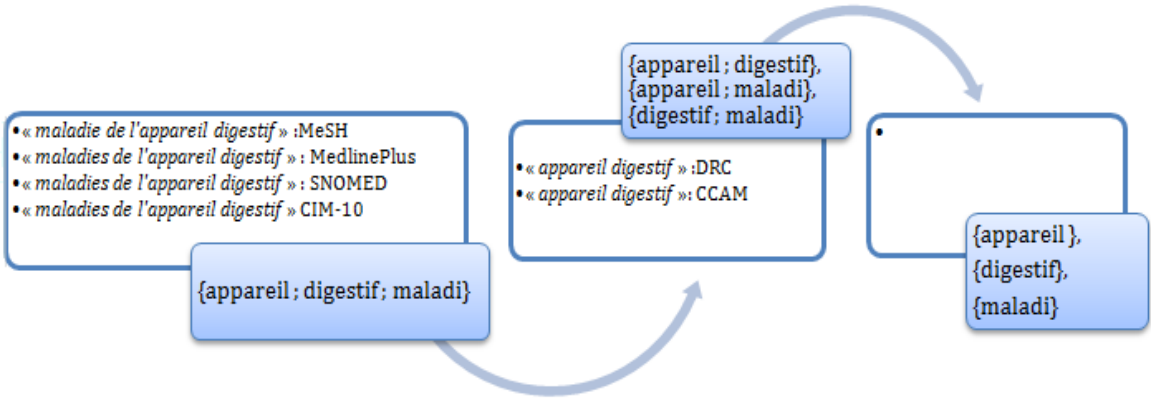


Figure 5.1.3.1.1. Identification des descripteurs des terminologies médicales

Comme avec la version mono-terminologique de l'indexation, cette phase est complétée par l'identification des termes devant être indexés avec les descripteurs ainsi identifiés, tels que

[101] Se référer à l'Annexe C, pour l'algorithme Lucene

l'affiliation des qualificatifs, l'application des règles d'indexation du MeSH et des actions pharmacologiques.

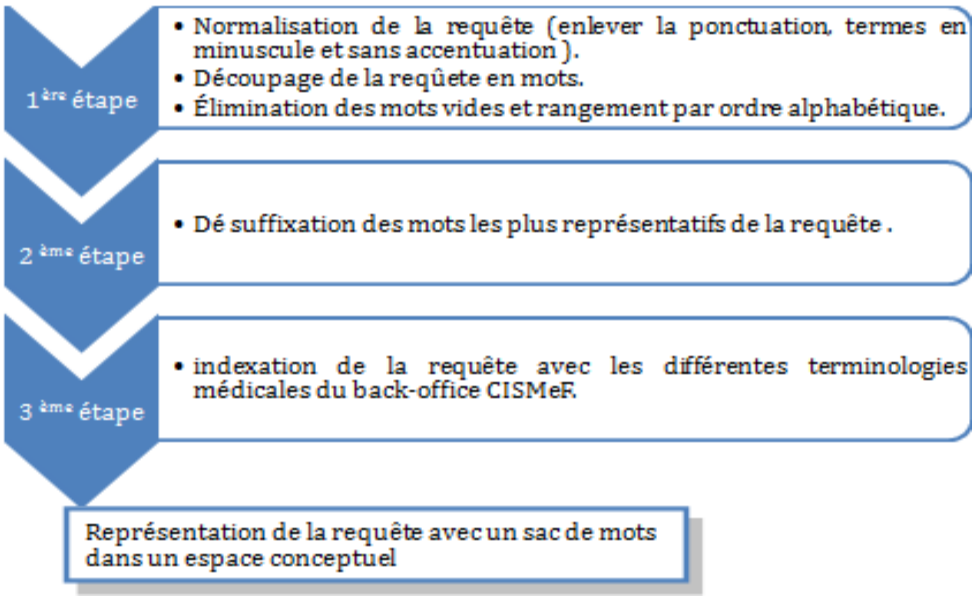


Figure 5.1.3.1.2. Résumé du traitement pour représenter la requête de l'utilisateur dans un monde multi-terminologique

Ainsi, à partir de la représentation de la requête, notre algorithme de recherche d'information multi-terminologique s'applique selon trois phases :

- 1ère phase : *Identification des descripteurs par rapport aux termes d'indexation de la ressource ou au niveau du titre de la ressource*. En effet, nous jugeons que retrouver des ressources qui sont indexées (ou leurs titres représentés) par des descripteurs identifiables au niveau de la requête de l'utilisateur est la première phase la plus pertinente en terme d'appariement termes/ressources.

  Par exemple, étant donnée la requête de l'utilisateur « asthme de l'enfant », le résultat de la requête booléenne est le suivant :

Requête : asthme de l'enfant
Résultat : ((asthme.mr[ART][CIM][CIP][DRC][MSH][SNO] ou asthme.ti) et (enfant.mr[MSH][SNO] ou enfant.ti)) ou (asthme chez l'enfant.mr[MED] ou asthme enfant.ti)
Note : *mr :* mot réservé (métaterme + descripteur + qualificatif + type de ressources), rechercher le terme comme un terme d'indexation de la ressource ; *ti :* rechercher le terme au niveau du titre de la ressource ; *ART :* la terminologie WHO-ART ; *CIM :* la classification CIM-10 ; *CIP :* la terminologie CISP2 ; *DRC :* la terminologie DRC ; *MSH :* le thésaurus MeSH ; *SNO :* la nomenclature SNOMED ; *MED :* la terminologie MedlinePlus.

  Dans ce cas, tous les termes de la requête de l'utilisateur sont identifiés comme

descripteurs dans une ou plusieurs terminologies.

Si un terme de la requête initiale n'a pas été identifié comme un descripteur d'une terminologie, la requête booléenne est transformée afin de rechercher le terme au niveau du titre de la ressource.

Par exemple,

Requête : développement psychomoteur de l'enfant
Résultat : (developpement de l'enfant.mr[MSH] et psychomoteur.ti) ou (developpement psychomoteur.mr[SNO] et enfant.mr[SNO]) Ou (developpement de l'enfant.mr[MED] et psychomoteur.ti)
Note : *mr :* mot réservé (métaterme + descripteur + qualificatif + type de ressources), rechercher le terme comme un terme d'indexation de la ressource ; *ti :* rechercher le terme au niveau du titre de la ressource ; *MSH :* the MeSH thesaurus ; *MSH :* le thésaurus MeSH ; *SNO :* la nomenclature SNOMED ; *MED :* la terminologie MedlinePlus.

Dans cet exemple, tous les termes de la requête ont été identifiés comme des descripteurs SNOMED « développement psychomoteur » et « enfant ». En prenant le thésaurus MeSH, le terme psychomoteur n'a pas été identifié comme un descripteur, donc la requête booléenne est complétée par une recherche dans le titre, d'où la recherche suivante : *developpement de l'enfant.mr[MSH]* ***et*** *psychomoteur.ti.*

- 2ème phase : *Identification des descripteurs au niveau des métatermes*. En effet, si un terme de la requête n'a pas été reconnu comme un descripteur ou présent dans le titre de la ressource, la recherche est faite au niveau des métadonnées de la ressource (les champs caractérisant la ressource tels que la description, l'éditeur, la date…) avec une mesure d'adjacence égale à 5. En d'autres termes, la distance en termes de nombre de mots entre les termes de la requête est égale à 5.

Par exemple,

Requête : association formotérol corticostéroïde
Résultat : (((corticosteroides.sr ou corticosteroide.ti)) et (association.mr[CIS][MSH] ou association.ti)) et (formoterol.mr[MSH] ou formoterol.ti) = 0 -> ((((corticosteroides.sr ou corticosteroide.tc)) et (association.mr[CIS][MSH] ou association.tc)) et (formoterol.mr[MSH] ou formoterol.tc)) ou (l'association formoterol corticosteroide.at)
Note : *mr :* mot réservé (métaterme + descripteur + qualificatif + type de ressources), rechercher le terme comme un terme d'indexation de la ressource ; *ti :* rechercher le terme au niveau du titre de la ressource ; *tc :* tous les champs, rechercher le terme au niveau des métadonnées ; *CIS :* la terminologie CISMeF ; *MSH :* le thesaurus MeSH.

Le résultat met en relief la recherche en deux phases.

- 3ème phase : *Identification des descripteurs en plein texte*. Dès lors, un terme de la requête n'est pas reconnu comme un descripteur, ni présent dans le titre de la ressource, ni au niveau des métadonnées de la ressource, la recherche est appliquée en plein texte avec une mesure d'adjacence égale à 10.

  Par exemple,

Requête : bronchite asthmatiforme
Résultat : ((bronchite.mr[ART][MED][MSH][SNO] ou bronchite.ti)) ET asthmatiforme.ti = 0 -> (((bronchite.mr[ART][MED][MSH][SNO] OU bronchite.tc)) ET asthmatiforme.tc) OU (bronchite asthmatiforme.at) = 0 -> bronchite asthmatiforme.aj
Note : *mr :* mot réservé (métaterme + descripteur + qualificatif + type de ressources), rechercher le terme comme un terme d'indexation de la ressource ; *ti :* rechercher le terme au niveau du titre de la ressource ; *tc :* tous les champs, rechercher le terme au niveau des métadonnées ; *at :* adjacence tous champs ; *aj :* adjacence plein texte ; *ART :* la terminologie WHO-ART ; *MSH :* le thésaurus MeSH ; *SNO :* la nomenclature SNOMED ; *MED :* la terminologie MedlinePlus.

  Pour cet exemple, nous avons un résultat grâce à la recherche en plein texte. Ce qui veut dire que les termes « *bronchite* » et « *asthmatiforme*» sont présents dans le texte du document et distants de moins de10 mots.

- 4ème phase : *la recherche d'information étendue à Google-CISMeF*. Les descripteurs identifiés de la requête sont appariés avec les concepts UMLS ayant le même CUI (Concept Unique Identifier). Un CUI regroupe tous les termes des différentes terminologies médicales qui partagent le même sens. L'expansion de requête peut être, par la suite, enrichie, par transitivité, par d'autres synonymes de concepts. Par exemple, le descripteur MeSH « *avortement provoqué* » est apparié avec le descripteur MedDRA « *interruption de la grossesse* » ou encore le descripteur de la CIM-10 « i*nterruption de la grossesse affectant le fœtus et le nouveau-né* » ayant le même CUI UMLS. Ainsi, la recherche dans Google permet de retrouver tous les documents, des sites éditeurs CISMeF, indexés par ces trois termes.

5.1.3.2 IMPLEMENTATION DE L'ALGORITHME

Le passage du monde mono-terminologique vers l'univers multi-terminologique et, par conséquent, la mise à jour de la base de données et l'implémentation de l'algorithme a été réalisé en collaboration avec une équipe de 8 ingénieurs de l'Institut National des Sciences Appliqués (INSA) de Rouen dans le cadre d'un PIC (Projet INSA Certifié) 2008-2009.

Dans le même cadre d'implémentation, l'algorithme de recherche d'information multi-terminologique est programmé en Java et disponible dans la version R&D de Doc'CISMeF.

5.1.3.3 EVALUATION DE LA PLUS VALUE DE LA MULTI-TERMINOLOGIE

Pour évaluer notre approche, nous avons réalisé une étude, fin 2009, permettant de mesurer la valeur ajoutée de l'univers multi-terminologique par rapport au monde mono-terminologique lors de la recherche d'information dans le catalogue CISMeF.

La figure 5.1.3.3 illustre un exemple de recherche d'information multi-terminologique au sein du catalogue CISMeF et met en relief l'apport d'une telle recherche. En effet, le descripteur CCAM « *JQQM003 - échographie de surveillance de la croissance fœtale avec échographie-doppler des artères utérines de la mère et des vaisseaux du fœtus* » présente pour l'utilisateur une information plus précise que le descripteur MeSH « *échographie prénatale* ».

Figure 5.1.3.3. Exemple du résultat de la recherche d'information multi-terminologique

5.1.3.3.1 Méthode

Pour cela, nous avons analysé les requêtes des utilisateurs les plus fréquentes de Doc'CISMeF (analyse des logs), les avons classifiées en requêtes à un seul terme, requêtes à deux termes et requêtes à trois termes. Ces types de requêtes mettent en relief la complexité croissante de l'algorithme de recherche d'information multi-terminologique. Avec des requêtes à plus de trois termes, nous n'avons pas eu un résultat significatif permettant d'évaluer notre stratégie de recherche.

L'étude (sakji et al. 2010b) est réalisée sur le corpus du catalogue CISMeF composé de 36.107 ressources indexées manuellement et 22.240 ressources indexées automatiquement[102].

La recherche est effectuée en deux temps :

1. lancer les requêtes en mono terminologie avec le thésaurus MeSH ;
2. lancer les requêtes en multi-terminologie avec toutes les terminologies présentes dans le back-office de CISMeF.

Néanmoins, face au problème d'interprétation de requêtes par le moteur de recherche Doc'CISMeF et, voulant se concentrer sur la valeur ajoutée de la multi-terminologie, nous avons dû transformer la requête interprétée selon notre algorithme de recherche. Pour cela nous avons restreint l'ensemble des requêtes sélectionnées à celles ayant des réponses selon la première phase de l'algorithme (identification des descripteurs et recherche dans le titre de la ressource) et ne gardant, par la suite, que l'identification des descripteurs.

Ainsi, nous avons comme résultat :

Requêtes à 3 termes	**Requête mono-terminologique**	**Requête multi-terminologique**	**Requête multi-terminologique non MeSH**
maladie de l'appareil digestif	maladie de l'appareil digestif.mr**[MSH]**	(maladie de l'appareil digestif.mr**[MSH]**) Ou (appareil digestif.mr**[DRC]** et maladie.ti) Ou (appareil digestif.mr**[CCA]** et maladie.ti) OU (maladies de l'appareil digestif.mr**[CIM]**) OU (maladies de l'appareil digestif.mr**[MED]**) OU (maladies de l'appareil digestif.mr**[SNO]**)	((appareil digestif.mr**[DRC]** et maladie.ti) Ou (appareil digestif.mr**[CCA]** et maladie.ti) OU (maladies de l'appareil digestif.mr**[CIM]**) OU (maladies de l'appareil digestif.mr**[MED]**) OU (maladies de l'appareil digestif.mr**[SNO]**)) ***sauf*** (maladie de l'appareil digestif.mr**[MSH]**)

L'évaluation a été réalisée par trois experts : une documentaliste-pharmacienne de l'équipe CISMeF, un médecin senior de santé publique et un médecin junior de médecine de travail (assistante hospitalière universitaire en médecine du travail). Nous avons essayé de choisir des évaluateurs du domaine aussi diversifiés (en spécialités) que possible afin, non seulement, d'avoir un jugement objectif, mais aussi de voir les différents points de vue des utilisateurs selon leurs contextes et leurs attentes en recherche d'information.

Pour déterminer l'apport de la multi-terminologie, les experts ont mesuré la qualité, en termes de besoin informationnel, des ressources distinctes entre les deux modes de recherche. Les résultats ont été répertoriés comme *pertinent* si la ressource correspond, bel et bien, au thème de la recherche, *non pertinent* si la ressource n'a pas de relation avec le sujet de la requête et *intermédiaire* sinon.

[102] Les chiffres datant de Décembre 2009, lors de la réalisation de l'étude

Ainsi, pour chaque type de requêtes (à 1 terme, à 2 termes ou à 3 termes), chaque évaluateur classe son jugement selon les trois modalités citées plus haut.

5.1.3.3.2 Résultats

La première colonne du tableau 5.1.3.3.2.1 met en évidence le nombre des ressources restituées par le mode de recherche mono-terminologique pour chaque type de requête. La deuxième colonne énumère le nombre des ressources recueillies par le mode de recherche multi-terminologique. Par construction, toutes les ressources retournées en mono-terminologie sont également restituées en multi-terminologie. Dans la troisième colonne, nous enregistrons le pourcentage des différences entre les deux modes de recherche en terme de couverture. Nous remarquons que le pourcentage le plus élevé est observé pour les requêtes à deux termes avec 44,88%.

	Nombre des ressources par		**Pourcentage des différences**
	Recherche mono-terminologique	**Recherche multi-terminologique**	
Requête à 1 terme	2.942	3.432	16,65%
Requête à 2 termes	3.353	4.858	44,88%
Requête à 3 termes	7.864	7.993	1,64%
Total	14.159	16.283	15%

Tableau 5.1.3.3.2.1. Nombre des ressources selon les différents modes de recherche et les différents types de requêtes ainsi que le pourcentage de différence entre les deux modes de recherche

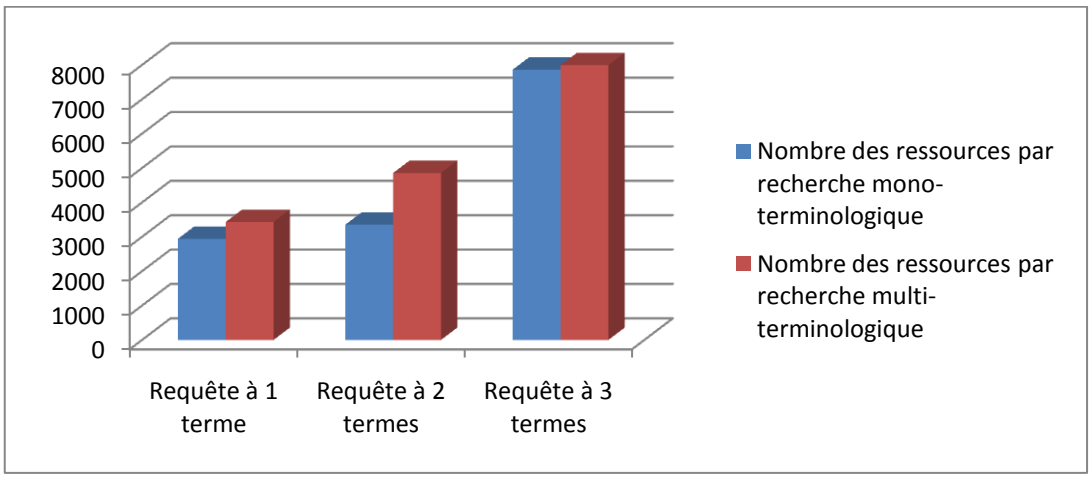


Figure 5.1.3.3.2.1. Illustration de la différence entre les deux modes de recherche selon chaque type de requête

Le tableau 5.1.3.3.2.2 décrit l'évaluation des trois spécialistes que nous avons considérés comme des gold standard. Leur évaluation se focalise sur les ressources qui ont été restituées par la recherche multi-terminologique et absentes pour la mono-terminologie. Les valeurs enregistrées dans le tableau représentent les pourcentages des ressources qui ont été jugées par les évaluateurs comme pertinentes, intermédiaires ou non pertinentes selon les trois types de requêtes et les trois spécialistes.

Pour les requêtes à 1 terme, le pourcentage des résultats pertinents est évalué à 67,11%, alors que l'intermédiaire était à 10,35% contre 21,43% pour les non pertinents.

Pour les requêtes à 2 termes, le résultat global était un peu différent dans la mesure où le meilleur pourcentage est toujours enregistré pour le *pertinent* avec 57,81% suivi du résultat intermédiaire avec 31 ,47% ensuite par le résultat *non pertinent* avec 10,71%.

Pour les requêtes à 3 termes, les ressources pertinentes enregistrent un taux de 43,66%, les ressources intermédiaires un taux de 32,44% et les non pertinentes sont à 23,9%.

La moyenne des résultats selon les trois types de requêtes est mise en relief au niveau du tableau 5.1.3.3.2.3 : d'une manière générale, le premier expert juge les résultats pertinents à 53,8% des cas, le deuxième expert à 68,3% et le troisième expert à 47,7% des cas.

	Requête à 1 terme			**Requête à 2 termes**			**Requête à 3 termes**		
	pertinent (%)	Intermédiaire (%)	*Non pertinent*(%)	*pertinent* (%)	Intermédiaire (%)	*Non pertinent*(%)	*pertinent* (%)	Intermédiaire (%)	*Non pertinent*(%)
Expert 1	73,03	5,31	21,66	47,17	33,20	19,63	41,12	21,44	37,44
Expert2	71,66	4,82	20,19	75,94	18,58	5,47	53,87	41,00	5,13
Expert 3	56,65	20,92	22,43	50,33	42,63	7,04	35,99	34,88	29,13
Moyenne	67,11	10,35	21,43	57,81	31,47	10,71	43,66	32,44	23,90

Tableau 5.1.3.3.2.2. Résultat de l'évaluation des ressources disparates entre la recherche d'information multi-terminologique et la recherche d'information mono-terminologique

Le résultat moyen par expert pour les différents types de requêtes			
	pertinent (%)	**Intermédiaire (%)**	**Non pertinent (%)**
Expert 1	53,8	20,0	26,3
Expert 2	68,3	21,5	10,2
Expert 3	47,7	32,8	19,5
Moyenne	56,6	24,7	18,7

Tableau 5.1.3.3.2.3. Évaluation des résultats de la recherche d'information multi-terminologique par expert

La figure ci-dessous met en relief les courbes décroissantes en termes de pertinence des résultats de l'évaluation selon les trois experts. L'évaluation est réalisée pour les ressources restituées distinctes entre les deux modes de recherche d'information : mono-terminologique et multi-terminologique.

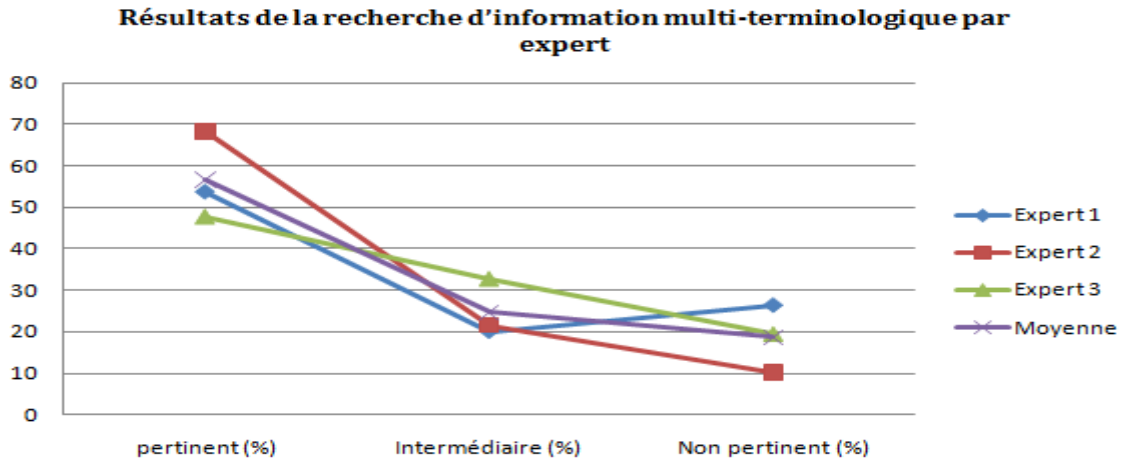


Figure 5.1.3.3.2.2. Évaluation des résultats de la recherche multi-terminologique

5.1.3.3.3 Discussion

Les résultats observés (cf. Tableau 5.1.3.3.2.1) montrent la valeur ajoutée de la recherche multi-terminologique à la mono-terminologique en terme de couverture : 16.283 ressources restituées par le premier mode de recherche vs. 14.159 par le deuxième, soit +15%.

Malgré la différence de jugement des trois experts dans certains cas, globalement, les résultats sont homogènes : nous avons, en tête de liste, les ressources jugées pertinentes (56,6%) suivies des moins pertinentes (24,7%) et enfin les non pertinentes (18,7%).

Pour cette évaluation, le rappel est incalculable étant donné que nous ne connaissons pas le nombre des ressources pertinentes pour une requête donnée, dans la base de données.

La pertinence de la multi-terminologie pour les requêtes à trois termes (43,66%) est assez

faible à cause de la difficulté de la mise en correspondance entre la représentation de la requête de l'utilisateur et les concepts des terminologies, alors qu'elle est meilleure pour les requêtes à un et deux termes (respectivement 67,11% et 57,81%).

L'évaluation de la valeur ajoutée de la multi-terminologie a été faite fin 2009. Depuis ce temps, nous avons ajouté plusieurs terminologies médicales à notre système d'information CISMeF et nous avons amélioré et enrichi l'indexation (manuelle et automatique) des ressources ce qui nous incite à refaire cette étude avec un ensemble plus important de requêtes.

En effet, la limite de cette évaluation est le nombre de requêtes étudiées, dû au fait que la validation des résultats est faite d'une manière exclusivement manuelle. Du coup, chaque expert dispose de plusieurs centaines de ressources à étudier et à juger.

Par ailleurs, étant donnée la connaissance peu développée des indexeurs concernant les nouvelles terminologies médicales intégrées dans notre système, le nombre des ressources indexées manuellement dans l'univers multi-terminologique demeure assez restreint, par rapport aux ressources indexées automatiquement, d'une part, et par rapport au nombre de ressources indexées manuellement dans le monde mono-terminologique, d'autre part.

Actuellement, parmi les 38.237 ressources du catalogue CISMeF indexées manuellement, 32.970 (86,22%) sont indexées par le thésaurus MeSH seulement, 3.866 (10,11%) sont indexées par deux terminologies, 1.397 (3,65%) sont indexées par trois terminologies et 4 (0,02%) sont indexées par quatre terminologies.

Le tableau qui suit résume le nombre des ressources indexées manuellement et automatiquement par les différentes terminologies médicales du back-office CISMeF :

Terminologies	**Nombre de ressources indexées manuellement**	**Nombre de ressources indexées automatiquement**
CCAM	345	4.642
CIM-10	3	5.956
CISP2	2	2.608
CLADIMED	2	5.438
Codes médicaments	1.462	15.314
DRC	1	11.331
LPP	4	4.956
MedDRA	11	11.165
MedlinePlus	2	6.167
MeSH	38.237	33.935
Orphanet	0	10.944

SNOMED	55	25.568
WHO-ART	3	4.594
WHO-ATC	4.785	12.937
WHO-CIF	0	2.485
WHO-ICPS	0	2.588

Pour les premiers pas, dans la démarche d'intégration de l'univers multi-terminologique dans le catalogue CISMeF, nous avons essayé de mettre au point cette « preuve de concept » afin de mesurer la présumée valeur ajoutée de la recherche d'information multi-terminologique.

Comme exemple illustrant la valeur ajoutée de la multi-terminologie, nous pouvons citer la recherche concernant le syndrome de Rokitansky ou MRKH. Il s'agit d'une maladie qui se manifeste par une absence partielle ou totale du vagin et de l'utérus.

En effet, la requête de l'utilisateur « mrkh » permet de récupérer quatre ressources indexées avec ce concept en utilisant le thésaurus MeSH. Les ressources restituées ayant l'abréviation « mrkh » dans le titre étant donné que ce terme n'est pas un concept MeSH (cf. Figure 5.1.3.3.3.1). À travers une recherche d'information employant toutes les terminologies médicales présentes dans le back-office CISMeF, nous avons un résultat constitué de six ressources indexées avec ce concept. L'identification des ressources est faite, en plus de la recherche en titre, grâce au descripteur MedDRA «Mayer-rokitansky-kuster-hauser syndrome » qui est un terme d'indexation (cf. Figure 5.1.3.3.3.2).

Figure 5.1.3.3.3.1. Résultat de la recherche d'information mono-terminologique

Figure 5.1.3.3.3.2. Résultat de la recherche d'information multi-terminologique

5.2 CLASSEMENT DU RESULTAT DE LA RECHERCHE D'INFORMATION

Classer les résultats de la recherche d'information implique un ordonnancement des documents, du plus ressemblant au besoin informationnel de la requête de l'utilisateur au moins pertinent.

La masse d'informations disponibles sur le Web peut conduire, dans la plupart du temps, à des résultats pléthoriques, ce qui déroute les utilisateurs. Face à ce constat, plusieurs travaux ont vu le jour concernant la mise au point des méthodes de tri automatiques, des résultats de la recherche (Courtois et al ., 1999).

Le but du classement est d'afficher dans les 10 à 20 premières réponses les documents répondant au mieux à la requête de l'utilisateur. Généralement, si nous ne trouvons pas ce que nous cherchons dans les toutes premières pages du résultat, nous reformulons notre question.

(Lardy, 2000) résume les méthodes de tri en trois catégories :

- ✓ le tri par pertinence (relevance ranking). Cette méthode repose sur des travaux de recherche déjà anciens de Robertson et Sparckjones (Robertson et al., 1976). Les résultats d'une requête sont affichés selon un ordre déterminé par le calcul d'un score pour chaque réponse. La pertinence est fondée sur :

    1. le poids d'un mot dans un document qui est déterminé par sa place dans le document : il est maximum pour le titre et le début du texte; à l'intérieur il est plus important si le mot est en majuscule ;

2. la densité qui est basée sur la fréquence d'occurrence dans un document par rapport à la taille du document. Si deux documents contiennent le même nombre d'occurrences, le document le plus petit sera favorisé ;
3. le poids d'un mot dans la base qui est basé sur la fréquence d'occurrence pour toute la base de données. Les mots peu fréquents dans le corpus sont favorisés. Les mots vides sont soit éliminés soit sous-évalués ;
4. la correspondance d'expression qui est basée sur la similarité entre la représentation de la requête et la représentation des documents. Un document contenant une expression identique à celle de la requête reçoit le poids le plus élevé ;
5. la relation de proximité qui est basée sur l'adjacence des termes de la requête dans le document.

- ✓ le tri par popularité avec 2 variantes : en fonction du nombre de liens pointant sur une page (algorithme PageRank ; méthode de Google). Google évalue l'importance d'une page par les liens qu'elle reçoit mais analyse en plus la page qui contient le lien.

  L'autre possibilité de tri par popularité est celle en fonction du nombre de visites et du temps passé (méthode de DirectHit[103]) ;

- ✓ le tri par calcul dynamique de catégories : classement des documents trouvés dans des dossiers (clustering) constitués automatiquement en fonction des réponses (méthode de NorthernLight[104]).

Dans la même perspective d'orienter l'utilisateur vers le résultat le plus pertinent par rapport sa requête initiale, nous citons le travail de (Sakji et al. 2008) qui définissent un contexte conceptuel fondé sur un treillis de Galois, construit à partir de pages web, en association avec des ontologies. L'utilisateur peut trouver les pages web qui répondent mieux à sa requête en naviguant dans le treillis grâce à la mesure de similarité proposée entre ses concepts.

S'inspirant des travaux de (Lardy, 2000), nous avons modifié le classement du résultat de la recherche d'information dans CISMeF, en prenant en compte le poids des termes d'indexation.

Même si la date de publication des ressources (et spécialement du domaine de la santé) est importante pour restituer, aux utilisateurs, les plus récentes, nous avons constaté que l'introduction de la notion de pondération des termes devient requise pour notre classement.

Avant cette thèse, les résultats de la recherche d'information dans le catalogue CISMeF étaient affichés uniquement par ordre chronologique. Une fois les documents restitués ayant des représentations correspondantes à celle de la requête de l'utilisateur, ils sont affichés du plus récent au plus ancien (date de publication). Ce critère d'affichage est inspiré de la stratégie de PubMed.

[103] Moteur de recherche DirectHit. URL : www.directhit.com
[104] Le portail de recherche NorthernLight. URL : www.northernlight.com

Pendant cette thèse, nous avons introduit de nouvelles heuristiques permettant un classement plus pertinent du résultat de la recherche. Ce tri, qui prend en compte les ressources restituées indexées manuellement mais aussi celles indexées automatiquement, repose sur un calcul de pertinence. Cette pertinence est fonction linéaire du nombre de descripteurs indexant la ressource et/ou présents dans le titre et de leur pondération (majeur/mineur).

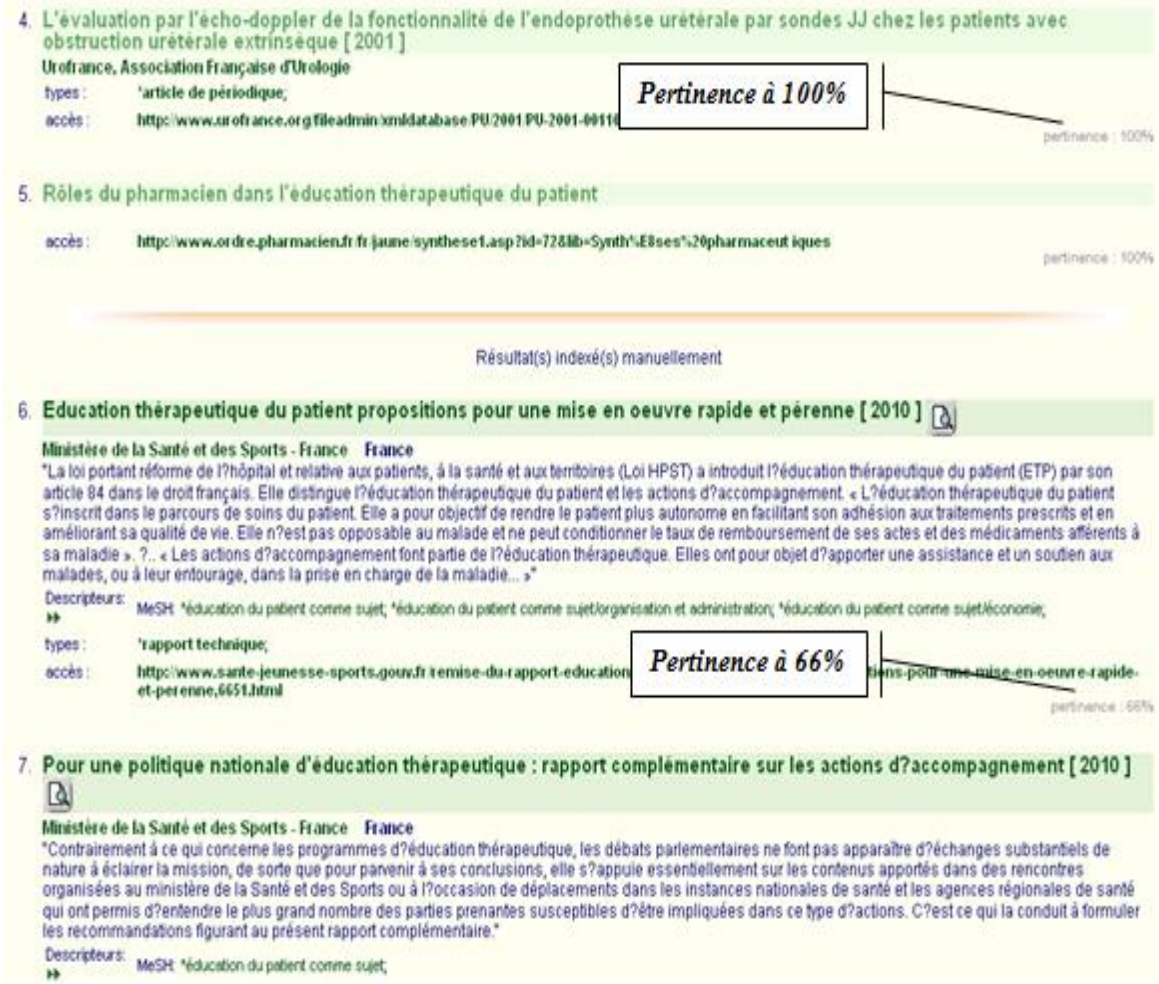


Figure 5.2. Classement du résultat de la recherche d'information selon la pertinence des documents restitués

CONCLUSION

À travers ce chapitre, nous avons relaté la nouvelle approche de l'équipe CISMeF fondée sur une recherche d'information multi-terminologique grâce aux différentes terminologies médicales intégrées dans le système d'information.

Les améliorations qui ont été apportées au catalogue CISMeF ont été rendues possible grâce à la modification de stratégie de recherche et à la mise en application les résultats de l'étude effectuée concernant la désuffixation.

Notre premier souci était d'adapter le catalogue CISMeF (Sakji et al. 2009a) aux besoins et aux connaissances terminologiques des utilisateurs, dont le nombre ne cesse d'augmenter depuis sa création en 1995.

À notre connaissance et jusqu'à aujourd'hui, une recherche d'information multi-terminologique, dans un site web de santé, est appliquée pour la première fois au sein de notre catalogue CISMeF.

CHAPITRE 6

TRAVAUX CONNEXES A LA THESE DANS LE CADRE DU PROJET PSIP

INTRODUCTION

Dans ce chapitre, nous décrivons les travaux connexes aux principaux thèmes de la thèse, notamment le passage du monde mono-terminologique vers l'univers multi-terminologique, la recherche d'information multi-terminologique et l'indexation automatique bi-terminologique des médicaments. Toutefois, ils restent au centre du domaine de la recherche d'information multi-terminologique. Notre participation à ces travaux a donné suite à d'autres perspectives prometteuses pour améliorer l'indexation et la recherche d'information médicale.

6.1 INTEGRATION DE NOUVELLES TERMINOLOGIES POUR F-MTI

Le F-MTI (French Multi-Terminology Indexer) a été conçu afin d'indexer les dossiers médicaux en utilisant plusieurs terminologies médicales à savoir la CIM-10, la CCAM, le thésaurus MeSH, la terminologie interne de la société Vidal ainsi que la nomenclature SNOMED (Pereira et al., 2009).

Dans le cadre du projet PSIP, nous étions amenés à enrichir cet outil afin d'améliorer l'indexation des comptes-rendus médicaux pour l'extraction et l'exploitation des données. Pour ce faire, nous avons intégré des terminologies médicales dédiées aux médicaments : la classification ATC (N=5.514), les noms commerciaux et la Dénomination Commune Internationale[105] des médicaments (N=22.662) ainsi que les concepts chimiques supplémentaires (N=7.104) et les actions pharmacologiques du MeSH traduits en français par l'équipe CISMeF et le thésaurus Orphanet pour les maladies rares (N=7.421).

[105] La dénomination commune internationale est utilisée pour faciliter l'identification des substances pharmaceutiques ou les ingrédients pharmaceutiques actifs. La dénomination commune est connue comme le nom générique des médicaments

Pour l'intégration des terminologies, nous avons eu besoin de formater leurs structures selon le dictionnaire déjà établi durant la thèse de Suzanne Pereira sans, pour autant, perdre les informations utiles de chaque terminologie. En effet, comme nous l'avons mentionné précédemment chacune est présente selon un format spécifique.

Ensuite, la principale tâche d'optimisation du temps de réponse a été réalisée par un ingénieur de l'équipe CISMeF qui a permis de diminuer celui-ci, d'une manière considérable. Dans le cadre de PSIP, nous avons lancé le F-MTI sur 4.000 comptes-rendus, le temps de traitement était d'environ 2 heures (soit 1,9 secondes par compte rendu au lieu de 45 secondes).

6.2 RECHERCHE D'INFORMATION SEMANTIQUE : APPLICATION DE SPARQL

Le but de cette section est de décrire une application qui a marqué notre recherche. Nous présentons un *« proof of the concept »* de la recherche d'information, en utilisant le langage SPARQL et, en particulier, l'implémentation faite par Oracle [106].

6.2.1 LE FORMAT RDF

Dans la cadre du web sémantique, le Consortium du World Wide Web (W3C) chargé de développer des technologies pour le Web, a validé une application du format XML pour la description du contenu sémantique, appelé RDF (Ressource Description Framework). RDF est un formalisme basé sur un modèle sémantique de graphes étiquetés et orientés. RDF est basé sur une relation de métadonnées sous la forme (propriété, valeur) qui décrivent une description des ressources. Ainsi, RDF décrit le graphe sous la forme d'un ensemble de triplets {ressource, propriété, valeur}. Les ressources sont des entités d'informations pouvant être référencées par un nom symbolique ou un identifiant (par exemple un URI : Unique Resource Identifier). Les propriétés sont les étiquettes des arcs orientés reliant un premier nœud étiqueté par une ressource à un second nœud qui peut être, soit une valeur atomique, soit une autre ressource.

Considérons cette phrase « Quatre-vingt treize *est un roman de* Victor Hugo publié en 1874, ayant pour thème la révolution française » exprimée en langage naturel. Une telle description peut être analysée en plusieurs phrases mettant en relief la paire (propriété, valeur) appliquée à un sujet, en d'autres termes une métadonnée et sa valeur :

1. "Quatre-vingt treize est un roman"
2. "Quatre-vingt treize est écrit par l'auteur Victor Hugo"
3. "Quatre-vingt treize est publié en 1874"
4. "Quatre-vingt treize a comme thème la révolution française"

La forme abstraite en triplets s'écrit sous cette forme :

1. (Quatre-vingt treize, type, roman)
2. (Quatre-vingt treize, auteur, Victor Hugo)

[106] Ces outils d'Oracle nous ont été fournis dans le cadre du projet PSIP dont Oracle est partenaire.

3. (Quatre-vingt treize, année de publication, 1874)
4. (Quatre-vingt treize, thème, révolution française)

Pour avoir ces informations sous la forme du RDF, il est important que le sujet en commun soit identifié par un URI, comme étant un identifiant unique. Un tel URI est présent, par exemple, dans la base de données DBpedia[107], qui fournit des descriptions RDF concernant les sujets des articles de Wikipedia. Le roman *Quatre-vingt treize* est identifié par l'URI suivant : http://dbpedia.org/resource/Ninety-Three. Par convention, cette même représentation peut être décrite comme suit : *dbpedia: Quatre-vingt treize*.

Le premier triplet définit le type du sujet. La valeur du « roman » est identifiée par la base DBpedia par l'URI: *http://dbpedia.org/class/yago/Novel106367879,*(*yago: Novel106367879).* Les types utilisés par DBpedia font référence à l'ontologie Yago ontology, qui est un représentant du vocabulaire générique Wordnet, dans lequel 106367879 identifie le concept "roman".

dbpedia: Quatre-vingt treize *rdf:type* *yago:Novel106367879*

6.2.2 APPLICATION DU FORMAT RDF AU CATALOGUE CISMEF

Le langage d'interrogation du RDF est basé sur la structure des triplets et la sémantique des vocabulaires. Parmi ces langages, nous pouvons citer SPARQL[108] considéré comme un langage standard de requêtes. SPARQL permet l'interrogation du graphe sémantique en sélectionnant les ressources qui répondent à une partie de la structure.

Pour la « preuve de concept » de notre recherche, nous avons transformé la base de données CISMeF en un graphe RDF, constitué d'un ensemble de triplets RDF qui décrivent les ressources intégrées dans le catalogue, ainsi que quelques terminologies du back-office afin de constituer une partie de l'univers multi-terminologique.

Notre première expérience a commencé avec Sésame[109], un serveur en libre accès, de stockage, d'inférence et interrogation des données RDF[110].

Ensuite, une collaboration avec Oracle, nous a permis d'utiliser les outils sémantiques d'Oracle tels que Joseki un moteur http open source qui supporte le langage SPARQL et les requêtes SPARQL permettant d'accéder aux modèles RDF Oracle stockés dans la base de données Oracle 11g[111].

Ainsi, après la construction de notre base de données sémantique et l'installation d'Oracle WebLogic[112], nous pouvons interroger nos données avec Joseki.

---

[107] DBpedia, querying Wikipedia as a data base. URL: http://wiki.dbpedia.org
[108] Sparql Query Language for RDF. URL: http://www.w3.org/TR/rdf-sparql-query
[109] URL : http://www.openrdf.org/
[110] Se référer à l'Annexe D, pour un exemple de requête
[111] RDF Semantic Data Management Using the Oracle Spatial 11*g* Option. URL: http://www.oracle.com/technology/obe/11gr1_db/datamgmt/nci_semantic_network/nci_Semantics_les01.htm
[112] Oracle WebLogic Suite 11g. URL : http://www.oracle.com/appserver/weblogic/weblogic-suite.html

Grâce au graphe RDF et les outils d'interrogation SPARQL, nous pouvons, par exemple, avoir les ressources décrites par le descripteur SNOMED « pression cardiovasculaire ».

La requête SPARQL s'écrit comme suit :

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX cismef: <http://www.chu-rouen.fr/cismef#>
    select ?s ?f
     where {
    ?s      cismef:decritPar      ?d.
    ?d      cismef:appartientA    cismef:Terminologie_SNOMED.
    ?d      rdfs:label            "pression cardiovasculaire"@fr.
    }
```

Vu la structure hiérarchique des terminologies médicales, notamment la SNOMED que nous avons utilisée pour notre exemple, nous pouvons retrouver aussi toutes les ressources indexées par les concepts SNOMED qui subsument « pression cardiovasculaire », à savoir « tension artérielle », « pression veineuse »…

```
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:publishing="http://www.mondeca.com/system/publishing#"
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX skos:<http://www.w3.org/2004/02/skos/core#>
PREFIX cismef: <http://www.chu-rouen.fr/cismef#>
    select ?s ?f
     where {
    ?s      cismef:decritPar      ?d.
    ?d      cismef:appartientA    cismef:Terminologie_SNOMED.
    ?d      rdfs:label            "pression cardio vasculaire"@fr.
    ?d      publishing:BT         ?f.
    }
```

Ce travail sur SPARQL sera poursuivi pendant six mois en 2011 dans le cadre d'un post-doc au sein de l'équipe CISMeF.

6.3 INDEXATION DES DOSSIERS MEDICAUX : ADAPTATION DE L'OUTIL DU PR PETER ELKIN

Le professeur Peter Elkin (Mount Sinaï School of Medicine (MSSM)) est un des six membres de l'advisory Board du projet PSIP. Les sujets de recherche de son équipe et de l'équipe

CISMeF étant très proches, il m'a proposé de passer trois mois au MSSM, NYC à la marge du projet PSIP. L'objectif était d'appliquer ses outils à une autre langue, en l'occurrence le français.

Dans le cadre de la tâche « semantic mining » du projet PSIP, (Elkin et al. 2008) présente ses travaux de recherche concernant la détection des maladies et des troubles dans les dossiers médicaux. Son système repose sur une indexation fondée sur les concepts de la nomenclature SNOMED CT (Systematized Nomenclature of Medicine Clinical Terms), une terminologie médicale couvrant la plupart des domaines de l'information clinique, tels que *les maladies*, *les résultats*, *les procédures*, *les micro-organismes*, *les produits pharmaceutiques* etc. Cette terminologie est la plus utilisée aux Etats-Unis.

L'identification des concepts dans les dossiers médicaux, comme étant des assertions négatives (tel que : no evidence of pneumonia) ou des assertions positives, est le but ultime du médecin lui permettant de mieux comprendre les implications des textes médicaux.

Grâce à notre collaboration avec l'équipe du professeur Elkin, nous avons essayé de mettre au point ce système pour nos dossiers médicaux d'autant que nous avons déjà eu une première expérience d'indexation des dossiers médicaux, avec quelques terminologies médicales durant la thèse de Suzanne Pereira (Pereira, 2008) et une continuation de ce travail avec la thèse en cours d'Ahmed Diouf Dirieh Dibad.

Pour ce faire, la tâche principale, durant ce stage de trois mois aux Etats-Unis, a été de comprendre le fonctionnement du système afin de permettre une adaptation française (Sakji et al., 2010a).

Le traitement de la version française du parseur s'est déroulé selon ces quatre étapes :

- ✓ la création du modèle du langage : cette phase nous a permis de détecter les différentes formes d'assertions : négatives, positives et incertaines, afin de les intégrer dans le processus de traitement ;
- ✓ la construction du serveur terminologique : étant donné que la SNOMED CT n'est pas disponible en français, nous avons utilisé la CIM-10. La correspondance qui a été créée entre les deux terminologies ce qui nous a permis d'effectuer une indexation équivalente dans les dossiers médicaux français traduits en anglais ;
- ✓ l'adaptation de la procédure du désuffixation : nous avons utilisé l'algorithme de Lucene vu ses performances par rapport à d'autres algorithmes de désuffixation ;
- ✓ la traduction des synonymes et des expressions régulières impliquées dans le traitement.

CONCLUSION

Dans ce chapitre, nous avons décrit les travaux connexes à cette thèse, dans le cadre d'une amélioration de la recherche d'information médicale multi-terminologique dans le catalogue CISMeF.

Par ailleurs, mon expérience aux Etats-Unis, m'a permise d'exploiter d'autres structures d'application, notamment la recherche d'information dans les dossiers médicaux des patients. En effet, repérer les principaux concepts d'indexation et les différencier en assertions négatives et positives étaient les principales tâches pour améliorer l'exploitation des dossiers médicaux. Ce champ de recherche est parmi les nouveaux thèmes abordés par l'équipe CISMeF, et que nous développerons prochainement.

CHAPITRE 7

PERSPECTIVES

À travers ce chapitre, nous mettons en relief nos perspectives et nos projets de recherche en continuation avec le travail concrétisé pendant cette thèse.

Pour chacune de nos approches, des améliorations sont nécessaires pour pallier les faiblesses identifiées et enrichir les approches et les méthodes appliquées.

- **Amélioration des travaux de la thèse**

En effet,

L'étude réalisée pour améliorer la recherche d'information par extension MeSH-ATC nous donne des perspectives prometteuses pour consolider l'approche de l'indexation par la classification ATC au sein du PIM. Ceci nous permettra de faire face aux quelques problèmes dus à l'attribution du bon code ATC d'indexation aux ressources. Les améliorations à moyen terme vont concerner :

- ✓ la prise en compte du contexte de la substance chimique lors de l'indexation du corpus du PIM par la classification ATC. Ceci serait appliqué en se référant aux niveaux supérieurs (1^{er}, $2^{ème}$, $3^{ème}$ et $4^{ème}$ niveaux) de la substance chimique elle-même ;
- ✓ l'indexation des ressources du PIM avec l'ATC multiple et le calcul du score des codes ATC candidats afin d'enlever l'ambiguïté détectée dans certains cas.

La limite de l'étude réalisée pour comparer la valeur ajoutée de la multi-terminologie était, en effet, le nombre de requêtes lancées dans le catalogue CISMeF, due à la validation manuelle des résultats. Comme depuis la réalisation de cette étude, nous avons enrichi notre serveur terminologique, nous pouvons refaire l'étude avec, notamment, un nombre plus important de requêtes. Ainsi, nous pouvons apporter les améliorations nécessaires à notre algorithme de recherche.

- **Pistes de réflexion et applications**

Le post-doc de six mois que je débuterai à la fin de ma thèse se focalisera sur la recherche d'information multi-terminologique dans un dossier électronique du patient. Je collaborai avec Ahmed Diouf Dirieh Dibad qui a débuté ses travaux de recherche sur ce sujet, mais aussi avec Tayeb Merabti (post Doc CISMeF depuis juin 2010), spécialisé en interopérabilité sémantique et Julien Gros jean, ingénieur de recherche. Pour cela nous nous concentrerons sur

le langage SPARQL, une expérience (« preuve de concept »), déjà faite dans le catalogue CISMeF.

Dans le cadre de ces travaux, un modèle de données générique a été déjà conçu pour représenter un dossier électronique de patient (DEP) dans un but de recherche d'information mais aussi pour des fins connexes comme l'exploration (une vue synthétique de l'historique du patient) et la classification des dossiers patients, etc.

Pour indexer les dossiers médicaux, F-MTI a été employé, jusqu'à présent, en utilisant quelques terminologies médicales telles que la CCAM, le MeSH, la SNOMED et la CIM-10.

Les Perspectives sont :

- enrichir le dictionnaire de données de F-MTI ;
- mettre en place les outils et les méthodes pour la mise en place d'un prototype de système de recherche d'information multi-terminologique dans un dossier de santé ;
- pratiquer le benchmarking entre SPARQL et les outils CISMeF ;
- exploiter la recherche d'information sur un ensemble de dossiers de santé ;
- explorer les dossiers de santé (résumé du dossier médical).

D'autre part, la collaboration avec le professeur Peter Elkin continue, dans le but d'améliorer la version française de son système (MCVS : Multi-threaded Clinical Vocabulary Server) en intégrant la version française de la SNOMED CT (au lieu de l'indexation par la CIM-10). Une comparaison entre MCVS et le F-MTI sera réalisée, afin d'apporter les améliorations nécessaires à l'un ou à l'autre.

CONCLUSION GENERALE

Notre problématique initiale était de mettre au point un modèle et une stratégie de recherche permettant une recherche d'information multi-terminologique appliquée à un site médical. La nouveauté de ce travail a été de prendre en compte le contexte et les connaissances des utilisateurs.

Pour ce faire, nous avons conçu et mis au point un modèle générique multi-terminologique au sein du back-office CISMeF fondé précédemment sur le thésaurus MeSH uniquement. La généricité du modèle nous a permise, par la suite, d'enrichir notre serveur terminologique à chaque fois que nous disposons et que nous avons besoin d'une nouvelle terminologie médicale. En effet, grâce au modèle, nous avons participé à la mise en œuvre du Portail Terminologique de Santé (PTS), un point d'accès vers une grande panoplie des terminologies. Ce portail constitue une plateforme pour rassembler ces dernières dans une même structure sans se soucier ni de leurs gestion ni de leurs mise à jour.

La migration vers l'univers multi-terminologique fondé sur plusieurs terminologies médicales s'est illustrée, dans un premier temps, au sein du Portail d'Information sur les Médicaments (PIM) par la mise au point une indexation automatique par la classification ATC, outre l'indexation par le thésaurus MeSH. Cela nous a permis d'avoir une indexation et une recherche d'information bi-terminologique. Nous avons conçu le PIM dans le cadre du projet PSIP afin de se restreindre au domaine médicamenteux. Le PIM a vu un succès progressif auprès des professionnels de santé qui ont un centre d'intérêt plus particulier pour les substances chimiques et les médicaments. En plus, l'affichage hiérarchique des informations de la substance chimique peut être considéré une bonne pédagogie pour les étudiants dans la mesure où ce choix permet de contextualiser l'information d'une part et de fournir des informations complémentaires tels que les organes sur lesquels la substance chimique agit, ou encore ses actions pharmacologiques et thérapeutiques.

Le PIM devrait vraisemblablement passer en accès libre à la fin du projet PSIP.

L'étude présentée à la fin du quatrième chapitre sur l'extension MeSH-ATC pour la recherche d'information a confirmé notre théorie, qu'en cas de confusion, il est recommandé de contextualiser la substance chimique.

Au sein du catalogue CISMeF, nous avons implanté notre nouvel algorithme avec la nouvelle structure multi-terminologique de la base de données. L'algorithme se différencie de ce qui précède par une recherche plus exhaustive à travers toutes les terminologies possibles qui représentent au mieux les ressources du catalogue CISMeF.

Les résultats de l'étude que nous avons menée pour mesurer la valeur ajoutée de la multi-terminologie par rapport à la mono terminologie, nous a révélés une amélioration globale de 15% et une satisfaction plus large dans les rangs des utilisateurs dans la mesure où chacun retrouve la (ou les) terminologie(s) qu'il maîtrise au mieux.

Ainsi, grâce à l'indexation automatique par la classification ATC des ressources du PIM et avec les différentes terminologies médicales disponibles en français pour les ressources de CISMeF, nous avons mis en pratique les termes de recherche avec leurs contextes d'utilisation. Ceci permet d'améliorer le résultat de la recherche d'information et la qualité des SRI, un reproche avancé par Blair (Blair, 1990) où il met l'accent sur la complexité des systèmes de recherche d'information pour fournir un bon résultat dans la mesure où ils nécessitent un langage précis pour mettre les termes dans leurs contextes.

Notre objectif, dans le futur proche, est d'améliorer les deux approches d'indexation et de recherche d'information, afin de mieux répondre aux besoins des utilisateurs.

BIBLIOGRAPHIE

Abdou, S., and Savoy J. "Searching in Medline : Query expansion and manual indexing evaluation." *Information processing and management,*, 2008: 781-789.

Aymé, S., Urbero B., Oziel D., Lecouturier E., and Biscarat AC. "Information on rare diseases: the Orphanet project." *Rev Med Interne*, 1998: 376S-377S.

Baeza-Yates, R., and Riberto-Neto B. *Modern Information retrieval.* New York: Addison-Wesley, 1999.

Baron, S., and Linden M. "The role of the International Classification of Functioning, Disability and Health, ICF' in the description and classification of mental disorders." *European Archives of Psychiatry and Clinical Neuroscience*, 2008: 81-85.

Bates, DW., Evans RS., Murff H., Stetson PD., Pizziferri L., and Hripcsak G. "Detecting Adverse Drug Events using Information Technology." *The Journal of the American Medical Informatics Association*, 2003: 115-128.

Baziz, M. *Indexation conceptuelle guidée par ontologie pour la recherche d'information.* Thèse de doctorat, Toulouse, France: Institut de Recherche en Informatique de Toulouse, 2005.

Bertillon, J. "Classification of the causes of death (abstract)." In *Transactions of the 15th International Congress on Hygiene Demography*. Washington, 1912.

Beuscart-Zephir, MC., Bjorn B., Cacciabue PC., and Koutkias V. "Definitions of core concepts in PSIP." Rapport interne, 2009.

Blair, DC. *Language and representation in information retrieval.* New York, NY, USA : Elsevier North-Holland, Inc, 1990.

Bodenreider, O. "The Unified Language System (UMLS): integrating biomedical terminology." *Nucleic Acids Research*, 2004: 267-270.

Borst, WN. *Construction of Engineering Ontologies.* Enschede: University of Tweenty, 1997.

Boughanem, M. «Introduction à la recherche d'information.» Dans *Recherche d'information: état des lieux et perspectives*, 19-44. Hermès-Lavoisier, 2008.

Boughanem, M. *Les Systèmes de Recherche d'Information: d'un modèle classique à un modèle connexionniste.* Thèse de Doctorat, Toulouse, France: Université Paul Sabatier, 1992.

Boughanem, M., and Soulé-Dupuy C. " A Connexionist Model for Information Retrieval." *DEXA*, 1992: 260-265.

Boughanem, M., et Savoy J. *Recherche d'information: état des lieux et perspectives.* Hermès-Lavoisier, 2008.

Boughanem, M., et Tamine L. «Connexionisme et génétique pour la recherche d'information.» Dans *Les systèmes de recherche d'informations*, 77-99. Hermès, 2004.

Bourda, Y., and Hélier M. "Applying IEEE Learning Object Metadata to Publishing Teaching Programs." *ED-MEDIA.* Seattle, 1999.

Bousquet, C., Henegar C., Lillo-le Louet A., et Jaulent MC. «Apport d'une modélisation ontologique pour la détection du signal en pharmacovigilance.» *15es journées francophones d'ingénierie des connaissances.* Lyon, 2004. 187-198.

Boyer C., Gaudinat A., Baujard V., Geissbühler A. "Health on the Net Foundation: assessing the quality of health web pages all over the world." *Studies in health technology and informatics (Stud Health Technol Inform)*, 2007: 1017-1021.

Bradford, R. "Relationship Discovery in Large Text Collections Using Latent Semantic Indexing." *In Proceedings of the 4th Workshop on Link Analysis, Counterterrorism and Security, SIAM Data Mining Conference,*. Bethesda, MD, 2006. 20-22.

Bratsas, C., Koutkias V., Kaimakamis E., Bamidis P., and Maglaveras N. "Ontology-based vector space model and fuzzy query expansion to retrieve knowledge on medical computational problem solutions." *International Conference of the IEEE Engineering in Medicine and Biology Society.* 2007. 3794-3797.

Brin, S., and Page L. "The anatomy of a large-scale hypertextuel web search engine." *Proceedings of the WWW7.* Amsterdam: Elsevier, 1998. 107-117.

Brini, AH. *Un Modèle de Recherche d'Information basé sur les réseaux possibilistes.* Rapport de thèse, Université Paul Sabatier de Toulouse, 2005.

Brown, EG. "Effects of coding dictionary on signal generation: a consideration of use of MedDRA compared with WHO-ART." *Drug Safety*, 2002: 445-52.

Chaumier, J. *Le traitement linguistique de l'information.* Paris: Entreprise moderne d'édition, 1988a.

Chaumier, J. *Travail et méthodes du/de la documentaliste: connaissances du problème, applications pratiques.* Paris: ESF, 1988b.

Chazard, E., Preda C., Merlin B., Ficheur G., et Beuscart R. «Détection et prévention des effets indésirables liés aux médicaments par data-mining.» *Ingénierie et Recherche BioMédicale*, 2009: 192-196.

Chevallet, JP. «Modélisation logique pour la recherche d'information.» Dans *Les systèmes de recherche d'information*, 105-138. Hermes, 2004.

Cornet, R., and de Keizer N. "Forty years of SNOMED: a literature review." *BMC Medical Informatics and Decision Making*, October 2008: online October 27.

Coté, RA. *SNOMED: Systematized Nomenclature of Medicine (2 volumes).* College of American Pathologists , 1986.

Courtois, M P., and Berry MW. "Results-ranking in Web search engines." *Online*, 1999: 39-46.

Crestani, F. "Implementation and evaluation of a relevance feedback device based on neural networks." In *From Natural to Artificial neural Computation: International Workshop on*

Artificial Neural Networks, volume 930 of Lecture Notes in Computer Science, 597–604. Springer-Verlag, 1995.

Crestani, F., and Lalmas M. "Logic and uncertainty in information retrieval." In *Lectures on information retrieval*, 179-206. Springer-Verlag New York, Inc., 2001.

Crestani, F., Lalmas M., van Rijsbergen CJ., and Campbell L. ""Is This Document Relevant? ... ProbablyProbably": A Survey of Probabilistic Models in Information Retrieval." *ACM Computing Surveys*, 1998: 528-552.

Croft, WB., and Harper DJ. "Using probabilistic models of document retrieval without relevance information." *Journal of Documentation*, 1979: 285-295.

Cuggia, M., Darmoni S., Garcelon N., Soualmia L., and Bourde A. "Doc'UMVF: tow search tools to provide quality-controlled teaching resources in French to students and teachers." *International Journal of Medical Informatics (IJMI)*, 2007: 357-362.

Darmoni S., Sakji S., Grosjean J., Beuscart MC. "Metamodel, Terminologies (for applicable data repositories of the scope of the PSIP project)." Deliverable of the PSIP project, 2010.

Darmoni, S., Sakji S., Pereira S., and Kergourlay I. *Final results of semantic mining.* Internal report, PSIP project, 2010.

Darmoni, S., Sakji S., Pereira S., and Kergourlay I. "First results of semantic mining." Deliverable of the PSIP project, 2009.

Darmoni, SJ., Amsallem E., Haugh MC., Lukacs B., Chalhoub C., and Leroy JP. "Level of evidence as a future gold standard for the content quality of health resources on the internet." *Methods of Information in Medicine*, 2003: 200-225.

Darmoni, SJ., et al. "Affiliation of a resource type to a MeSH term in a quality-controlled health gateway." *12th World Congress on Health and Medical Informatics (Medinfo).* 2007. 407-411.

Darmoni, SJ., Leroux V., Thirion B., Santamaria P., and Gea M. "Netscoring : critères de qualité de l'information de santé sur internet." *Les enjeux des industries du savoir*, 1999: 29-44.

David, C., Giroux L., Bertrand-Gastaldy S., and Lanteigne D. "Indexing as problem solving- a cognitive approach to consistency." *Canadian Association of Information Sceince (CAIS/ACSI).* 1995.

De Loupy, C. «L'apport de connaissances linguistiques en recherche documentaire.» *Traitement Automatique du Langage Naturel: TALN'01* . 2001.

Deerwester, S., Dumais ST., Furnas GW., Landauer TK., and Harshman R. "Indexing by latent semantic indexing." *Journal of the American Society for Information Science* , 1990: 391–407.

Dekkers, M., and Weibel S. "State of the Dublin Core Metadata Initiative." 2003. URL: http://www.dlib.org/dlib/april03/weibel/04weibel.html (accessed July 30, 2010).

Despres, S., and Szulman S. "Réseau terminologique versus Ontologie." *Toht.* 2008. 17-34.

Dirieh Dibad, AD, Sakji S., Prieur E., Pereira S., Joubert M., and Darmoni SJ. "Recherche d'information multi-terminologique en contexte : Etude préliminaire." *13èmes Journées Francophones d'Informatique Médicale (JFIM).* 2009. 101-112.

Dittmar, PG., Stobaugh RE., and Watson CE. "The chemical abstracts service chemical registry system. I.General Design." *Journal of Chemical Information and Computer Sciences (J Chem Inf Comput Sci),* 1976: 111–121.

Douyère, M., et al. "Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway." *Health information and libraries journal (Health Info Libr J),* 2004: 253-61.

Dumais, ST. "Latent Semantic Indexing (LSI): TREC-3 Report." *Overview of the Third Text REtrieval Conference.* 1995. 219-230.

Elkin, PL., et al. "NLP-based identification of pneumonia cases from free-text radiological reports." *American Medical Informatics Association/ Annual Symposium proceedings*, 2008: 172-176.

Eysenbach, G., Yihune G., Lampe K., Cross P., and Brickley D. "A metadata vocabulary for self- and third-party labeling of health web-sites: Health Information Disclosure, Description and Evaluation Language (HIDDEL)." *American Medical Informatics Association/ Annual Symposium proceedings*, 2001: 169-173.

Family Medicine Research Center. 2010. http://www.fmrc.org.au (accessed Septembre 01, 2010).

Gaudinat, A., et al. "Health search engine with e-document analysis for reliable search results." *International Journal of Medical Informatics (IJMI),* 2006: 73-85.

Gay, CW., Kayaalp M., and Aronson R. "Semi-automatic indexing of full text biomedical articles." *American Medical Informatics Association/ Annual Symposium proceedings*, 2005: 271-275.

Gehanno, JF., Kerdelhue G., Sakji S., Massari P., Joubert M., and Darmoni SJ. "Relevance of Google-customized search engine vs. CISMeF quality-controlled health gateway." *Studies in health technology and informatics (Stud Health Technol Inform),* 2009: 312-316.

Gehanno, JF., Thirion B., and Darmoni SJ. "Evaluation of meta-concepts for information retrieval in a quality-controlled Health Gateway." *American Medical Informatics Association/ Annual Symposium proceedings*, 2007: 269-273.

Greenwood, M. "Medical statistics from Graunt to Farr." Cambridge, 1948.

Gruber, T. "A translation Approach to portable ontology specification." *Knowledge Acquisition,* 1993: 199-220.

Hanser, S., Zaiss A., and Schulz S. "Comparison of ICHI and CCAM basic coding system." *Studies in health technology and informatics (Stud Health Technol Inform),* 2006: 795-800.

Hatcher, E., and Gospodnetic O. *Lucene in Action.* Manning Publications, 2004.

Hull, DA. "Stemming algorithms : A case study for detailed evaluation." *Journal of the American Society of Information Science*, 1996: 70-84.

ISO, 1087-1:2000. "Terminology work-vocabulary- part 1: theory and application." 2000.

Jansen, BJ., and Spink A. "How are we searching the World Wide Web? A comparison of nine search engine transaction logs." *Information Processing and Management*, 2006: 248-263.

Joubert, M., A., Gaudinat, Boyer C., Fieschi M., and HON Foundation Council members. "WRAPIN: a tool for patient empowerment within EHR." *Studies in health technology and informatics (Stud Health Technol Inform)*, 2007: 147-151.

Joubert, M., Aymard S., Fieschi D., and Fieschi M. "ARIANE: un moteur de recherche de deuxième génération dans le domaine de la santé." *Informatique et santé*, 2002: 73-80.

Joubert, M., Dufour J., Aymard S., Falco L., Staccini P., and Fieschi M. "Le projet CoMeDIAS: Accès à des bases de données hétérogènes au moyen de services internet." *Informatique et santé*, 2003: 200-205.

Keselman, A., Browne AC., and Kaufman DR. "Consumer health information seeking as hypothesis testing." *Journal of the American Medical Informatics Association (JAMIA)*, 2008: 484-495.

Koch, T. "Quality-controlled subject gateways: definitions, typologies, empirical overview, Subject gateways." *Online Information Review*, 2000: 24-34.

Kwok, KL. "A neural network for probabilistic information retrieval." *Proceedings of ACM SIGIR, Conference on Research and development in Information Retrieval.* 1989. 21-30.

Lamy, JB., Duclos C., et Venot A. «De l'analyse d'un corpus de texte à la conception d'une interface graphique facilitant l'accès aux connaissances sur le médicament.» *20ème Journées Francophones d'Ingénierie des Connaissances: Actes d'IC.* 2009. 265-276.

Lamy, JB., et al. "Towards iconic language for patient records, drug monographs, guidelines and medical search engines." *Studies in health technology and informatics (Stud Health Technol Inform).* 2010. 156-160.

Lardy, JP. "Méthodes de tri des résultats des moteurs de recherche." 2000. URL: http://halshs.archives-ouvertes.fr/docs/00/06/20/56/HTML/ (accessed July 29, 2010).

Laskri, T., and Meftouh K. "Extraction automatique du sens d'une phrase en langue Française par une approche neuronale." *JADT 2002 : 6es Journées internationales d'Analyse statistiques des Données Textuelles.* 2002. 413-422.

Lassila, O., and Mr Guiness D. "The role of frame-based representation on the Semantic Web." Technical report KSL-01-02, 2001.

Le Loarer, P. " Indexation automatique, recherche d'information et évaluation." *Collection Sciences de l'information. Série Etudes et techniques*, 1994: 149-201.

Lefevre, P. *La recherche d'information : du texte intégral au thésaurus.* Editions Hermès, 2000.

Leininger, K. "Interindexer consistency in psycINFO." *Journal of Librarianship and Information Science*, 2000: 4-8.

Leonard, LE. "Inter-indexer consistency studies, 1954-1975 : a review of the literature and summary of study results." *University of Illinois Graduate School of Library Science Occasional Papers*, 1977.

Lerat, P. *Les langues spécialisées.* . Paris: Presses Universitaires de France., 1995.

Letord C., Sakji S., Pereira S., Dahamna B., Kergourlay I., Darmoni SJ. "Recherche d'information multi-terminologique : application à un portail d'information sur le médicament en Europe." *Ingénierie et Recherche Biomédicale / BioMedical Engineering and Research (IRBM)*, 2008: 350-356.

Lewis, DD. "An evaluation of phrasal and clustered representations on a text categorization task." *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval.* Copenhagen, Denmark: ACM, 1992. p.37-50.

Lu, Z., Kim W., and Wilbur WJ. "Evaluation of query expansion using MeSH in PubMed." *Information Retrieval,*, 2009: 69-80.

Luhn, HP. "The automatic creation of literature abstracts." *IBM Journal of research and development,*, 1958: 159-165.

Lussier, YA., Rothwell DJ., and Côté RA. "The SNOMED model: a knowledge source for the controlled terminology of the computerized patient record." *Methods of Information in Medicine*, 1998: 161-164.

Maniez, J. *Les langages documentaires et classificatoires : conception, construction et utilisation dans les systèmes documentaires.* Paris : Éditions d'organisation, 1987.

Markey, K. "Interindexer consistency tests : a literature review and report of a test of consstency in indexing visual materials." *Library and Information Science Research*, 1984: 155-177.

Maron, M., and Kuhns J. "On relevance, probabilistic indexing and information retrieval." *Journal of the Association for computing Machinery*, 1960: 216-244.

Martinet, J., Chiaramella Y., et Mulhem P. «Un modèle vectoriel étendu de recherche d'information adapté aux images.» *20ème Congrès INFORSID'02 (Informatique des Organisations et Systèmes d'Information et de Décision).* Nantes, France, 2002. 337-348.

Mayer, MA., Darmoni SJ., Fiene M., Eysenbach G., Kohler C., and Roth-Berghofer T. "MedCIRCLE - modeling a collaboration for internet rating, certification, labeling and evaluation of health information on the semantic world-wide-web." *Medical Informatics Europe*, 2003: 667-672.

McCray, AT., Ide NC., Loane RR., and Tse T. "Strategies for supporting consumer health information seeking." *International Congress on Medical Informatics (Medinfo)*, 2004: 1152-1156.

Merabti, M. *Méthodes pour la mise en relations des terminologies médicales: Contribution à l'interopérabilité sémantique Inter et Intra terminologique.* Rapport de thèse, Université de Rouen, 2010.

Miller, N., Lacroix EM., and Backus JE. "MEDLINEplus: building and maintaining the National Library of Medicine's consumer health Web service." *Bulletin of the Medical Library Association (Bull Med Libr Assoc)*, 2000: 11-17.

Morel, F. «Pourquoi un dictionnaire des résultats de consultation en médicine générale?» *La Revue du praticien. Médecine générale*, 1996: 83-86.

Morimoto, T., Gandhi TK., Seger AC., Hsieh TC., and Bates DW. "Adverse drug events and medication errors: detection and classification methods." *Quality & safety in health care (Qual Saf Health Care)*, 2004: 306-314.

Mothe, J. "Search mechanisms using a neural network-Comparison with the vector space model." *4th RIAO Intelligent Multimedia Information Retrieval Systems and Management.* New York, 1994. 275-294.

Müller, H., et al. "Overview of the ImageCLEFmed 2007 Medical Retrieval and Annotation Tasks." *Working Notes of the 2007 CLEF Workshop*, 2007.

Nakache, D., et Métais E. «Evaluation : nouvelle approche avec juges.» *INFORSID'05 XXIII e congrès.* Grenoble, 2005. 555-570.

National Coordinating Council for Medication Errors Reporting and Prevention NCCMERP. "Taxonomy of Medication Errors." *Pharmacien hospitalier (Pharm hosp).* 2002.

Neches, R., Fikes RE., Finin T., Gruber TR., Senator T., and Swartout WR. "Enabling technology for knowledge sharing." *AI Magazine*, 1991: 36-56.

Névéol, A. *Automatisation des tâches documentaires dans un catalogue de santé en ligne.* Rapport de thèse, Rouen: INSA de Rouen, 2005.

Névéol, A., Mork J., Aronson A., and Darmoni S. "Evaluation of French and English MeSH indexing systems with a parallel corpus." *American Medical Informatics Association/ Annual Symposium proceedings*, 2005: 565-569.

Nie, J. «Un modèle logique général pour les systèmes de recherche d'information. Application au prototype RIME.» Rapport de thèse, Université Joseph Fourrier, 1990.

Organisation Mondiale de la Santé. *CIM-10 : Classification statistique internationale des maladies et des problèmes de santé connexes, dixième révision, volume 1.* Genève: OMS, 1993, 1335p.

Organisation Mondiale de la Santé. *Manuel de classement statistique international des maladies, traumatismes et causes de décès. Sixième révision des nomenclatures internationales de maladies et causes de décès adoptée en 1948, volume 2, index alphabétique.* Genève: OMS, 1950b.

Organisation Mondiale de la Santé. *Manuel de classement statistique international des maladies, traumatismes et causes de décès. Sixième révision des nomenclatures*

internationales de maladies et causes de décès adoptée en 1948, volume1. Genève: OMS, 1950a, 382p.

Paternostre, M., Francq P., Lamoral J., Wartel D., et Saerens M. «Carry, un algorithme de désuffixation pour le français.» Rapport Technique, Université libre de Bruxelles, 2002, http://beams.ulb.ac.be/beams/documents/carryfinal.pdf.

Pereira, S. *Indexation Multi-Terminologique de Concepts en Santé.* Rapport de thèse, Rouen: Université de Rouen, 2008.

Pereira, S., et al. «F-MTI : outil d'indexation multi-terminologique : application à l'indexation automatique de la SNOMED.» *13ème Journées Francophones d'Informatique Médicale (JFIM).* 2009. 57-67.

Porter, MF. "An algorithm for suffix stripping." *Program*, 1980: 130-137.

Prie, Y. "Sur la piste de l'indexation conceptuelle de documents. Une approche par l'annotation." *Document Numérique, numéro spécial "L'indexation"*, 2000: 11-35.

PubMed help. *How PubMed works: automatic term mapping.* 2005. URL: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=helppubmed&part=pubmedhelp#pubmedhelp.How_PubMed_works_aut (accessed July 26, 2010).

Rajashekar, TB., and Croft WB. "Combining automatic and manual index representations in probabilistic retrieval." *Journal of the American Society for Information Science*, 1995: 272-283.

Rector, AL. "Thesauri and formal classifications: Terminologies for people and machines." *Methodsof Information in Medicine*, 1998: 501-509.

Robertson, SE. "The probability ranking principle in IR." In *Readings in information retrieval*, 281-286. Morgan Kaufmann Publishers Inc., 1997.

Robertson, SE., and Sparckjones K. "Relevance weighting of search terms." *Journal of the American society for Information Science*, 1976: 129-146.

Robertson, SE., and Walker S. "Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval." *Proceedings of SIGIR 1994.* 1994. 232-241.

Runciman, W., Hibbert P., Thomson R., Van Der Schaaf T., Sherman H., and Lewalle P. "Towards an International Classification for Patient Safety: key concepts and terms." *International journal for quality in health care (Int J Qual Health Care)*, 2009: 18-26.

Safran, C. "A Concept-Based Information Retrieval Information Approach for User-oriented Knowledge Transfer." Rapport de thèse, Institute for Information Systems and Computer Media (IICM), 2005.

Sakji, S., Aufaure MA., Polaillon G., Le Grand B., et Soto M. «Une mesure de similarité contextuelle pour l'aide à la navigation dans un treillis.» *Extraction et Gestion des Connaissances (EGC).* 2008. 103-114.

Sakji, S., Darmoni S., and Elkin P. "Evaluation of a French – English Intelligent Natural Language Processor." *MedInfo*, 2010(a).

Sakji, S., et al. "Automatic indexing in a drug information portal." *Studies in health technology and informatics (Stud Health Technol Inform)*, 2009(b): 112-122.

sakji, S., Massari P., Letord C., Rollin L., Joubert M., and Darmoni S. "Evaluation of multi-terminology information retrieval in a medical catalog." *Methods of Information in Medicine* , 2010(b): soumis.

Sakji, S., Thirion B., Dahamni B., et Darmoni SJ. «Recherche des sources d'information institutionnelle de santé françaises Le site Internet CISMeF.» *Presse Médicale*, 2009(a): 1443-1450.

Salton, G. "The SMART Retrieval System: Experiments in Automatic Document Processing." *Prentice-Hall.* 1971.

Salton, G., and McGill MJ. *Introduction to modern information retrieval.* New York: McGraw-Hill, Inc., 1983.

Salton, G., Wong A., and Yang CS. "A vector space model for automatic indexing." *Commun. ACM*, 1975: 613–620.

Santé Canade,. *Santé Canada.* September 01, 2010. URL: www.sc-hc.gc.ca (accessed September 01, 2010).

Savoy, J. «Indexation manuelle et automatique : une évaluation comparative basée sur un corpus en langue française.» *COnférence en Recherche d'Information et Applications (CORIA).* 2005. 9-24.

Schmid, H. "Probabilistic part-of-speech tagging using decision trees." *International Conference on New Methods in Language Processing.* Manchester, UK, 1994.

Schmitt, E., Antier D., Bernheim C., Dufay E., Husson MC., and Tissot E. "Dictionnaire français de l'erreur médicamenteuse." 2006.

Skrbo, A., Begović B., and Skrbo S. "Classification of drugs using the ATC system (Anatomic, Therapeutic, Chemical Classification) and the latest changes." *Medicinski arhiv (Med Arh)*, 2004: 138-141.

Smeaton, AF. "Progress in the Application of Natural Language Processing to Information Retrieval Tasks." *Computer Journal*, 1992: 268-278.

Smeaton, AF. "Using NLP or NLP resources for information retrieval tasks." In *Natural Language Information Retrieval*, 99-111. 1999.

Soler, JK., Okkes I., Wood M., and Lamberts H. "The coming of age of ICPC: celebrating the 21st birthday of the International Classification of Primary Care." *Family Practice (Fam Pract)*, 2008: 312-317.

Soualmia, L. *Etude et Evaluation d'Approches Multiples d'Expansion de Requêtes pour une Recherche d'Information Intelligente : Application au Domaine de la Santé sur Internet.* Rapport de thèse, Rouen: INSA de Rouen, 2004.

Soualmia, L., Barry C., and Darmoni SJ. "Knowledge-Based Query Expansion over a Medical Terminology Oriented Ontology." *Artificial Intelligence in Medicine*, November 11, 2003: 209-213.

Soualmia, L., Dahamna B., Thirion B., and Darmoni SJ. "Strategies for health information retrieval." *Studies in health technology and informatics (Stud Health Technol Inform)*, 2006: 595-600.

Strzalkowski, T. "Natural language processing in large-scale text retrieval tasks." In *TREC*, 173-188. 1992.

Studer, R., Benjamins VR., and Fensel D. "Knowledge Engineering: Principles and Methods." *Data and Knowledge Engineering*, 1998: 161-197.

Tamine-Lechani, L., Zemirli N., et Bahsoun W. «Approche statistique pour la définition du profil d'un utilisateur de système de recherche d'information.» Dans *Information - Interaction - Intelligence*. Cépaduès Editions, 2007.

Thirion, B., Douyère M., Soualmia L., Dahamna B., Leroy JP., and Darmoni SJ. " Metadata element sets in the CISMeF quality-Controlled Health Gateway." *International Conference on Dublin Core and Metadata Applications.* Shanghai, China, 2004.

Thirion, B., Robu I., and Darmoni SJ. "Optimization of the PubMed Automatic Term Mapping." *Studies in health technology and informatics (Stud Health Technol Inform)*, 2009: 238-42.

Tricot, A. «Recherche d'information et apprentissage avec documents électroniques.» Dans *Lire, écrire, communiquer, apprendre avec Internet*, 441-462. Solal, 2006.

Vallet, D., Miriam Fernandez M., and Castells P. "An ontology-based information retrieval model." *European Semantic Web Conference (ESWC).* 2005. 455–470.

Vallez, M., and Pedraza-Jimenez R. *Natural Language Processing in Textual Information Retrieval and Related Topics.* http://www.hipertext.net. 2007.

Van Rijsbergen, CJ. "A new Theoritical Framework for Information Retrieval." *Proceedings of SIGIR-86, 9th ACM Conference on Research and Development in Information Retrieval.* Pisa, 1986. 194-200.

Van Rijsbergen, CJ. *Information Retireval.* Butterworths, 1979.

Van Slype, G. *Les langages d'indexation : conception, construction et utilisation dans les systèmes documentaires.* Paris : Éditions d'organisation, 1987.

WHO Collaborating Centre for Drug Statistics Methodology. *ATC/DDD methodology.* November 19, 2009. URL: http://www.whocc.no/atc_ddd_methodology/history/.

Wilbur, WJ., and Kim, W. "The dimensions of indexing." *American Medical Informatics Association/ Annual Symposium proceedings*, 2003: 714-719.

World Alliance & WHO Health Information Systems Department. "International Classification for Patient Safety." Statement of Purpose, 2009.

World Health Organizations,. *Classification statistique internationale des maladies et des problèmes de santé connexes, Dixième révision.* September 01, 2010. http://apps.who.int/bookorders/anglais/detart1.jsp?sesslan=1&codlan=2&codcol=15&codcch=754 (accessed September 01, 2010).

Zaiss, A., and Hanser S. "The French Common Classification of Procedures CCAM. An option for Germany." *Bundesgesundheitsblatt Gesundheitsforschung Gesundheitsschutz*, 2007: 944-52.

Zeng, Q., Crowell J., Plovnick R., Kim E., Ngo L., and Dibble E. "Assisting consumer health information retrieval with query recommendations." *Journal of the American Medical Informatics Association (JAMIA)*, 2006: 80-90.

Zipf, GK. *Human Behaviour and the Principle of Least Effort: An Introduction to Human Ecology.* Cambridge: Addison-Wesley, 1949.

Zweigenbaum, P. «Encoder l'information médicale: des terminologies aux systèmes de représentation des connaissances.» *Innovation Stratégique en Information de Santé (ISIS)*, 1999: 27-47.

Zweigenbaum, P., et al. «UMLF : construction d'un lexique médical francophone unifié.» *Actes des Journées Francophones d'Informatique Médicale (JFIM)*, 2003.

LISTE DE PUBLICATIONS

Sakji, S., Darmoni S., Elkin P. Evaluation of a French – English Intelligent Natural Language Processor. *Methods of Information in Medicine,* 2010. (soumis).

Sakji, S., Massari P., Letord C., Rollin L., Joubert M., Darmoni S. Evaluation of multi-terminology information retrieval in a medical catalog. *Methods of Information in Medicine. 2010.* (soumis).

Elkin PL., Trusko BE., Koppel R., Speroff T., Mohrer D., Sakji S., Gurewitz I., Tuttle M., Brown SH. Secondary use of clinical data. *Studies in health technology and informatics (Stud Health Technol Inform).* 2010: 14-29.

Sakji, S., Gicquel Q., Pereira S., Kergoulay I., Proux D., Darmoni SJ., Metzger MH. Evaluation of a French Medical Multi-Terminology Indexer for the Manual Annotation of Natural Language Medical Reports of Healthcare-Associated Infections. *13th International Congress on Medical Informatics.* 2010: 252-256

Merlin, B., Chazard E., Pereira S., Serrot E., Sakji S., Beuscart R., Darmoni SJ. Can F-MTI semantic-mined drug codes be used for Adverse Drug Events detection when no CPOE is available? *13th International Congress on Medical Informatics,* 2010: 1025-1029

Sakji S., Thirion B., Dahamna B., Darmoni SJ. Recherche des sources d'information institutionnelle de santé françaises Le site Internet CISMeF. *Presse Médicale*, 2009 : 1443-1450.

Darmoni, SJ., Sakji S., Pereira S., Merabti, T., Prieur E., Joubert M., Thirion B. Multiple terminologies in an health portal: automatic indexing and information retrieval. *Artificial Intelligence in Medicine,* Verona, Italy, July, Lecture Notes in Computer Science, 2009 : 255-259.

Pereira, S., Sakji S., Névéol A., Kergoulay I., Kerdelhué G., Serrot E., Joubert M., Darmoni SJ. Abstract multi-terminology indexing for the assignment of MeSH descriptors. *American Medical Informatics Association/ Annual Symposium proceedings*, 2009: 521-525.

Sakji, S., Dirieh Dibad,AD., Kergourlay I., Joubert M., Darmoni SJ. Information Retrieval in Context Using Various Health Terminologies. *International Conference on Research Challenges in Information Science IEEE,* Fez, Morocco, April, 2009 : 453-458.

Sakji S., Letord C., Dahamna B., Kergourlay I., Pereira S., Joubert M., Darmoni, SJ. Automatic indexing in a drug information portal. *Studies in health technology and informatics (Stud Health Technol Inform).* 2009: 112-122.

Sakji, S., Letord C., Pereira S., Dahamna B., Joubert M., Darmoni, SJ. Drug Information Portal in Europe: information retrieval with multiple health terminologies. *Studies in health technology and informatics (Stud Health Technol Inform)*. 2009: 497-501.

Dirieh Dibad, AD., Sakji S., Prieur E., Pereira S., Joubert M., Darmoni, SJ. « Recherche d'information multi-terminologique en contexte : Etude préliminaire ». *13ème journées francophones d'informatique médicale (JFIM)*. 2009 : 101-112.

Letord, C., Sakji S., Pereira S., Dahamna B., Kergoulay I., Darmoni SJ. « A Drug Information Portal in Europe ». *American Medical Informatics Association/ Annual Symposium proceedings.* 2009:p.931.

Letord, C., Sakji S., Pereira S., Dahamna B., Kergourlay I., Darmoni, SJ. « Recherche d'information multi-terminologique : application à un portail d'information sur le médicament en Europe ». *Ingénierie et Recherche Biomédicale / BioMedical Engineering and Research.* 2008 : 350-356.

Sakji, S. « Recherche multi-terminologique de l'information de santé sur l'Internet ». *5ème édition de la COnférence en Recherche d'Information et Applications (CORIA).* 2008 : 409-416.

ANNEXE A

❖ Description OWL du modèle de la terminologie ATC

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE rdf:RDF [
 <!ENTITY itm "http://www.mondeca.com/system/itm#">
 <!ENTITY owl "http://www.w3.org/2002/07/owl#">
 <!ENTITY rdf "http://www.w3.org/1999/02/22-rdf-syntax-ns#">
 <!ENTITY rdfs "http://www.w3.org/2000/01/rdf-schema#">
 <!ENTITY stms "http://www.chu-rouen.fr/stms">
 <!ENTITY xsd "http://www.w3.org/2001/XMLSchema#">]>

<rdf:RDF xml:base="&stms;"
      xmlns:itm="&itm;"
      xmlns:owl="&owl;"
      xmlns:rdf="&rdf;"
      xmlns:rdfs="&rdfs;">

<!-- Ontology Information -->
 <owl:Ontology rdf:about="">
  <itm:defaultLanguage xml:lang="en">fra</itm:defaultLanguage>
  <rdf:type rdf:resource="&owl;Thing"/>
  <owl:versionInfo xml:lang="fr">Ontologie STMS - Version 1.14 - 2007-12-
04</owl:versionInfo>
 </owl:Ontology>

<!-- Classes -->
 <owl:Class rdf:about="#ATCCaracteristiqueChimique">
  <rdfs:label xml:lang="fr">ATC Characteristique Chimique</rdfs:label>
  <rdfs:subClassOf rdf:resource="#ATCConcept"/>
 </owl:Class>

 <owl:Class rdf:about="#ATCConcept">
  <rdfs:subClassOf
rdf:resource="http://www.mondeca.com/system/publishing#Descriptor"/>
 </owl:Class>

 <owl:Class rdf:about="#ATCEffetPharmacologique">
  <rdfs:label xml:lang="fr">ATC Effet Pharmacologique</rdfs:label>
  <rdfs:subClassOf rdf:resource="#ATCConcept"/>
 </owl:Class>

 <owl:Class rdf:about="#ATCIndicationTherapeutique">
  <rdfs:label xml:lang="fr">ATC Indication Thérapeutique</rdfs:label>
  <rdfs:subClassOf rdf:resource="#ATCConcept"/>
 </owl:Class>
```

```
<owl:Class rdf:about="#ATCOrganeAnatomique">
 <rdfs:label xml:lang="fr">ATC Organe Anatomique</rdfs:label>
 <rdfs:subClassOf rdf:resource="#ATCConcept"/>
</owl:Class>

<owl:Class rdf:about="#ATCSubstanceChimique">
 <rdfs:label xml:lang="fr">ATC Substance Chimique</rdfs:label>
 <rdfs:subClassOf rdf:resource="#ATCConcept"/>
</owl:Class>
```

❖ Modélisation UML de la classification ATC

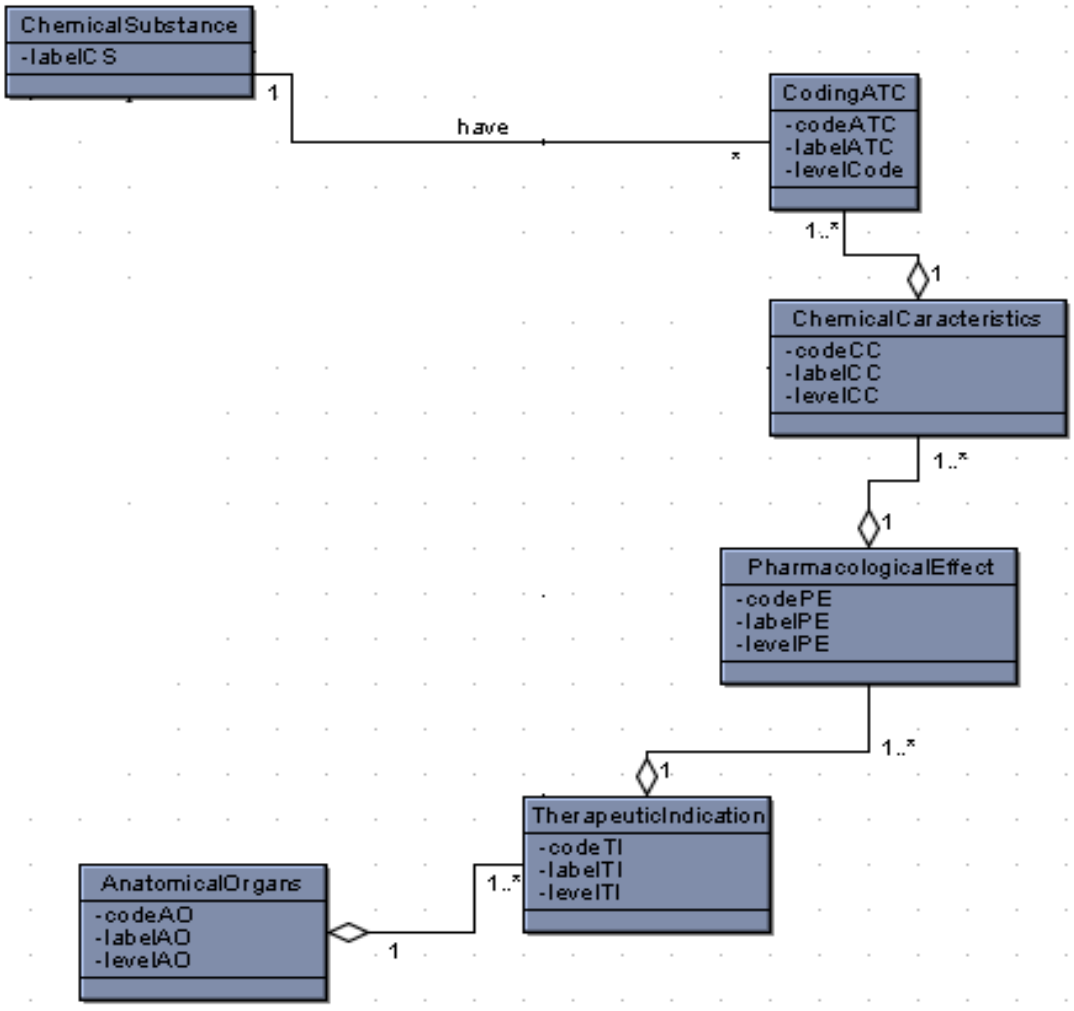


Figure A.1. Diagramme de classe de la classification ATC

❖ Modélisation UML de la CIM-10

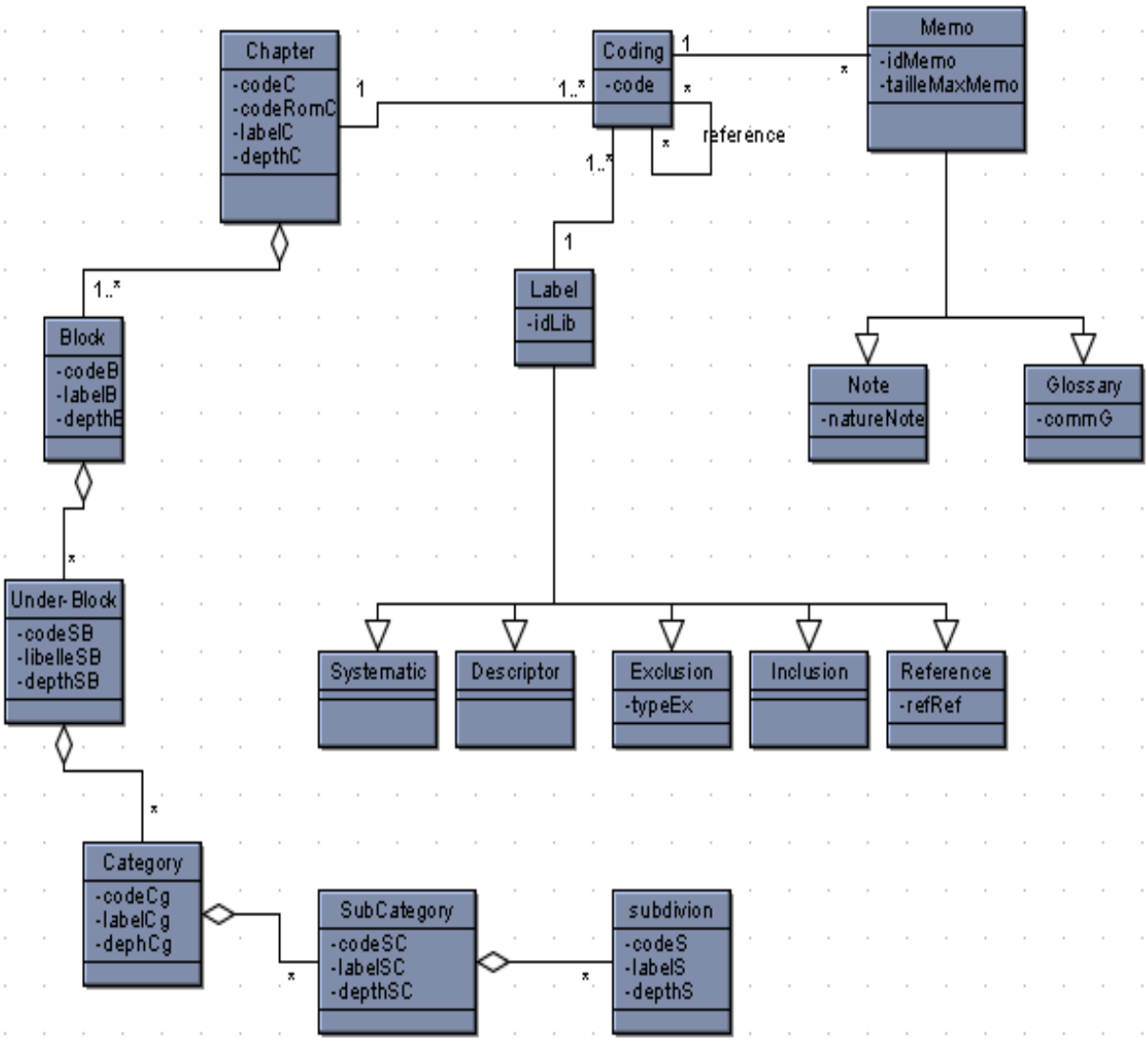


Figure A.2. Diagramme de classe de la CIM-10

ANNEXE B

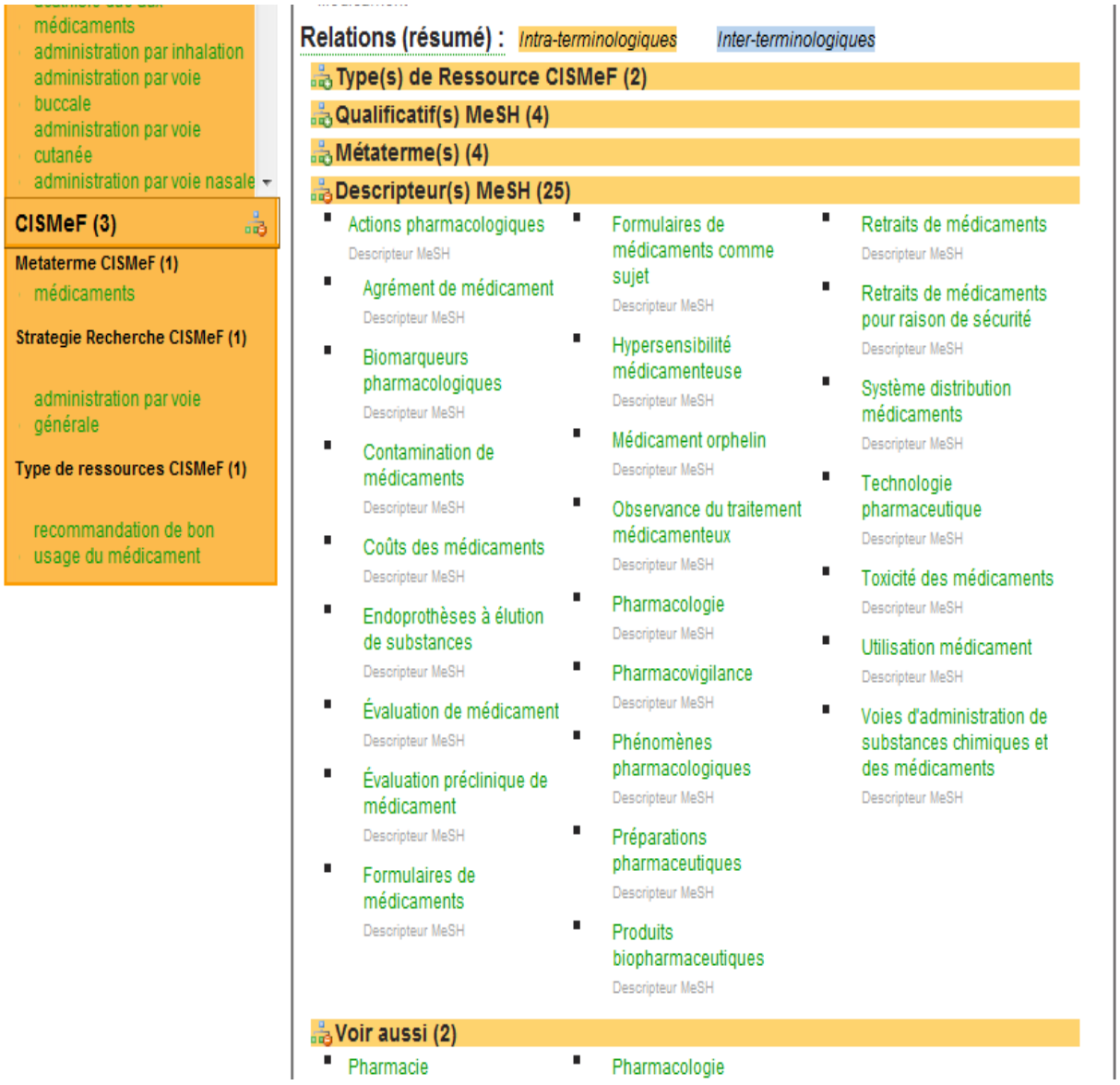


Figure B.1. La liste des descripteurs MeSH en relation avec le métaterme « *médicaments* »

Vu la structure hiérarchique du thésaurus MeSH, s'ajoutent à cette liste tous les descripteurs qui sont hiérarchiquement inférieurs à ces derniers :

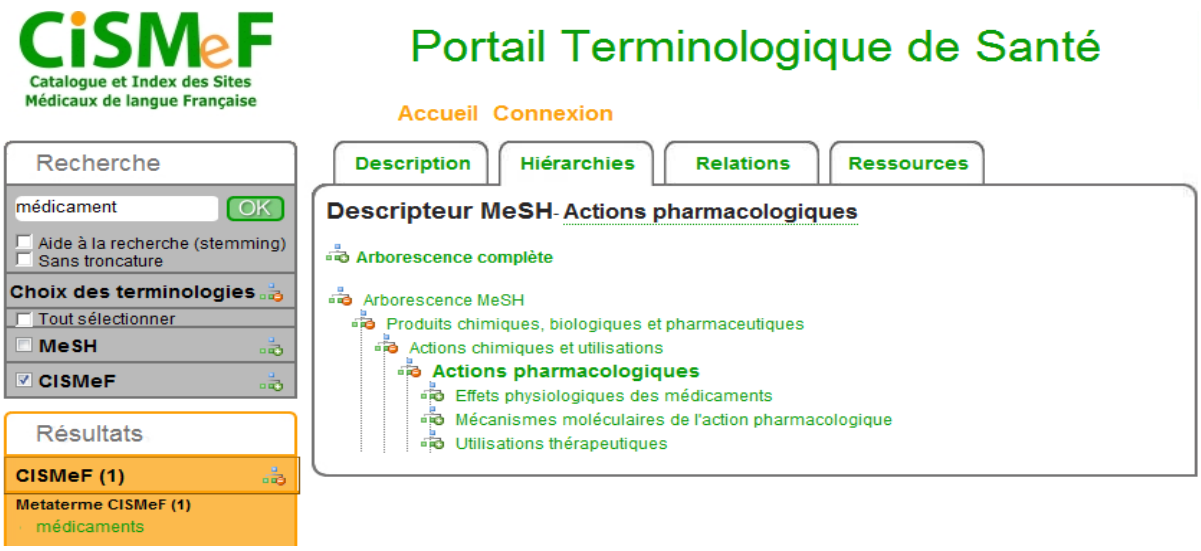


Figure B.2. La hiérarchie du descripteur « actions pharmacologiques »

ANNEXE C

Letters in French include the following accented forms,

â à ç ë é ê è ï î ô û ù

The following letters are vowels:

a e i o u y â à ë é ê è ï î ô û ù

Assume the word is in lower case. Then put into upper case ***u*** or ***i*** preceded and followed by a vowel, and ***y*** preceded or followed by a vowel. ***u*** after ***q*** is also put into upper case. For example,

jouer -> joUer

ennuie -> ennuIe

yeux -> Yeux

quand -> qUand

(The upper case forms are not then classed as vowels)

If the word begins with two vowels, *RV* is the region after the third letter, otherwise the region after the first vowel not at the beginning of the word, or the end of the word if these positions cannot be found. (Exceptionally, ***par***, ***col*** or ***tap***, at the beginning of a word is also taken to define *RV* as the region to their right.)

For example,

a i m e r a d o r e r v o l e r t a p i s

|...| |.....| |.....| |...|

*R*1 is the region after the first non-vowel following a vowel, or the end of the word if there is no such non-vowel. *R*2 is the region after the first non-vowel following a vowel in *R*1, or the end of the word if there is no such non-vowel.

For example:

f a m e u s e m e n t

|.....R1.....|

|...R2....|

Note that *R*1 can contain *RV* (*adorer*), and *RV* can contain *R*1 (*voler*).

Below, 'delete if in *R*2' means that a found suffix should be removed if it lies entirely in

*R*2, but not if it overlaps *R*2 and the rest of the word. 'delete if in *R*1 and preceded by *X*' means that *X* itself does not have to come in *R*1, while 'delete if preceded by *X* in *R*1' means that *X*, like the suffix, must be entirely in *R*1.

Start with step 1

Step 1: Standard suffix removal

Search for the longest among the following suffixes, and perform the action indicated.

ance iqUe isme able iste eux ances iqUes ismes ables istes

delete if in *R*2

atrice ateur ation atrices ateurs ations

delete if in *R*2

if preceded by ***ic***, delete if in *R*2, else replace by ***iqU***

logie logies

replace with ***log*** if in *R*2

usion ution usions utions

replace with ***u*** if in *R*2

ence ences

replace with ***ent*** if in *R*2

ement ements

delete if in *RV*

if preceded by ***iv***, delete if in *R*2 (and if further preceded by ***at***, delete if in *R*2), otherwise,

if preceded by ***eus***, delete if in *R*2, else replace by ***eux*** if in *R*1, otherwise,

if preceded by ***abl*** or ***iqU***, delete if in *R*2, otherwise,

if preceded by ***ièr*** or ***Ièr***, replace by ***i*** if in *RV*

ité ités

delete if in *R*2

if preceded by ***abil***, delete if in *R*2, else replace by ***abl***, otherwise,

if preceded by ***ic***, delete if in *R*2, else replace by ***iqU***, otherwise,

if preceded by ***iv***, delete if in *R*2

if ive ifs ives

delete if in *R*2

if preceded by ***at***, delete if in *R*2 (and if further preceded by ***ic***, delete if in *R*2, else replace by ***iqU***)

eaux

replace with ***eau***

aux

replace with ***al*** if in *R*1

euse euses

delete if in *R*2, else replace by ***eux*** if in *R*1

issement issements

delete if in *R*1 and preceded by a non-vowel

amment

replace with ***ant*** if in *RV*

emment

replace with ***ent*** if in *RV*

ment ments

delete if preceded by a vowel in *RV*

In steps 2*a* and 2*b* all tests are confined to the *RV* region.

Do step 2*a* if either no ending was removed by step 1, or if one of endings ***amment***, ***emment***, ***ment***, ***ments*** was found.

Step 2*a*: Verb suffixes beginning ***i***

Search for the longest among the following suffixes and if found, delete if preceded by a non-vowel.

îmes ît îtes i ie ies ir ira irai iraIent irais irait iras irent irez iriez irions irons iront is issaIent issais issait issant issante issantes issants isse issent isses issez issiez issions issons it

(Note that the non-vowel itself must also be in *RV*.)

Do step 2*b* if step 2*a* was done, but failed to remove a suffix.

Step 2*b*: Other verb suffixes

Search for the longest among the following suffixes, and perform the action indicated.

ions

delete if in *R*2

é ée ées és èrent er era erai eraIent erais erait eras erez eriez erions erons eront ez iez

delete

âmes ât âtes a ai aIent ais ait ant ante antes ants as asse assent asses assiez assions

delete

if preceded by ***e***, delete

(Note that the ***e*** that may be deleted in this last step must also be in *RV*.)

If the last step to be obeyed — either step 1, 2*a* or 2*b* — altered the word, do step 3

Step 3

Replace final ***Y*** with ***i*** or final ***ç*** with ***c***

Alternatively, if the last step to be obeyed did not alter the word, do step 4

Step 4: Residual suffix

If the word ends ***s***, not proceeded by ***a***, ***i***, ***o***, ***u***, ***è*** or ***s***, delete it.

In the rest of step 4, all tests are confined to the *RV* region.

Search for the longest among the following suffixes, and perform the action indicated.

ion

delete if in *R*2 and preceded by ***s*** or ***t***

ier ière Ier Ière

replace with ***i***

e

delete

ë

if preceded by ***gu***, delete

(***ion*** is removed only when it is in *R*2 — as well as being in *RV* — and proceeded by ***s*** or ***t*** which must be in *RV*.)

Always do steps 5 and 6.

Step 5: Undouble

If the word ends ***enn***, ***onn***, ***ett***, ***ell*** or ***eill***, delete the last letter

Step 6: Un-accent

If the word ends ***é*** or ***è*** followed by at least one non-vowel, remove the accent from the ***e***.

And finally:

Turn any remaining ***I***, ***U*** and ***Y*** letters in the word back into lower case.

ANNEXE D

```
<rdf:RDF xml:base="file:/C:/Users/sakji.saoussen/Documents/saoussen.owl">
  <!-- Ontology Information -->
  <owl:Ontology rdf:about=""/>
  <!-- Classes -->
  <owl:Class rdf:about="http://www.chu-rouen.fr/cismef#Descripteur"/>
  <owl:Class rdf:about="http://www.chu-rouen.fr/cismef#Editeur"/>
  <owl:Class rdf:about="http://www.chu-rouen.fr/cismef#Fiche"/>
  <owl:Class rdf:about="http://www.chu-rouen.fr/cismef#SynonymeDescripteur"/>
  <owl:Class rdf:about="http://www.chu-rouen.fr/cismef#Terminologie"/>
  <!-- Annotation Properties -->
  <owl:AnnotationProperty rdf:about="http://www.w3.org/2000/01/rdf-schema#label"/>
  <!-- Datatype Properties -->
  <owl:DatatypeProperty rdf:about="http://www.chu-rouen.fr/cismef#URLFiche"/>
  <owl:DatatypeProperty rdf:about="http://www.chu-rouen.fr/cismef#idDescripteur"/>
  <owl:DatatypeProperty rdf:about="http://www.chu-rouen.fr/cismef#idEditeur"/>
  <owl:DatatypeProperty rdf:about="http://www.chu-rouen.fr/cismef#idSynonyme"/>
  <owl:DatatypeProperty rdf:about="http://www.chu-rouen.fr/cismef#idTermino"/>
  <owl:DatatypeProperty rdf:about="http://www.chu-rouen.fr/cismef#numFiche"/>
  <owl:DatatypeProperty rdf:about="http://www.chu-rouen.fr/cismef#remarque"/>
  <owl:DatatypeProperty rdf:about="http://www.chu-rouen.fr/cismef#sousTitreFiche"/>
  <owl:DatatypeProperty rdf:about="http://www.chu-rouen.fr/cismef#titreFiche"/>
  <!-- Object Properties -->
  <owl:ObjectProperty rdf:about="http://www.chu-rouen.fr/cismef#appartientA"/>
  <owl:ObjectProperty rdf:about="http://www.chu-rouen.fr/cismef#avoirEditeur"/>
  <owl:ObjectProperty rdf:about="http://www.chu-rouen.fr/cismef#decritPar"/>
  <!-- Instances -->
  <cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_CIM_LIB_11252"/>
  <cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_CIM_LIB_9672"/>
  <cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_MSH_D_000925"/>
  <cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_MSH_D_001794"/>
  <cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_MSH_D_004211"/>
<cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_SNO_F_0" cismef:idDescripteur="SNO_F_0">
  <cismef:appartientA rdf:resource="http://www.chu-rouen.fr/cismef#Terminologie_SNOMED"/>
- <rdfs:label xml:lang="fr">
    Section 0 Fonctions biologiques en général, états du malade et diagnostics infirmiers
  </rdfs:label>
</cismef:Descripteur>
<cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_SNO_F_00" cismef:idDescripteur="SNO_F_00">
  <cismef:appartientA rdf:resource="http://www.chu-rouen.fr/cismef#Terminologie_SNOMED"/>
  <rdfs:label xml:lang="fr">00 Fonctions biologiques: termes généraux</rdfs:label>
</cismef:Descripteur>
<cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_SNO_F_00000" cismef:idDescripteur="SNO_F_00000">
  <cismef:appartientA rdf:resource="http://www.chu-rouen.fr/cismef#Terminologie_SNOMED"/>
  <rdfs:label xml:lang="fr">fonction biologique</rdfs:label>
</cismef:Descripteur>
<cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_SNO_F_00001" cismef:idDescripteur="SNO_F_00001">
  <cismef:appartientA rdf:resource="http://www.chu-rouen.fr/cismef#Terminologie_SNOMED"/>
  <rdfs:label xml:lang="fr">fonction corporelle générale normale</rdfs:label>
</cismef:Descripteur>
<cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_SNO_F_00002" cismef:idDescripteur="SNO_F_00002">
  <cismef:appartientA rdf:resource="http://www.chu-rouen.fr/cismef#Terminologie_SNOMED"/>
  <rdfs:label xml:lang="fr">fonction corporelle générale anormale</rdfs:label>
</cismef:Descripteur>
<cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_SNO_F_00003" cismef:idDescripteur="SNO_F_00003">
  <cismef:appartientA rdf:resource="http://www.chu-rouen.fr/cismef#Terminologie_SNOMED"/>
  <rdfs:label xml:lang="fr">fonction corporelle générale augmentée</rdfs:label>
</cismef:Descripteur>
<cismef:Descripteur rdf:about="http://www.chu-rouen.fr/cismef#Descripteur_SNO_F_00004" cismef:idDescripteur="SNO_F_00004">
  <cismef:appartientA rdf:resource="http://www.chu-rouen.fr/cismef#Terminologie_SNOMED"/>
  <rdfs:label xml:lang="fr">fonction corporelle générale diminuée</rdfs:label>
</cismef:Descripteur>
```

Figure D.1. Les ressources de la base de données CISMeF en format RDF

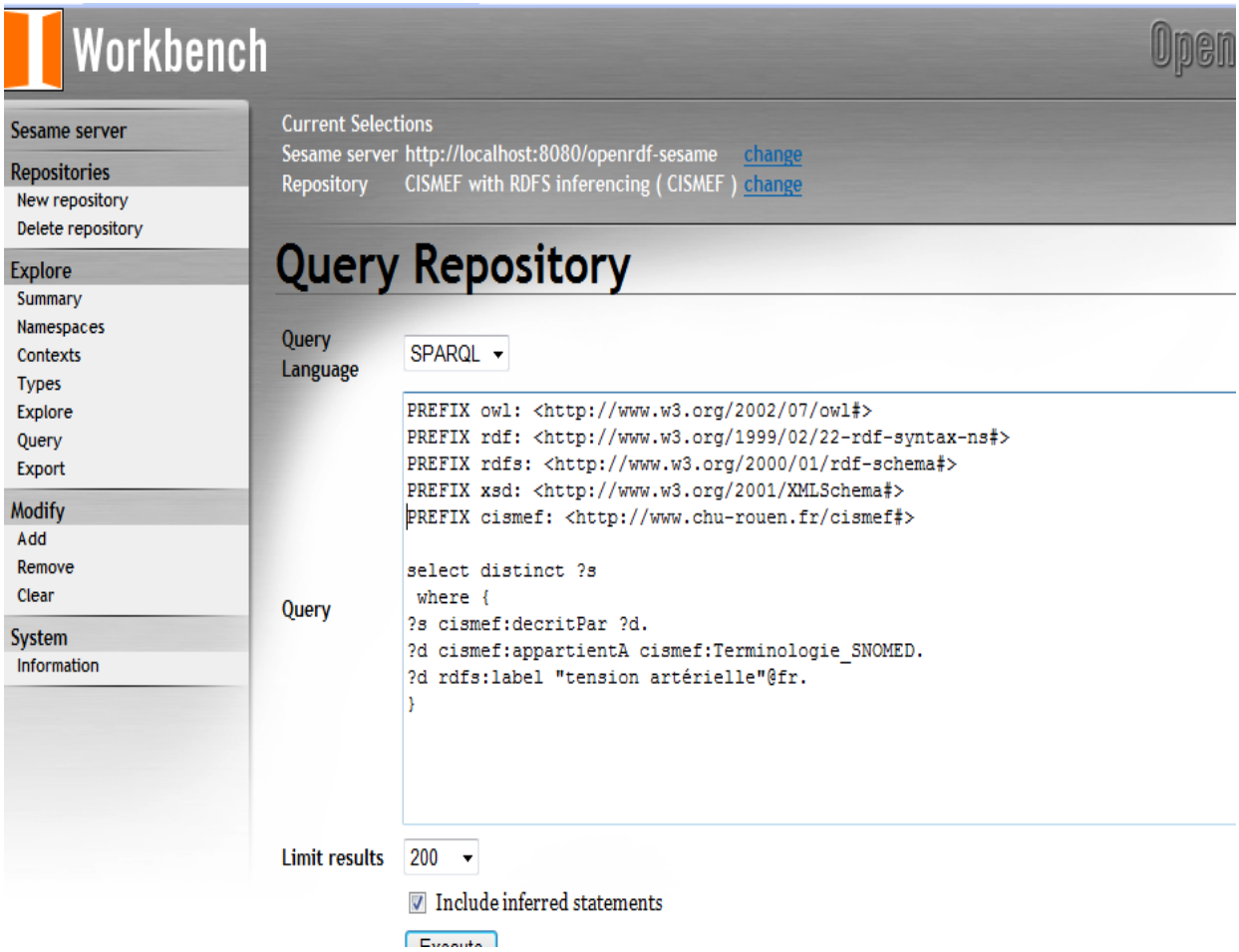


Figure D.2. Exemple de requête SPARQL en utilisant l'interface de Sésame : Trouver les ressources indexées par le descripteur SNOMED « tension artérielle »

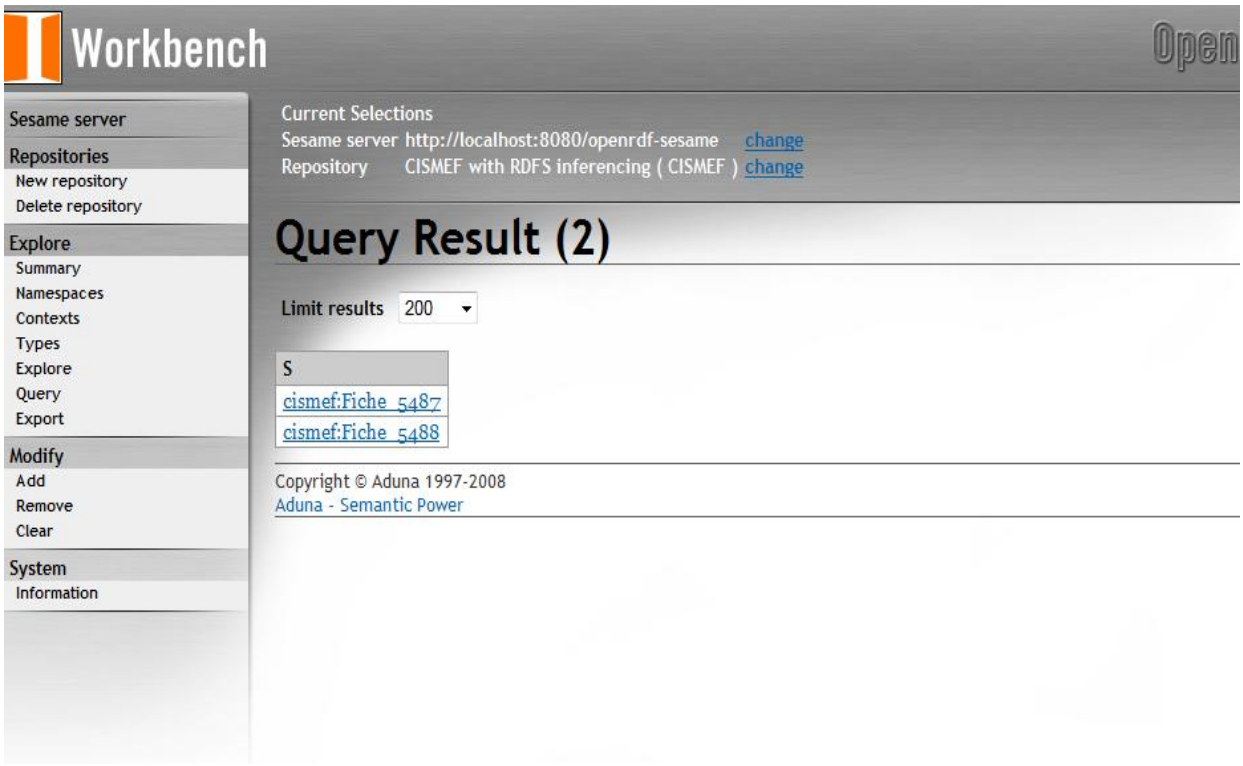
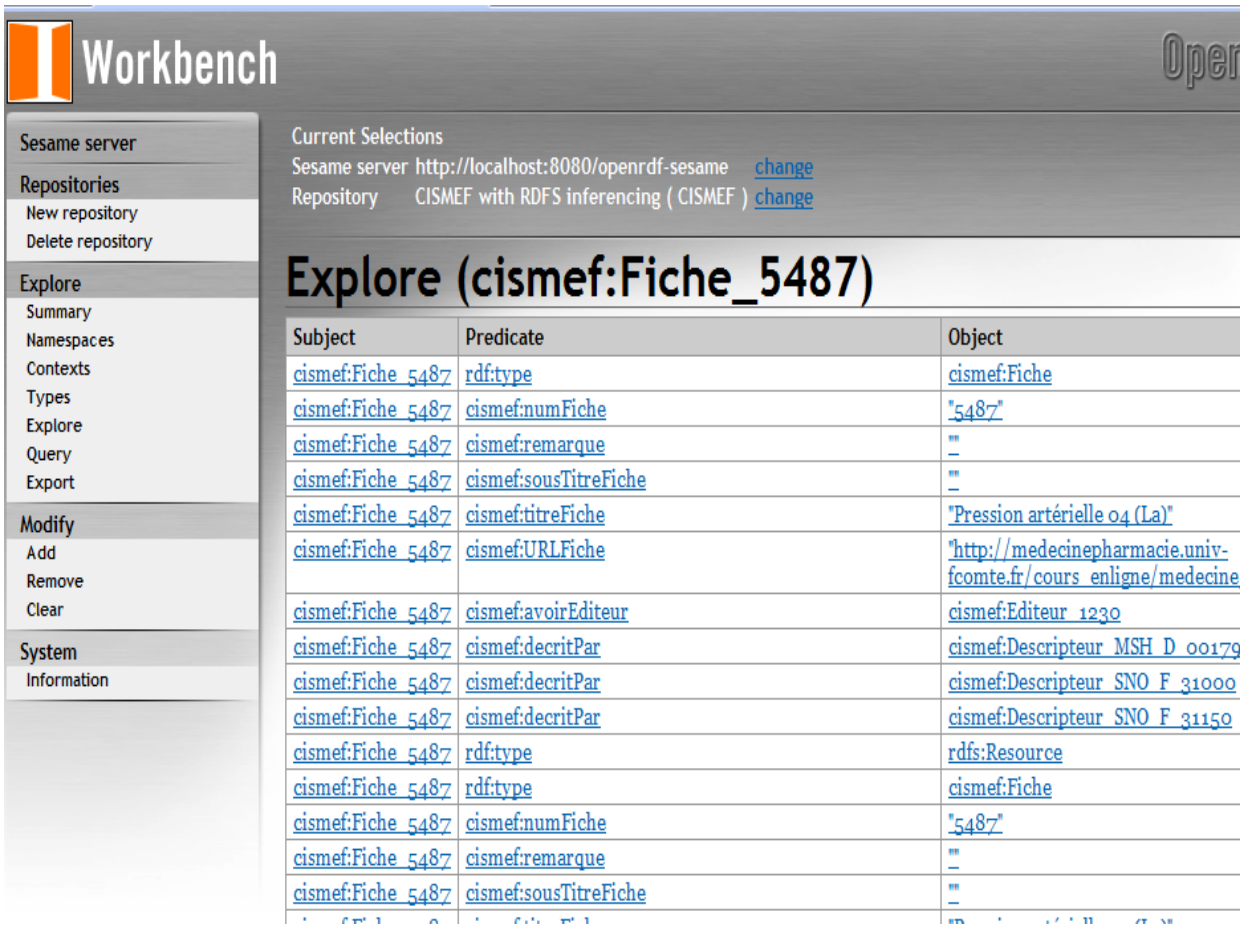


Figure D.3. Résultat de la requête :
Les ressources du catalogue CISMeF indexées par le descripteur SNOMED « tension artérielle »



Subject	Predicate	Object
cismef:Fiche_5487	rdf:type	cismef:Fiche
cismef:Fiche_5487	cismef:numFiche	"5487"
cismef:Fiche_5487	cismef:remarque	""
cismef:Fiche_5487	cismef:sousTitreFiche	""
cismef:Fiche_5487	cismef:titreFiche	"Pression artérielle 04 (La)"
cismef:Fiche_5487	cismef:URLFiche	"http://medecinepharmacie.univ-fcomte.fr/cours_enligne/medecine
cismef:Fiche_5487	cismef:avoirEditeur	cismef:Editeur_1230
cismef:Fiche_5487	cismef:decritPar	cismef:Descripteur_MSH_D_00179
cismef:Fiche_5487	cismef:decritPar	cismef:Descripteur_SNO_F_31000
cismef:Fiche_5487	cismef:decritPar	cismef:Descripteur_SNO_F_31150
cismef:Fiche_5487	rdf:type	rdfs:Resource
cismef:Fiche_5487	rdf:type	cismef:Fiche
cismef:Fiche_5487	cismef:numFiche	"5487"
cismef:Fiche_5487	cismef:remarque	""
cismef:Fiche_5487	cismef:sousTitreFiche	""

Figure D.4. Les informations en RDF de la première ressource du résultat : la ressource n°5487