Université de Rouen **U.F.R. des sciences et techniques**
École doctorale sciences physiques, mathématiques et de l'information pour l'ingénieur

Méthodes pour la mise en relations des terminologies médicales : contribution à l'interopérabilité sémantique Inter et Intra terminologique

THÈSE

présentée et soutenue publiquement le 24 Juin 2010

pour l'obtention du

Doctorat de l'université de Rouen

(spécialité informatique)

par

Tayeb Merabti

Composition du jury

Directeur de thèse :	Stefan Darmoni
Co-encadrants :	Thierry Lecroq Michel Joubert
Rapporteurs :	Pierre Zweigenbaum Jean-Marie Rodrigues

Laboratoire d'informatique, de traitement de l'information et des systèmes

Résumé

Depuis une vingtaine d'années, l'accès et l'utilisation des données médicales sont devenus des enjeux majeurs pour les professionnels de santé comme pour le grand public. Dans ce contexte, plusieurs terminologies médicales spécialisées ont été créées. Ces terminologies ont pour la plupart des formats de représentation et visées différentes : la nomenclature SNOMED 3.5 pour le codage d'informations cliniques, les classifications CIM10 et CCAM pour le codage épidémiologique puis médico-économique, le thésaurus MeSH pour la bibliographie... Devant ce constat et la nécessité grandissante de permettre la coopération de différents acteurs de la santé et des systèmes d'information associés, il apparait nécessaire de rendre les terminologies « interopérables ». Notre travail qui s'inscrit dans le cadre du projet ANR InterSTIS (Interopérabilité Sémantique des Terminologies dans les Systèmes d'informations de Santé Français), vise à mettre en œuvre des méthodes permettant de contribuer à l'interopérabilité entre les différentes terminologies francophones qui seront intégrées dans un même serveur Multi-Terminologique. De plus, nous utilisons nos différents algorithmes conjointement avec le métathésaurus UMLS afin d'apporter une plus grande couverture au niveau des relations entre les terminologies. Nous bénéficions, notamment, dans le cadre de cette thèse d'une expérience riche dans le domaine du Traitement Automatique de la Langue (TAL) issue des précédents travaux de recherche dans les équipes CISMeF et LERTIM.

Abstract

Since twenty years ago, access and use of medical data become major issues for health professional and lay people. In this context, multiple health terminologies were be developped. These terminologies have mostly different format and purpose : SNOMED International for clinical coding, CCAM and ICD10 used for epidemiological and medico-economic purposes, MeSH thesaurus for bibliographic databases. According to this and the growing need to allow cooperation between differnt health actors and related information systems. It is necessary to allow interoprablity between health terminologies. This work take place in a more global InterSTIS project (french acronym of Semantic Interoperability of terminologies in French Health Information Systems) funded by the French National Research Agency. The goal of InterSTIS is to make interoprable the main French medical terminologies within a "Health Multi-Terminology Server" (HMTS). We use also UMLS to provide a large coverage of relations between terminologies. We enjoy in the case of this PhD of an extensive experience in the Natural Language Processing field from a multiple CISMeF and LERTIM research projects.

Avant-propos

Cette thèse est le résultat de trois années d'efforts, des dizaines de nuits blanches, de plusieurs milliers de lignes de codes, de quelques billions de cycles CPU et de milliers de cafés. Il est aussi le fruit de rencontre avec de nombreuse personnes qui m'ont appris et surtout donnée beaucoup. Je tiens à exprimer tout d'abord mes remerciements aux membres du jury :

- Monsieur Stefan Darmoni, mon directeur de thèse, pour m'avoir accueilli dans sa formidable équipe CISMeF depuis mon stage de Master. Je lui adresse un grand merci pour tout le temps qu'il a investi pour que ce projet de recherche soit de qualité et pour que je puisse mener mon travail dans les meilleurs des conditions. J'espère avoir toujours autant de volonté et d'enthousiasme que lui pour mener mes recherches futures.
- Monsieur Michel Joubert, d'avoir co-encadré ce travail de thèse et grâce à qui j'ai beaucoup appris, autant sur le plan scientifique que personnel.
- Monsieur Thierry Lecroq pour son co-encadrement et son soutien scientifique, grâce à qui j'ai pu travailler sur de nouvelles perspectives de recherche et pour avoir toujours pris le temps de relire mes articles les bons comme les moins bons.
- Je tiens à remercier les Professeurs Jean-Marie Rodrigues et Pierre Zweigenbaum d'avoir accepté de servir de rapporteurs de cette thèse. Je suis flatté que ces distingués chercheurs aient bien voulu s'intéresser aux travaux que je présente dans cette thèse.

Je tiens à remercier l'ensemble de l'équipe CISMeF (Ahmed, Aurélie, Badisse, Benoît, Catherine, Elise, Gaëtan, Josette, Julien, Ivan, Lina, Romain, Saoussen, Suzanne et Zied).
Je remercie également les gens qui ont participé de près ou de loin à cette thèse : Ana Rath, Cedric Bousquet, Hocine Abdoune et Eric Sadou.
J'exprime ma sincère gratitude pour monsieur Djelloul Ziadi qui dès ma soutenance d'ingéniorat n'a pas cessé de m'encourager et de me pousser pour que je puisse terminer cette thèse.
Un grand merci à mes parents, pour leur présence et leur soutien. Mes deux frères Ah-

med et Hadj pour leurs encouragements et leur soutien aussi. Les mots me manquent pour exprimer toute ma reconnaissance pour eux.
Je remercie ma femme qui depuis notre union n'a pas cessé de me soutenir et de m'encourager. J'espère que je ferai autant pour elle afin qu'elle puisse terminer sa thèse.
Je remercie aussi les nouveaux membres de ma famille pour leurs encouragements : Abd el Halim, Fatima, Ahmed, Memen...
Je tiens à remercier aussi mes amis : Khaled, Mohamed M, faissal, Mohamed D, Senouci...
Enfin, mes ultimes remerciements vont à mon créateur, le tout puissant pour m'avoir donné la force et la volonté afin d'accomplir ce modeste travail.

Table des matières

Liste des tableaux

Table des figures

Chapitre 1

Introduction

1.1 Contexte général

Cette thèse s'inscrit dans le contexte général de l'informatique médicale. Notre champ de recherche s'intéresse plus particulièrement au traitement automatique des données médicales. Ces données peuvent être de nature variée : textes libres, bases de données médicales... À l'origine non structurées, elles sont pour la plupart stockées dans des bases de données sous forme exploitable pour permettre leur utilisation.
Depuis une vingtaine d'années l'accès et l'utilisation des données médicales est devenu un enjeu majeur pour les professionnels de santé comme pour le grand public. Dans ce contexte, plusieurs terminologies médicales spécialisées ont été créées. Ces terminologies ont pour la plupart des formats de représentations et visées différentes : la nomenclature SNOMED 3.5 pour le codage d'informations cliniques, les classifications CIM10 et CCAM pour le codage épidémiologique puis médico-économique, le thésaurus MeSH pour la bibliographie... Face à la multiplication de ces terminologies, les limites actuelles des outils ne proviennent pas de leurs performances à stocker et traiter rapidement de gros volumes de données, mais de leur incapacité à prendre en compte les divergences « syntaxiques » et « structurelles (sémantiques) » entre ces données.
Devant ce constat et la nécessité grandissante de permettre la coopération de différents acteurs de la santé et des systèmes d'information associés, il apparait nécessaire de rendre les terminologies « interopérables ».
Ainsi, il est indispensable de mettre en place un modèle commun de représentation des termes, quels que soient leurs terminologie ou référentiel d'origine, ainsi que les méthodes permettant de mettre en relation les termes d'une terminologie vers ses équivalents, directs ou indirects, dans d'autres terminologies.
Le projet InterSTIS (Interopérabilité Sémantique des Terminologies dans les Systèmes

d'informations de Santé Français)[1], a pour but de fédérer et de rendre interopérables les principales terminologies médicales au sein d'un « Serveur Multi-Terminologique de Santé » (SMTS).
Notre travail qui s'inscrit dans le cadre de ce projet, vise à mettre en œuvre des méthodes permettant de contribuer à l'interopérabilité entre les différentes terminologies francophones qui seront intégrées dans le SMTS.
Plusieurs travaux ont été menés par différentes équipes afin de mettre en place des plate-formes pour permettre l'interopérabilité entre terminologies. L'UMLS (Unified Medical Language System) développé par « US National Library of Medicine » depuis 1986, est le parfait exemple de ce type de plate-formes. Actuellement, il est considéré comme la plus large base de données terminologiques existante (section 3.2).
Toutefois, l'UMLS ne rend pas les terminologies intégrées interopérables au sens « sémantique ». Il intègre les différentes terminologies telles qu'elles se présentent sans établir de liens entre les termes de celles-ci autrement que par le rattachement de termes équivalents à un même identifiant ou par des relations explicites opérées manuellement Imel (2002). D'autres travaux ce sont intéressés à la problématique de mettre à disposition des serveurs de terminologies dans le domaine de santé Rector *et al.* (1997); Chute *et al.* (1999). De ces études, nous pouvons citer : le système GALEN GAL (2005) (General Architecture for Language and Nomenclatures), SYMBIOmatrics[2](SYnergies in Medical Informatics and Bioinformatics), le projet SemanticHEALTH[3].
Le SMTS est un serveur multi-termiologique développé par trois partenaires (MONDECA, CISMeF et LERTIM), et qui va permettre l'intégration et la gestion de toutes les terminologies médicales francophones disponibles, le SMTS sera décrit en détail dans la cadre de cette thèse (section 3.3). Cependant, dans le cadre du projet InterSTIS, 6 terminologies ont été incluses dans le SMTS : SNOMED International, CIM10, CCAM, MeSH, SNOMED, CISP2 et TUV. D'autres projets de recherches auxquels l'équipe CISMeF participe permettent d'intégrer d'autres terminologies au sein de ce serveur.
La mise en relation entre différentes terminologies est une tâche fastidieuse à réaliser. Et cela indépendamment du domaine de la recherche, que ce soit dans la science de l'information Zeng et Chan (2004); W3C (2004), les bases de données Doan *et al.* (2004) ou les ontologies Euzenat et Shvaiko (2007). En plus des hétérogénéités des terminologies, deux autres problèmes rendent l'interopérabilité entre les terminologies difficile : la première réside dans le traitement informel des relations dans les terminologies, ce qui conduit à des définitions ambiguës Sarker *et al.* (2003), malheureusement, ce problème demeure difficile à résoudre parce qu'il nécessite des modifications dans les logiques de construction de chaque terminologie : les relations hiérarchiques, les relations de

[1] ANR-07-TECSAN-010
[2] http://www.symbiomatrics.org
[3] http://www.semantichealth.org

synonymie... Le deuxième problème est l'automatisation des méthodes permettant de mettre en relation les termes de différentes terminologies. En effet, la plupart des alignements existant entre les terminologies sont établies manuellement. Dans le cadre de cette thèse, nous décrirons deux ensembles d'alignement manuels (ORPHANET vers CIM10 et ATC vers MeSH). Ces alignements sont très chronophages et nécessitent beaucoup de temps de travail, en plus, il sont très dépendants des terminologies alignées. L'exemple de la correspondance manuelle entre ATC et MeSH a nécessité plus de 6 hommes.mois. A l'évidence, il n'est pas possible à l'échelle d'une équipe comme CISMeF ou même à l'échelle d'un consortium comme InterSTIS d'éffectuer 190 alignements manuels entre 20 terminologies $\frac{N(N-1)}{2}$, en revanche, l'humain peut se focaliser sur ceux qu'il juge pertinent : SNOMED-CIM10 ATC-MeSH, par exemple.

1.1.1 Objectif

Dans ce travail, nous cherchons principalement à apporter une contribution à cette deuxième problématique liée à l'automatisation des méthodes d'alignements afin de mettre en relation les terminologies médicales francophones. Nous pensons que les outils de traitement automatique de la langue (TAL) peuvent être très utiles à ce niveau. Nous bénéficions, notamment, dans le cadre de cette thèse d'une expérience riche dans le domaine issue des précédents travaux de recherches dans les équipes CISMeF et LERTIM. De plus, nous utilisons nos différents algorithmes conjointement avec le métathésaurus UMLS afin d'apporter une plus grande couverture au niveau des relations entre les terminologies. Outre les méthodes d'alignements proposées, cette thèse va contribuer à poser les premiers jalons d'une possible approche permettant l'interopérabilité « sémantique » entre les terminologies médicales francophones, de plus, tous les alignements réalisés dans le cadre de cette thèse sont (seront) utilisés dans tous les travaux futurs qui nécessitent l'utilisation conjointe de plusieurs terminologies médicales : l'indexation multi-terminologique Pereira (2007) et la recherche d'information multi-terminologique Sakji (2008); Dirieh Dibad *et al.* (2009).

1.1.2 Organisation du mémoire

Dans ce mémoire, nous exposons en premier lieu le contexte des travaux effectués, en particulier les différents travaux de recherches entamés par les équipes CISMeF et LERTIM. Nous passerons en revue tous les travaux passés et futurs qui sont relatifs de près ou de loin aux besoins exprimés dans le cadre de ce travail. Nous présentons aussi, le projet InterSTIS qui finance ma thèse de recherche depuis 2007.

Le deuxième chapitre introduit toutes les terminologies francophones utilisées dans la plupart de nos travaux, il touche aussi à la problématique de l'intégration des terminologies au sein d'un même serveur multi-terminologique. Nous détaillerons dans cette partie principalement le serveur multi-terminologique de santé, le cœur des différents projets de recherches entamés il y a trois ans dans plusieurs laboratoires de recherches spécialisés dans le traitement de l'information médicale.
Le troisième chapitre aborde l'analyse de l'état de l'art relatif à nos travaux de recherches. Nous proposons une classification des différentes méthodes d'alignements inspirée de Euzenat et Shvaiko (2007) et leurs travaux sur les alignements entre les ontologies.
La suite de la thèse est consacrée aux différentes méthodes utilisées et implémentées dans le cadre de ce travail, nous détaillerons notre algorithme d'alignement lexical lorsque nous entamerons la partie de notre thèse consacrée à la projection du thésaurus ORPHANET vers F_UMLS (les terminologies francophones de l'UMLS). Nous introduisons aussi dans cette partie une approche mixte fondée sur les outils TAL et les relations hiérarchiques pour aligner les termes ORPHANET vers F_UMLS. La deuxième partie de ce chapitre est consacrée à la projection de la classification ATC vers UMLS. Dans cette partie, en plus de nos méthodes et outils, nous utilisons l'outil MetaMap pour aligner les termes en anglais de l'ATC vers UMLS puis comparer les résultats des deux méthodes. Nous terminerons ce chapitre en proposant une méthodologie permettant de aligner la classification CCAM vers les termes de l'UMLS. La méthode proposée dans cette partie est assez différente des autres méthodes car nous nous basons sur la structure des codes de la CCAM pour appliquer notre méthode.
Le chapitre suivant est consacré à la présentation des résultats des différentes méthodes utilisées pour mettre en relation des terminologies francophones vers F_UMLS.

Dans le chapitre 6, nous proposons une méthode d'interopérabilité entre terminologies fondée sur UMLS afin de projeter les relations de la terminologie SNOMED CT entre trois terminologies francophones. Nous verrons que cette méthode va permettre de lier différentes terminologies (CIM10, SNMI et MeSH) avec des relations issues d'une autre terminologie (SNOMED CT). Le chapitre suivant dresse les différents résultats obtenus par la projection des relations SNOMED CT.
Le chapitre 8 résume et permet de discuter les principaux résultats et d'évoquer les différentes problématiques ainsi que les différentes perspectives de cette thèse. Nous terminons avec deux derniers chapitres consacrés aux perspectives et à la conclusion. Des annexes sont aussi fournies où nous présentons deux études de cas, une sur le Serveur Multi-Terminologique de Santé et l'autre sur le Portail Terminologique de Santé développé par CISMeF.

Chapitre 2

Contexte de travail et projet de recherche

Dans ce chapitre, nous présentons le contexte des travaux effectués, en particulier les différents travaux de recherches entamés par les équipes CISMeF et LERTIM. Nous passerons en revue tous les travaux passés et futurs qui sont relatifs de près ou de loin aux besoins exprimés dans le cadre de ce travail. Nous présentons aussi, le projet InterSTIS qui finance cette thèse de recherche depuis 2007.

2.1 L'équipe CISMeF

2.1.1 Travaux de l'équipe CISMeF

L'équipe CISMeF est dirigée par le professeur Stéfan Darmoni et Benoît Thirion le conservateur de la bibliothèque médicale du CHU de Rouen. L'équipe est composée actuellement de quatre documentalistes experts dans la description et l'indexation dans le domaine de la santé, trois ingénieurs de recherche, et trois doctorants. La figure 2.1 illustre les différents rôles de chacun d'eux dans les projets de l'équipe. De nombreux travaux ont été entrepris par l'équipe CISMeF dans le domaine de la recherche d'information en santé et dans l'indexation.

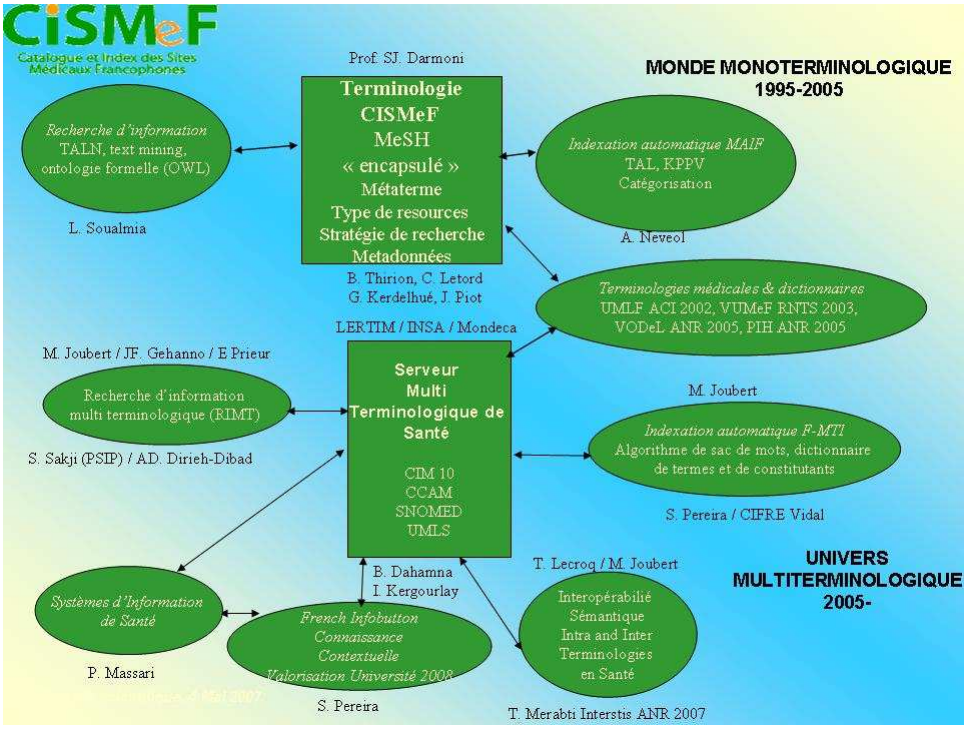

FIG. 2.1 – Organisation des projets de l'équipe CISMeF

2.1.2 Présentation du projet CISMeF

CISMeF (http://www.chu-rouen.fr/cismef ou http://www.cismef.org) est l'acronyme de Catalogue et Index des Sites Médicaux Francophones sur l'Internet. Il s'agit d'un portail de santé qui a été conçu pour cataloguer et indexer les sources d'information institutionnelles de santé françaises les plus importantes (N= 60 000) et ce afin de permettre une recherche plus pertinente pour les professionnels de santé, les étudiants mais aussi les patients, leurs familles, et d'une façon encore plus large le cyber-citoyen. CISMeF adhère aux principes de qualité de l'information de santé sur l'Internet définis par la Fondation Health on the Net (HON) depuis plus de 10 ans maintenant Darmoni *et al.* (1999).
Le site CISMeF (voir figure 2.2) est un site populaire avec un nombre d'utilisateurs (pendant 10 ans 1995-2005) se connectant à CISMeF d'environ 20 000 par jour ouvré. CISMeF utilise deux outils standards pour organiser l'information : le thésaurus MeSH (qui va être décrit en détail dans les sections suivantes) pour indexer les ressources, ainsi qu'un ensemble de méta-données extraites du noyau de Dublin Core Dekkers et Weibel (2003). Les métadonnées se réfèrent aux informations décrivant des ressources Web et dont les plus importantes sont le titre, l'identifiant, la date, le contenu, les mots clefs et le type de ressources. Pour décrire les ressources pédagogiques, l'équipe a ajouté huit méta-données spécifiques à CISMeF telles que : pays, institution...

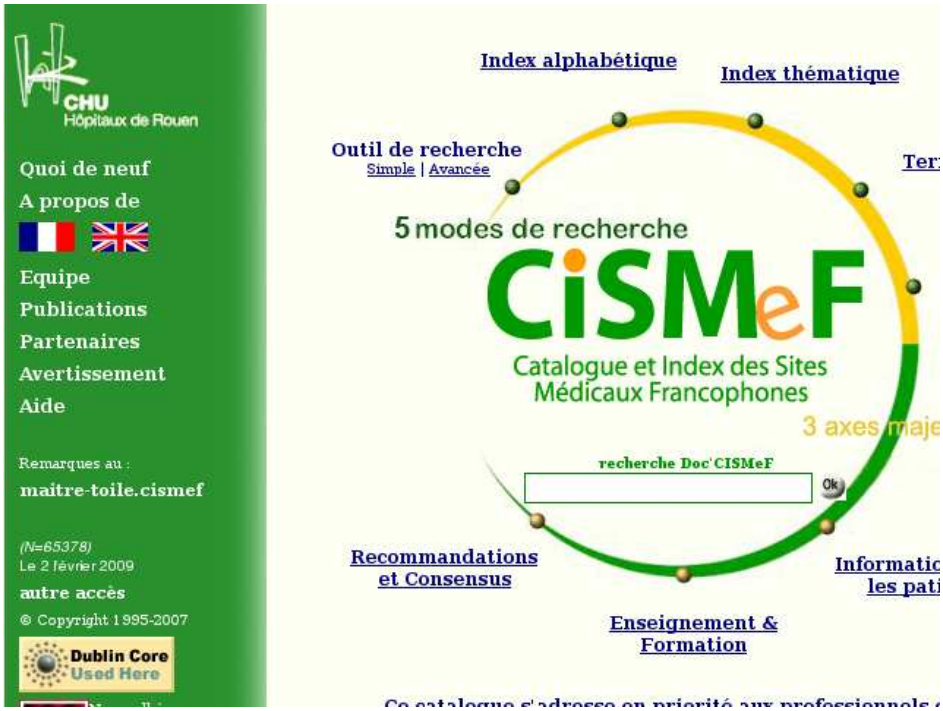

FIG. 2.2 – Le portail CISMeF

2.1.3 Les différents travaux de l'équipe CISMeF

Au centre des activités de l'équipe CISMeF se trouve la terminologie CISMeF. C'est en effet, sur cette terminologie que reposent les principaux travaux de recherche d'information dans le moteur et le catalogue de CISMeF. La terminologie CISMeF est utilisée principalement pour :

- la description des ressources : indexation des ressources avec les termes appartenant à la terminologie ;
- l'interprétation des requêtes des utilisateurs : traduction à l'aide des termes appartenant à la terminologie.

L'essentiel du travail de l'équipe consiste en la maintenance, la mise à jour du catalogue ainsi que son amélioration et son évolution, tant en termes de technologies utilisées que de recensement de nouvelles ressources et de facilité d'utilisation pour l'utilisateur.
L'ajout de nouvelles ressources (un exemple d'une ressource CISMeF est donné dans la figure 2.3) au catalogue s'effectue en quatre étapes :

1. recensement des ressources à l'aide d'une veille quotidienne ;
2. sélection des ressources selon des critères de qualité fondés sur le NetScoring (critères de qualité de l'information de santé sur Internet) Darmoni *et al.* (1999) ;
3. la description de chaque ressource CISMeF à l'aide d'une notice pour faciliter

la recherche dans le moteur de recherche CISMeF. Un ensemble de métadonnées est associé à chaque ressource par les indexeurs Darmoni *et al.* (1999, 2001) (figure 2.4). Ces métadonnées proviennent de plusieurs référentiels dont 11 champs (parmi les 15) du Dublin Core Dekkers et Weibel (2003); Thirion *et al.* (2004) pour les champs *auteur, date, description, format, identification, langue, éditeur, type de ressource, droits, sujet* et *titre*. Pour décrire les ressources pédagogiques, onze éléments de la catégorie « Education » du IEEE 1484 LOM (Learning Object Metadata) Bourda et Hélier (1999) sont utilisés en plus des autres métadonnées. Par ailleurs, des métadonnées spécifiques à CISMeF, ont été ajoutées pour décrire la qualité ou la localisation de la ressource : *institution, ville, province, pays, type d'accès, partenariat, coût et public ciblé*. Deux champs supplémentaires ont été créés pour les ressources destinées aux professionnels de santé : indication du niveau de preuve et la méthode utilisée pour l'établir Darmoni *et al.* (2003a). Les métadonnées HIDDEL (High Information Description Disclosure Evaluation Language) ont été introduites dans CISMeF dans le cadre du projet européen MedCircle (mars 2002 - septembre 2003) Mayer *et al.* (2003), qui avait pour but d'évaluer la qualité de l'information de santé afin de guider les utilisateurs vers des sources fiables.

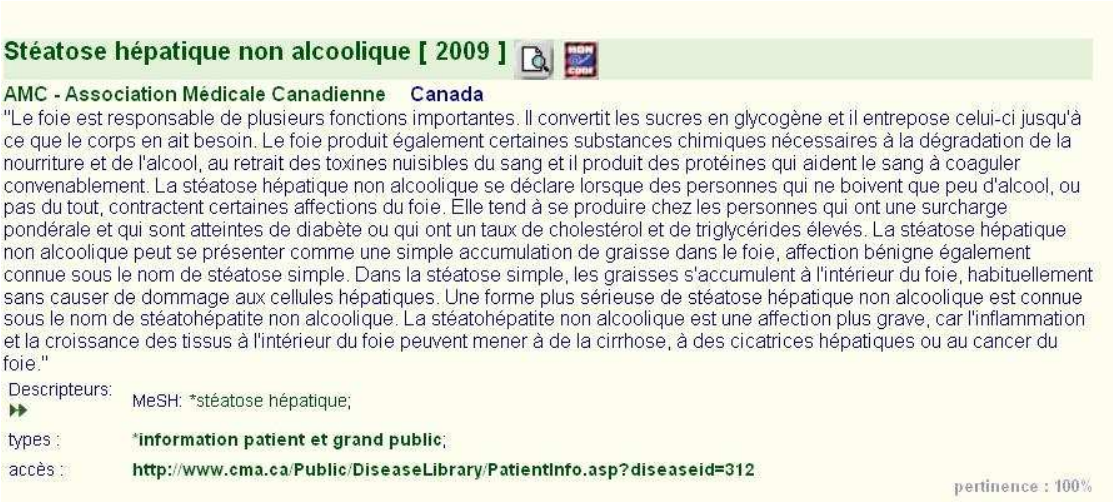

FIG. 2.3 – Exemple d'une ressource CISMeF

Cependant, il existe plusieurs niveaux d'indexation (assigner des mots clés à un document).

Niveau 1 : Une indexation purement manuelle pour les ressources de haute importance comme les recommandations par exemple. Un total de 36 439 ressources sont indexées manuellement par 12 992 mots clés MeSH différents dans CISMeF.

Niveau 2 : Une indexation supervisée qui consiste en une indexation automatique effectuée par un programme informatique sur le titre de la ressource. Les indexeurs sont ensuite chargés de valider et modifier à la main si nécessaire cette indexation. Elle est destinée aux ressources de qualité mais moins urgentes que celle du premier niveau. Un total de 8 878 ressources supervisées existe dans CISMeF, en utilisant 4 700 mots clés MeSH différents.

Niveau 3 : Une indexation purement automatique (sans validation humaine *a posteriori*) sur le titre pour les ressources de priorité faible dont la qualité ne nécessite pas une indexation précise. Un total de 25 583 ressources est indexées automatiquement dans CISMeF, en utilisant 7 939 mots clés MeSH différents.

4. La dernière étape consiste à mettre en ligne la ressource sur le catalogue.

FIG. 2.4 – Exemple d'une notice décrite par les différentes métadonnées

L'outil de recherche intégré au site CISMeF est Doc'CISMeF (voir figure 2.5). Cet outil donne un accès précis et rapide aux ressources : il permet de faciliter la saisie des requêtes par les utilisateurs afin d'obtenir une série de ressources susceptibles de contenir l'information recherchée. Ces ressources étaient affichées par ordre chronologique, mais depuis 2009, Doc'CISMeF permet un affichage combiné par ordre chronologique

et par pertinence. Cette dernière est calculée suivant le nombre de mots de la requête se trouvant dans les mots clés d'indexation et dans le titre. Ainsi, les ressources récentes avec une valeur maximale de pertinence sont affichées en premier. D'autre part, différents modes de recherche d'information sont possibles :

Une recherche simple : elle permet une saisie de requête sous forme d'expressions libres en français ou en anglais.

Une recherche avancée : elle permet des recherches poussées facilitées par l'utilisation d'un formulaire contenant des listes déroulantes permettant de combiner plusieurs champs (mots clés, type de ressources, ...) avec des opérateurs booléens (ET, OU, SAUF).

Une recherche *via* le serveur de terminologie : elle permet une recherche d'information à partir d'un mot clé sélectionné dans le serveur terminologique. Cette recherche peut être affinée (grâce à l'association de qualificatifs).

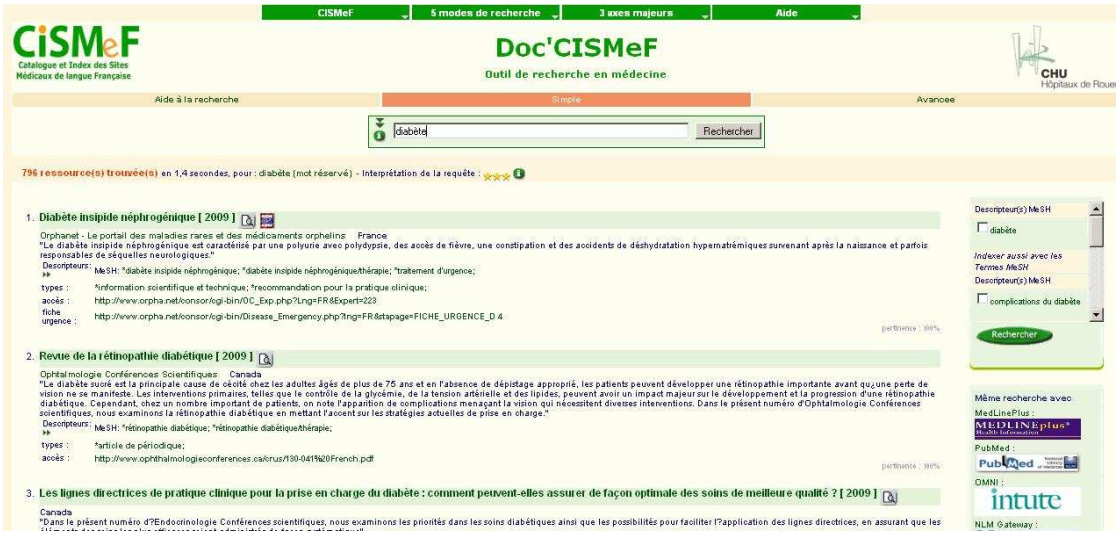

FIG. 2.5 – Exemple de recherche simple avec Doc'CISMeF

Par ailleurs, CISMeF permet aussi l'accès à d'autres sites spécialisés dans la recherche dans le domaine de la santé. L'accès à ces sites est réalisé de manière contextuelle dans CISMeF (l'onglet à droite de la figure 2.5). Cependant, plus récemment en 2009, CISMeF a développé « CISMeF InfoRoute » un outil en cours d'évaluation permettant un accès contextuel à plusieurs sites de santé regroupés par leur contexte d'utilisation. Par exemple, le contexte « Outils de recherche »(les sites : CISMeF, PubMed[1], Intute[2], ...), le contexte « Médicaments » (les sites : PIM (Portail d'Information sur le Médi-

[1] http://www.ncbi.nlm.nih.gov/pubmed/
[2] http://www.intute.ac.uk/

cament)[3], HAS (Haute Autorité de Santé)[4], AFSSAPS (Agence Française de Sécurité SAnitaire des Produits de Santé)[5], ...).
Il parait logique pour l'équipe CISMeF d'être impliquée dans des travaux touchant à la terminologie médicale notamment pour le MeSH mais aussi d'autres terminologies françaises telles que la CIM10 OMS (1993), la SNOMED Côté *et al.* (1993) et la CCAM Rodrigues *et al.* (2005a) ou toutes terminologies avec lesquelles des correspondances peuvent se faire. Les principaux travaux de CISMeF touchent à deux problématiques : a) l'amélioration de la terminologie CISMeF, b) faciliter la recherche d'information au sein du catalogue.

Amélioration de la terminologie CISMeF : l'équipe CISMeF fait évoluer jour après jour celle-ci Douyère *et al.* (2004). En effet, plusieurs collaborations avec d'autres équipes ont servi à l'enrichissement de la terminologie. Ainsi, l'équipe CISMeF a participé aux projets UMLF Zweigenbaum *et al.* (2003) (développement d'un lexique médical en français)[6] et VUMeF Darmoni *et al.* (2003b) de 2003 à 2007 en collaboration notamment avec l'équipe LERTIM et la société Vidal. Le but de ce projet était d'enrichir les terminologies médicales françaises dans l'UMLS (détaillé dans le chapitre 3). CISMeF a aussi collaboré avec la société Memodata (PME spécialiste des dictionnaires) dans le projet VODEL[7] en vue d'enrichir le catalogue de nombreuses définitions et traductions de plusieurs langues. D'autres travaux ont aussi été menés pour mieux comprendre le langage médical courant utilisé par les usagers non spécialistes du domaine dans l'élaboration de leurs requêtes Darmoni *et al.* (2002) notamment MEDLINEPlus.

L'indexation automatique au sein du catalogue : plusieurs travaux visant à améliorer la recherche des utilisateurs ont été effectués parmi lesquels ont peut citer, le projet Cogni'CISMeF pour un dialogue homme-machine et le système KnowQuE (Knowledge-based Query Expansion) Soualmia (2004); Soualmia *et al.* (2009) pour une recherche d'information implicite.
L'indexation manuelle des ressources constitue la base de la recherche d'information dans CISMeF, elle est très importante et malheureusement très coûteuse en temps (chronophage). En effet, l'indexation manuelle des ressources demande une analyse fine du document et de la terminologie ainsi que des bonnes connaissances métier. La forte expansion des ressources médicales de qualité sur Internet a poussé l'équipe CISMeF à chercher à augmenter sa productivité en disposant d'outils automatiques d'indexation. Les travaux d'Aurélie Névéol dans le cadre de sa thèse Névéol *et al.* (2005); Névéol (2005) ont mené à l'élaboration du système MAIF (MeSH Automatic Indexing in French) : un système d'indexation auto-

[3] http://doccismef.chu-rouen.fr/servlets/PIM
[4] www.has-sante.fr/portail/jcms/j_/accueil
[5] http://www.afssaps.fr
[6] http://www-test.biomath.jussieu.fr/umlf/
[7] http://www.rntl.org/projet/resume2005/vodel.htm

matique pour le MeSH, suivi par le développement pendant la thèse de Suzanne Pereira de F-MTI (French-Mutli Terminological Indexer) : un système d'indexation mutli-terminologique pour les terminologies en français.

2.1.4 CISMeF : d'un univers mono-terminologique vers un univers multi-terminologique

Dès 2005, lors du début de la thèse de S. Pereira, une décision stratégique de l'équipe CISMeF a permis le passage d'un monde mono-terminologique à un univers multi-terminologique (ECMT) (voir figure 2.1) Darmoni *et al.* (2009b) par la mise au point d'un extracteur de concept multi-terminologique et le développement d'un Serveur Multi-Terminologique de Santé (SMTS) qui rassemble plusieurs terminologies médicales francophones (voir section 3.3) et par une Recherche d'Information Multi-Terminologique (RIMT) (Thèse de S. Sakji).
La thèse de S. Pereira Pereira (2007) constitue le premier travail utilisant un environnement multi-terminologique. Cette thèse a eu pour objectif la réalisation et l'évaluation d'un outil d'indexation multi-terminologique « F-MTI » (French-Mutli Terminological Indexer) Pereira *et al.* (2009b,a). En plus du MeSH, l'outil F-MTI utilise déjà plusieurs terminologies médicales pour l'indexation des ressources médicales. F-MTI fonctionne en deux temps : une extraction des concepts des terminologies étudiées, puis une restriction vers les terminologies choisies Pereira *et al.* (2008). Cette restriction s'effectue via les relations entre terminologies.
Depuis 2007, plusieurs travaux orientés sur la problématique de la multi-terminologie ont été lancés. Le projet ANR InterSTIS « Interopérabilité Sémantique des Terminologies dans les systèmes d'Informations de Santé français » est l'un des projets lancés autour du SMTS. Ma thèse de recherche s'inscrit dans le cadre de ce projet que nous allons décrire dans la section 2.4.
Depuis 2009, l'équipe a aussi développé un outil de recherche d'information multi-terminologique Sakji *et al.* (2009a); Dirieh Dibad *et al.* (2009) dans un double contexte :

- **Documentation :** Commencée en 2007 avec la thèse de Saoussen Sakji, également encadrée par SJ. Darmoni et M. Joubert. Elle est la continuité de la thèse de Lina Saoualmia Soualmia (2004) sur la recherche d'information mono-terminologique. L'objectif du travail de Saoussen Sakji est la mise en œuvre d'un outil de recherche d'information multi-terminologique sur le catalogue CISMeF Sakji (2008). Ce travail a permis aussi de modifier le modèle terminologique de CISMeF pour la prise en charge de plusieurs terminologies médicales. L'outil développé est en cours d'évaluation dans le moteur de recherche Doc'CISMeF. De plus, dans un cadre pharmacologique une recherche bi-terminologique a été

élaborée Lethord *et al.* (2008) dans le cadre de cette thèse sur le PIM (Portail d'Information sur le Médicament) Sakji *et al.* (2009b); Lethord *et al.* (2008). Cette thèse s'inscrit dans le cadre du projet PSIP (Patient Safety through Intelligent Procedures in medication) Beuscart *et al.* (2009), un projet européen visant à une meilleure connaissance des effets indésirables liés aux médicaments, commencé en 2008 impliquant 13 partenaires.

– **Dossier Électronique du Patient :** Commencé en 2008 avec la thèse de Ahmed Diouf Dirieh-Dibad, encadré par Stéfan Darmoni, Philippe Massari et Elise Prieur. Le but de cette thèse étant aussi la recherche d'information mais dans un autre contexte: permettre une recherche d'informations multi-terminologique (RIMT) sur les dossiers électroniques Sakji *et al.* (2009a); Dirieh Dibad *et al.* (2009). Une modélisation formelle du Dossier Électronique du Patient (DEP) a été réalisée pour permettre la RIMT.
Pour ces deux thèses, une collaboration avec ORACLE a été entreprie pour utiliser les outils « sémantiques » notamment SPARQL Prud'hommeaux et Seaborne (2008).

Cependant, pour faciliter l'accès à toutes ces terminologies, CISMeF a développé un Portail de Terminologies de Santé (PTS) qui représente une porte d'entrée à ces dernières (voir figure 2.6), sans se soucier ni de la gestion ni de la mise à jour (ce qui est le cas pour le SMTS). Ce portail permettra aussi d'intégrer les relations entre terminologies trouvées dans le cadre de cette thèse. Nous présentons dans l'annexe B une étude de cas montrant l'utilité des relations inter-terminologiques pour faciliter la navigation dans le PTS. Le dernier projet est le projet ALADIN-DTH (Assistant de Lutte Automatisée et de Détection des Infections Nosocomiales à partir de Documents Textuels Hospitaliers)[8] Metzger *et al.* (2009). ALADIN vise à développer un outil de détection automatique des infections nosocomiales à partir des documents médicaux du dossier patient rédigés en langage naturel Proux *et al.* (2010). Dans le cadre de ce projet, CISMeF a développé un outil permettant de retourner sous forme structurée tous les termes de toutes les terminologies (voir figure 2.7) à partir d'une requête en langage naturel (avec ou sans expansion).

Dans la partie perspective, nous détaillons le projet PlaIR (Plateforme d'Indexation Régionale) démarré en 2009. Ce projet a pour objectif de mutualiser l'ensemble des travaux des laboratoires LITIS (Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes) et LiDiFra (Linguistique, Didactique, Francophone) portant sur l'indexation et la recherche d'information, que ce soit dans un univers de documents électroniques avec des vocabulaires contrôlés liés à des domaines métiers (comme dans les sciences de la santé, le droit ou les sciences de l'ingénieur) ou dans un univers de

[8] www.aladin-project.eu

FIG. 2.6 – Exemple de recherche dans le PTS

```
<?xml version="1.0" encoding="ISO-8859-1" ?>
- <interpretation>
  - <descripteurs>
      <des ter="CIM" id="CIM_LIB_4747" code="J45">asthme</des>
      <des ter="DRC" id="DRC_D_717" code="D717">asthme</des>
      <des ter="DRC" id="DRC_RCE_27" code="RCE_27">asthme</des>
      <des ter="MSH" id="MSH_D_001249" code="D001249">asthme</des>
      <des ter="MSH" id="MSH_D_001991" code="D001991">bronchite</des>
      <des ter="MSH" id="MSH_D_002648" code="D002648">enfant</des>
      <des ter="SNO" id="SNO_D_251000" code="D2-51000">bronchite asthmatique</des>
      <des ter="SNO" id="SNO_S_10170" code="S-10170">enfant</des>
    </descripteurs>
  - <expansions>
      <des ter="CCA" id="CCA_GLRP_001" code="GLRP001">GLRP001 - Séance de réentraînement à l'exercice d'un enfant
        asthmatique, sur machine</des>
      <des ter="CIM" id="CIM_LIB_16709" code="J45.9">J45.9 - bronchite asthmatique SAI</des>
      <des ter="CIM" id="CIM_LIB_16695" code="J44.8">J44.8 - bronchite asthmatique (obstructive) SAI chronique</des>
      <des ter="CIM" id="CIM_LIB_27093" code="J44">J44 - bronchite asthmatique (obstructive) chronique</des>
      <des ter="SNO" id="SNO_D_230500" code="D2-30500">bronchite chronique obstructive</des>
    </expansions>
  </interpretation>
```

FIG. 2.7 – Fichier XML retourné par l'interpréteur de la requête « bronchite asthmatique chez l'enfant »

documents papier numérisés en texte intégral sans domaine métier ciblé (comme dans le cas des documents d'archives et du patrimoine).

2.2 Travaux de recherche au sein du LERTIM

Présentation de l'équipe et de ses travaux de recherche

Le LERTIM (Laboratoire d'Enseignement et de Recherche sur le Traitement de l'Information Médicale) est un laboratoire spécialisé dans le traitement de l'information médicale. Le laboratoire est localisé à la Faculté de médecine de Marseille, Université de Méditerranée. Le laboratoire a été labellisé par le Ministère de la recherche : équipe d'accueil (EA 3283).
La recherche au laboratoire s'intéresse à l'élaboration de systèmes d'information hospitaliers performants (adaptés et évolutifs) Fieschi (2005).
L'activité du LERTIM concerne, entre autre, la biostatistique, l'aide à la décision, les systèmes d'information médicaux et de santé, les systèmes d'information pour la formation à distance et le soutien méthodologique en recherche clinique.
Le LERTIM s'intéresse aussi à la représentation et la modélisation des connaissances pour faciliter l'accès aux connaissances et leur acquisition. Les recherches dans ce domaine visent à élaborer des méthodes et développer des outils permettant un couplage entre connaissances médicales et informations sur le patient afin d'améliorer la décision médicale et la prise en charge du patient. Le projet ASTI Bouaud *et al.* (2002) se proposait de concevoir et d'évaluer une 2^{e} génération de système informatisé d'aide à la prescription. Une série de projets, ARIANE Joubert *et al.* (2002), VUMeF Darmoni *et al.* (2003b), COMeDIAS Joubert *et al.* (2003) et WRAPIN Joubert *et al.* (2007), ont eu pour but de permettre aux professionnels de santé d'accéder à des bases d'informations du domaine biomédical (bases de données patients, banques de données sur les médicaments, guides de bonnes pratiques, bibliographie) dans le système d'information de leur entreprise ou sur le net grâce à un ensemble de services en partenariat avec Health On the Net[9] en particulier. Enfin, le projet InterSTIS (voir section 2.4) dont fait partie ce travail de recherche, a pour but de rendre les principales terminologies médicales francophones interopérables.

[9] http://www.hon.ch/

2.3 Travaux de recherche au sein de l'équipe TIBS

2.3.1 Présentation de l'équipe

Le LITIS EA 4108[10](Laboratoire d'Informatique, de Traitement de l'Information et des Systèmes) est l'unité de recherche dans le domaine des Sciences et Technologies de l'information et de la Communication (STIC) de Haute-Normandie. C'est un laboratoire pluridisciplinaire associant praticiens et théoriciens à la jonction de l'informatique, de la reconnaissance de formes, du traitement du signal et des images, de la médecine et des mathématiques.
L'équipe TIBS[11](Traitement de l'Information en Biologie - Santé) est une équipe de l'axe « Traitement des Masses de Données » du Laboratoire LITIS. L'équipe est née de la fusion de deux équipes : GCSIS (Gestion de la Connaissance et Systèmes d'Information en Santé) dirigée par le Professeur Stefan Darmoni et ABISS (Atelier Biologie, Informatique, Statistique, Sociolinguistique) en 2007. Les principaux axes de recherche de la nouvelle équipe se penchent sur les problématiques de la recherche, de l'indexation et de l'extraction des informations pertinentes, en prenant comme champ d'application les données biologiques et les systèmes d'information en santé.

2.3.2 Travaux de l'équipe

En plus de ma thèse de recherche qui a démarré avec la naissance de l'équipe TIBS, j'ai travaillé sur plusieurs problématiques connexes à ma thèse. Cependant trois travaux principaux ont primé :

Distance sémantique entre ressources : Ce travail de recherche a été commencé lors de mon stage de master ITA (Informatique Théorique et Applications) à l'université de Rouen Merabti (2007). L'idée était de concevoir un algorithme « CISMeF related resources » (CISMeF_RRA) Merabti *et al.* (2008) permettant de calculer la similarité entre les ressources du catalogue CISMeF. Cet algorithme s'inspire largement de la fonction développée par PubMed « Related Articles » Kim *et al.* (2001). Notre algorithme combine deux distances pour le calcul de similarité entre les ressources : lexicales (sur l'ensemble des mots de titre et résumé) et sémantiques (relations sémantiques entre les mots d'indexation de chaque ressource).

[10] http://www.litislab.eu
[11] http://www.chu-rouen.fr/tibs/

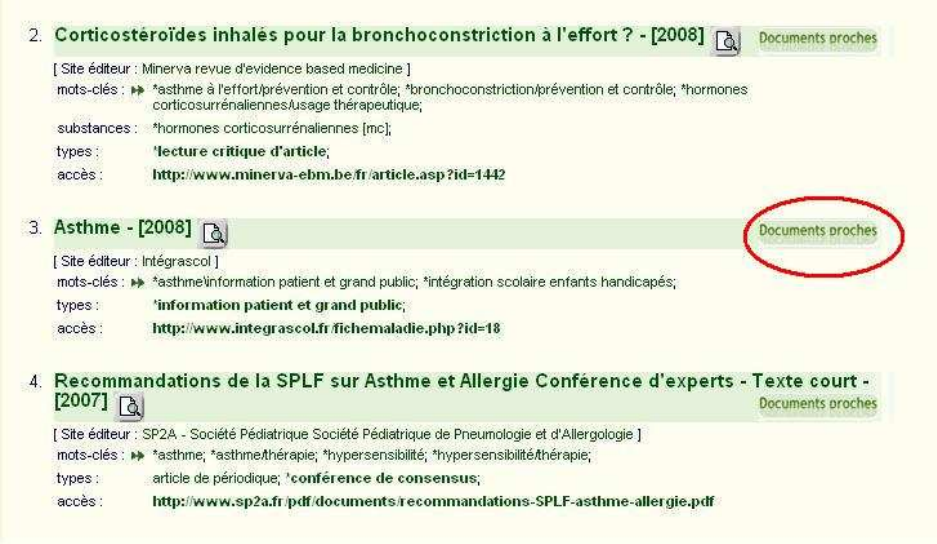

FIG. 2.8 – Ressources proches dans CISMeF

Identification des répétitions dans les navigations dans CISMeF : Ce travail a été fait dans le cadre d'un stage de master ITA en 2008 de Mohamed El-Abed El-Abed (2008). L'idée sous-jacente est la même que celle du travail précédent, qui à partir de la consultation d'une ou plusieurs ressources, propose une liste de liens susceptibles de contenir l'information recherchée par l'utilisateur. Le travail présente un algorithme d'extraction de comportements récurrents durant la consultation de ressources au sein du catalogue de santé CISMeF Pauchet *et al.* (2009). Nous avons proposé pour cela d'utiliser la structure de données appelée arbres des suffixes Weiner (1973); McCreight (1976), appliquée aux fichiers log de CISMeF. Parallèlement à cela, nous nous intéressons à l'identification de ressources pertinentes pour une requête donnée, en construisant un ensemble de ressources syntaxiquement et sémantiquement proche des ressources consultées au cours de la navigation. Son principe reste identique au précédent.

Détection et désambiguïsation des abréviations : dans le cadre du stage de master ITA en 2008 d'Ismail Mansour Mansour (2008), nous avons travaillé sur un algorithme de détection automatique des abréviations ambiguës dans les ressources médicales. L'algorithme que nous avons proposé est fondé sur la structure de données des arbres de suffixes.

2.4 Le projet InterSTIS (Interopérabilité Sémantique des Terminologies dans les systèmes d'Information de Santé français)

Le projet InterSTIS (Interopérabilité Sémantique des Terminologies dans les systèmes d'Information de Santé français) a été financé par l'appel à propositions TecSan 2007 lancé par l'Agence Nationale pour la Recherche (ANR) pour trois ans (janvier 2008- décembre 2010), a pour but de fédérer et de rendre interopérables les principales terminologies médicales au sein d'un « serveur terminologique multi-sources » (STMS) (voir section 3.6).
Le consortium du présent projet est constitué de trois sociétés industrielles spécialisées, entre autre, dans la représentation des connaissances, le langage naturel, et le langage médical :

- VIDAL SA, partenaire coordinateur, Paris
  http://www.vidal.fr
- Mondeca, Paris
  http://www.mondeca.com
- Memodata, Caen
  http://www.memodata.com

Quatre équipes hospitalo-universitaires spécialisées dans les terminologies médicales, les systèmes d'information et la diffusion d'information de santé :

- LERTIM[12], Faculté de Médecine, Université de la Méditerranée, Marseille
  http://cybertim.timone.univ-mrs.fr
- CISMeF, CHU de Rouen
  http://www.chu-rouen.fr/cismef
- DSPIM, Faculté de Médecine, Université Jean Monnet, Saint Etienne
  http://dossier.univ-st-etienne.fr/dspim/www/
- LabSTIC, Faculté de Médecine, Université de Nice-Sofia Antipolis
  http://portail.unice.fr/jahia/page4693.html

Une équipe du CNRS spécialisée dans le traitement du langage naturel :

- LIMSI, CNRS et Université Paris-Sud 11, ORSAY
  http://www.limsi.fr

Une fondation spécialisée dans la compréhension du langage naturel et la recherche d'information certifiée dans le domaine de la santé :

[12]Responsable scientifique.

– HON, partenaire associé, Genève
http://www.hon.ch/

Les objectifs d'InterSTIS se déclinent dans trois principales directions :

1. Modélisation des terminologies médicales francophones utilisées dans le STMS (voir section 3.3.2).
2. Intégration des terminologies médicales dans le STMS. En plus de l'intégration cette tâche va permettre l'alignement entre terminologies à l'intérieur du STMS.
3. Intégration et extension d'un lexique médical francophone.

Mon travail de recherche commencé en 2007 est entièrement financé par le projet InterSTIS. Un site internet a été mis en place www.interstis.org (figure 2.9) pour permettre aux participants de suivre régulièrement l'évolution du projet.

FIG. 2.9 – Le site InterSTIS

2.5 Synthèse

Nous avons présenté dans ce chapitre le contexte générale de cette thèse. Nous avons décrit brièvement les différents travaux de chacune des équipes impliqués. Les équipes CISMeF et LERTIM travaillent depuis quelques années sur des problématiques proches liées principalement aux terminologies médicales. Elles ont également lancé plusieurs collaborations sur différents projets (les projets UMLF et VUMeF). Nous avons vu aussi que la fusion des deux équipes GCSIS (Gestion de la Connaissance et Systèmes d'Information en Santé) et ABISS (Atelier Biologie, Informatique, Statistique, Socio-linguistique) en 2007, a permis l'ouverture sur de nouveaux axes de recherche dans les domaines de la recherche, de l'indexation et de l'extraction des informations pertinentes, en prenant comme champ d'application les données biologiques et les systèmes d'information en santé.
Le projet InterSTIS (Interopérabilité Sémantique des Terminologies dans les Systèmes d'information de Santé Français), a pour but de fédérer et de rendre interopérables les principales terminologies médicales au sein d'un « Serveur Multi-Terminologique de Santé » (SMTS).
Notre travail qui s'inscrit dans le cadre de ce projet, vise à mettre en œuvre des méthodes permettant de contribuer à l'interopérabilité entre les différentes terminologies francophones qui seront intégrées dans le SMTS.
Dans le prochain chapitre, nous allons décrire les différentes terminologies médicales utilisées dans le cadre de cette thèse. Nous décrirons en détail la problématique liée à l'intégration des terminologies au sein d'un même serveur multi-terminologique.

Chapitre 3

État de l'art

Dans ce chapitre, nous décrivons les terminologies francophones utilisées dans de nos travaux de recherche. Ce chapitre traite la problématique de l'intégration des terminologies au sein d'un même serveur multi-terminologique. Nous détaillerons dans cette partie principalement le serveur multi-terminologique de santé, le cœur des différents projets de recherches entamés il y a trois ans dans plusieurs laboratoires de recherche spécialisés dans le traitement de l'information médicale. Dans la deuxième partie de ce chapitre nous listons les principaux termes utilisés pour définir le mécanisme de mise en relation des terminologies. Nous proposons aussi une classification des différentes méthodes d'alignements inspirée de Euzenat et Shvaiko (2007) et leurs travaux sur les alignements entre les ontologies

3.1 Éléments de représentation

Le langage médical est caractérisé par un vocabulaire extrêmement riche et difficile à manipuler. Les termes utilisés sont souvent très imprécis et font rarement l'objet de définitions rigoureuses. Dans ce type de langage, il existe plusieurs façons d'exprimer la même chose (synonymies), ainsi que plusieurs interprétations possibles pour des termes similaires. Cette situation n'empêche par le personnel médical de communiquer mais complique considérablement l'automatisation de ces communications. Ainsi, pour traiter l'information médicale avec une « machine », il faut fournir un modèle formel Zweigenbaum (1999). Ce modèle est formé de l'ensemble des termes du langage et des relations qui permettent de relier des concepts généraux à des concepts plus spécifiques. Plusieurs modèles existent, les principaux (pour le domaine médical) sont

la terminologie et l'ontologie. Dans une terminologie, on s'intéresse aux mots et aux relations entre eux ; la relation structurante de base est la relation d'hyperonymie et son inverse l'hyponymie, tandis que dans une ontologie, on s'intéresse aux concepts et aux relations entre eux Smith *et al.* (2005).

3.1.1 Terminologies

Dans Roche (2005), une terminologie est définie comme un ensemble de mots. Une définition plus précise de la terminologie est donnée dans Lefevre (2000) : « Les terminologies sont 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 ». Formellement, Smith (2006) définit une terminologie comme un triplet ordonné :
$T = \langle N, L, v \rangle$ où :

- N représente aussi un ensemble de triplets $\langle p, S_p, d \rangle$ appelés des noeuds où p représente le libellé unique (nommé aussi un terme préféré), S_p un ensemble de synonymes $s, s', s'', \ldots$ et d une définition optionnelle attachée au noeud.
- L un ensemble de pairs ordonnées $\langle r, L_r \rangle$ appelées des liens où r représente une relation de type (« is_a » ou « $part\_of$ »), et L_r représente une paire ordonnée $\langle s, s' \rangle$ de termes. Ainsi, « s, r, s' » représente une relation dans la terminologie entre s et s'.
- v est un nombre qui représente la version de la terminologie.

La norme ISO ISO (2000) propose la meilleure et la plus simple des définitions « ensemble de désignations propre à une langue de spécialistes », les « désignations » peuvent être des termes (avec plusieurs statuts : termes préférentiels, synonymes, noms, symboles...). De ce fait, le contenu et la structure d'une terminologie dépendent de la fonction pour laquelle cette terminologie va être utilisée. Dans une terminologie médicale (ou système terminologique médical), des termes précis sont utilisés pour spécifier les concepts du domaine. Des relations peuvent aussi exister entre les termes. Par exemple, des relations de généralisation-spécialisation sont prises en compte par plusieurs terminologies permettant de hiérarchiser les termes du plus global au plus précis. Dans une terminologie, les concepts peuvent être désignés par plusieurs termes différents. Nous parlerons dans ce cas aussi d'un « système de concept » qui est défini dans ISO (2000) comme un ensemble de concepts structurés selon des relations entre eux. **Un terme préférentiel** désigne le nom du concept et plusieurs synonymes. En plus, les terminologies peuvent être multilingues (toutes les formes équivalentes sous le même concept) dans des langues différentes.
Mori *et al.* (1998) décrit l'évolution des terminologies en terme de trois générations.

- Terminologie de première génération « First generation »: Elles sont caractérisées par une organisation fixe (hiérarchie simple) et une simple représentation comme une liste indexée d'une façon alphabétique. Par exemple, la classification CIM10 OMS (1993) ou le thésaurus MeSH;
- Terminologie de deuxième génération « Second generation »: Elles sont caractérisées par une organisation dynamique (hiérarchie multiple) avec une indexation multiple. Par exemple, le dictionnaire médical des activités de règlementation MedDRA Brown *et al.* (1999) ou la SNOMED International Côté *et al.* (1993);
- Terminologie de troisième génération « Third generation »: Elles sont fondées sur un modèle formel avec des symboles permettant de dénoter des concepts et un ensemble de règles permettant de les manipuler. Par exemple, la SNOMED CT Spackman (2000), GALEN Rector *et al.* (1993).

D'un autre côté, une classification des différentes terminologies médicales a été définie dans la littérature de Keizer *et al.* (2000) ou dans les différentes normes du domaine ISO (2007, 2000). Ces classifications ont été établies en fonction des différents objectifs pour le traitement de l'information en plus d'un certain nombre de caractéristiques propres à chacune des terminologies :

Vocabulaire contrôlé Un ensemble de termes sans organisation logique (en général) accompagnés de leurs définitions. Cette définition englobe les termes « dictionnaires terminologiques », « vocabulaires » et « gloassaires » définies dans ISO (2000).

Classification Une classification représente un ensemble de termes organisés et hiérarchisés en classes et sous-classes de Keizer *et al.* (2000). Cette définition donne une vision plus simple de celle donnée dans ISO (2007) où elle définit une classification comme « un ensemble exhaustif de catégories mutuellement exclusives permettant le regroupement des données à un niveau de spécialisation spécifique ». La structure de la classification et la granularité des classes dépend des objectifs pour lesquels elle a été conçue. L'ATC (classification Anatomique, Thérapeutique) (voir section 4.3.1), la CCAM (Classification Commune des Actes Médicaux) (voir section 4.4.1) et la CIM10 (Classification Internationale des Maladies version 10) sont de bons exemples de classifications hiérarchiques médicales.

Nomenclature Elle désigne un ensemble de termes techniques, présentés selon un classement méthodique. Cette définition est la même utilisée dans ISO (2000) pour désigner une nomenclature. La nomenclature vise à recenser les termes d'un domaine de façon exhaustive. Les termes de la nomenclature peuvent être répartis selon plusieurs axes. Cette répartition permet de composer un concept complexe par combinaison de plusieurs concepts. Une nomenclature importante dans le domaine clinique à laquelle nous nous intéressons ici (voir section 6) est la Nomenclature Systématique des Médecines Humaine et Vétérinaire de Côté

et al. (1993).

Thésaurus Est un ensemble structuré de termes d'un vocabulaire. Les termes sont organisés de manière conceptuelle et reliés entre eux par des relations sémantiques. Trois types de relations entre les termes existent : relation hiérarchique (spécialisation-généralisation, tout-partie), relation d'équivalence (synonymes) et relation d'association pour les sujets connexes. La terminologie MeSH introduite lors de la présentation du projet CISMeF et définie en détail dans la section 3.1.3, est un thésaurus.

Lors de l'utilisation des codes pour désigner chaque terme dans ces types de terminologie, nous parlerons alors d'un système de codage. Désigner comme un système terminologique, un système de codage est défini dans ISO (2007) comme une combinaison d'un ensemble de concepts, d'un ensembles de codes et d'au moins d'un schéma de mapping entre codes et concepts. Notons aussi que la notion d'ontologie (définie dans la section 3.1.2) est utilisée comme synonyme pour un certain type de terminologies.

La tableau 3.1 Névéol (2005) résume les principales caractéristiques de chaque type de terminologie.

Type de terminologie	**Caractéristiques**
Vocabulaire contrôlé	définition des termes
Classification	structuration liens nommés entre les termes
Nomenclature	exhaustivité structuration
Thésaurus	normalisation des termes réduction des ambigüités

TAB. 3.1 – Les types de terminologies et leurs caractéristiques

3.1.2 Ontologie

L'ontologie comme discipline philosophique est définie comme la science qui s'occupe de ce qui est, des genres et des structures des objets, des propriétés, des événements, des relations dans tous les secteurs de la réalité Smith (2003). Depuis environ deux décennies, la communauté informatique a commencé à s'intéresser aux ontologies.

Leur importance est largement reconnue dans divers domaines de recherche Guarino (1998), tels que l'ingénierie des connaissances Gruber (1993); Uschold et Grüninger (1996) et la représentation des connaissances Guarino (1995); Sowa (2000).

La première définition de l'ontologie dans le domaine informatique est donnée par Gruber comme « a specification of a conceptualization » Gruber (1993). Bien que Smith pense que la contribution de Gruber soit la première tentative de définition crédible, elle laisse cependant la place à d'autres interprétations possibles Smith et Welty (2001). Selon Smith, des systèmes d'information tels que des catalogues, des glossaires, des thésaurus satisfont la définition de Gruber. Néanmoins, elle exprime une idée intuitive qui reste vraie pour le sens de l'ontologie, tel qu'il est employé dans la grande majorité des travaux. Zweigenbaum (1999) présente l'ontologie comme l'aboutissement formel de la définition d'une terminologie.
D'une manière générale, une ontologie fournit les moyens d'exprimer les concepts d'un domaine en les organisant hiérarchiquement et en définissant leurs propriétés sémantiques dans un langage de représentation des connaissances formel Bourigault *et al.* (2004). La relation hiérarchique « généralisation-spécialisation » est unique, ce qui permet de définir clairement la subsomption entre concepts. Des exemples d'ontologies sont les ontologies GALEN (General Architecture for Language and Nomenclatures) Rector *et al.* (2003) et FMA (Foundational Model of Anatomy) Rosse et Mejino (2003).

3.1.3 Les principales terminologies médicales

Dans le cadre de cette thèse, nous avons utilisé un certain nombre de terminologies médicales de différents types. La plupart sont traduites en français. Nous définissons dans cette section six terminologies importantes pour la suite de nos travaux car elles sont incluses dans UMLS et traduites en français (F_UMLS): MeSH (Medical Subject Headings), CIM10 (Classification Internationale des Maladies version 10), CISP2 (Classification Internationale des Soins Primaires, deuxième version), SNOMED 3.5 (Systematized Nomenclature Of MEDicine), MedDRA (Medical Dictionary for Regulatory Activities), WHO-ART (World Health Organisation - Adverse Reaction Terminology). D'autres terminologies vont être définies dans différentes sections en fonction de leur utilisation.

MeSH (Medical Subject Headings) :

Une première liste officielle de sujets a été publiée par la NLM (National Library of Medecine) états-unienne en 1954. La première version du MeSH a été publiée en 1960 pour indexer les articles scientifiques dans le système bibliographique biomédical automatisé de stockage et de recherche MEDLARS Austin (1968) (devenu depuis MEDLINE regroupant plus de 18 millions d'articles) Bachrach et Charen (1978). Le projet CISMeF utilise la terminologie MeSH pour l'indexation des ressources francophones disponibles gratuitement sur Internet. Le MeSH est traduit en 11 langues (français, anglais, espagnol, ...). Toutes ces traductions sont présentes dans UMLS. L'INSERM (Institut National de la Santé Et de la Recherche Médicale) a élaboré une version française du MeSH[1]. Une nouvelle version apparaît tous les ans, la dernière en date est la version 2010. Il existe au maximum 11 niveaux hiérarchiques dans le MeSH avec des relations de « spécialisation-généralisation » et « tout-partie » divisée en 15 arborescences thématiques auxquelles correspond un code spécifique : « A » pour « anatomie », « B » pour « organisme », « C » pour « maladie » etc. La figure 3.1 présente un extrait de l'arborescence : C « maladie ». Dans sa version 2010, le MeSH comporte 25 588 mots clés, 84 qualificatifs ainsi que 186 702 concepts chimiques supplémentaires. Les qualificatifs sont des termes qui peuvent être associés à un mot clé afin d'en préciser le sens Darmoni *et al.* (2007). Par exemple, « cancer des os/traitement médicamenteux » permet de restreindre le cancer des os au seul aspect du traitement médicamenteux (qualificatif). Les qualificatifs sont organisés hiérarchiquement du plus générique au plus précis.

Deux autres types de relations existent :

- La relation « voir aussi » permet de naviguer d'un mot clé à l'autre et de relier des termes proches.
- La relation « ne pas confondre » permet de préciser le sens et de lever les ambiguïtés.

Les types de publication sont des termes utilisés pour l'indexation du contenant dans la NLM. Ces termes ont servi de référentiel de départ pour créer les types des ressources utilisés dans CISMeF.

MedDRA (Dictionnaire médical des activités de réglementation) :

La terminologie MedDRA, une initiative de la conférence internationale sur l'harmonisation (ICH) OMS (1993), est un dictionnaire uniformisé de terminologies médi-

[1] http://ist.inserm.fr/mesh/html/mesh.html

FIG. 3.1 – Extrait de l'arborescence C (Maladies) du MeSH

cales Brown *et al.* (1999). Il est destiné au partage de renseignements de réglementation à l'échelle internationale sur les produits médicaux destinés à l'usage humain. Depuis janvier 2003, la terminologie médicale de MedDRA sert aux échanges électroniques d'informations et d'observations de pharmacovigilance à l'échelle internationale. MedDRA est aussi utilisé pour les effets secondaires dus aux instruments médicaux. MedDRA est disponible en plusieurs langues dont le français, l'anglais, l'espagnol ou le japonais. Le support de maintenance de MedDRA est assuré par le MSSO (Maintenance and Support Services Organization).[2]
MedDRA est construit selon une hiérarchie constituée de 26 classes de haut niveau (SOC) permettant de définir et traduire les renseignements médicaux selon 5 niveaux de précision :

Classe Organes/System Organ Class (SOC): il s'agit du plus haut niveau de la hiérarchie qui offre le plus large concept pour le regroupement des données par :

- Étiologie (Infections and infestations)
- Site d'atteinte (Gastrointesinal disorders)
- Action (surgical and medical procedures)

Termes de haut niveau/High Level Term (HLT): regroupent des termes préférés (PT) ayant en commun un lien anatomique, physiopathologique, étiologique ou fonctionnel.

Le terme préféré/Preferred Terms (PT) : est un terme décrivant un concept médical unique. Il doit être le moins ambigu et le plus spécifique et auto-descriptif possible. Un PT doit être relié à au moins un SOC.

Groupes de termes de haut niveau/High Level Group Term (HLGT) : regroupent plusieurs HLT ayant un lien anatomique, physiopathologique, étiologique ou fonctionnel.

Termes de bas niveau/Low Level Terms (LLT) : est le niveau préférentiel de codage, il couvre en effet le plus grand nombre d'entrées possibles. Chaque LLT est relié à un seul PT.

Le tableau 3.2 donne des exemples pour chaque type de termes ainsi que le nombre de termes dans MedDRA suivant chaque type.

Une classe organe regroupe l'ensemble des concepts liés à un organe. Chaque terme préféré est associé à une classe organe unique et peut appartenir de façon optionnelle à une ou plusieurs classes organes secondaires. Par exemple la néphropathie diabétique appartient à la classe organe des troubles rénaux mais il existe un lien secondaire vers la classe organe des troubles métaboliques. Les deux types de termes HLT et HLGT sont

[2] www.meddramsso.com

Type de terme	**Exemple de terme**	**Nombre de termes dans MedDRA**
System Organ Class (SOC)	Troubles du foie et des voies biliaires	26
High Level Group Term (HLGT)	Maladies hépatobiliaires	332
High Level Term (HLT)	Hépatite	1 682
Preferred Term (PT)	Adipose douloureuse de Dercum	17 867
Low Level Terms (LLT)	Syndrome abdominal aigu	56 580

TAB. 3.2 – Exemples et nombre de termes MedDRA suivant chaque type de terme

utilisés uniquement pour l'extraction de données et leur présentation. Ils ne sont pas utilisés pour le codage. Le schéma de la figure 3.3 représente la distribution hiérarchique dans la terminologie MedDRA. De plus, le dictionnaire MedDRA intègre des Requêtes Standard MedDRA (RSM) (SMQ en anglais). Les RSM sont des regroupements de termes qui se rapportent à un domaine médical spécifique (voir figure 3.2).

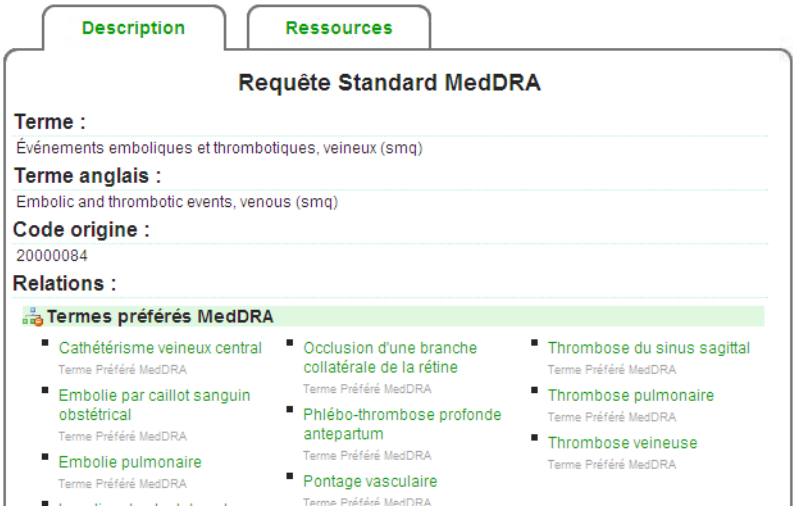

FIG. 3.2 – Exemple d'une requête Standard MedDRA

WHO-ART (World Health Organisation - Adverse Reaction Terminology) :

WHO-ART est une terminologie utilisée principalement pour le codage des effets indésirables des médicaments. Développée et maintenue par l'OMS (Organisation Mondiale de la Santé) WHO (1992), la structure de WHO-ART est assez simple pour permettre l'intégration de nouveaux termes correspondant à des médicaments ou à de nouvelles indications. La terminologie WHO-ART est structurée hiérarchiquement suivant quatre niveaux :

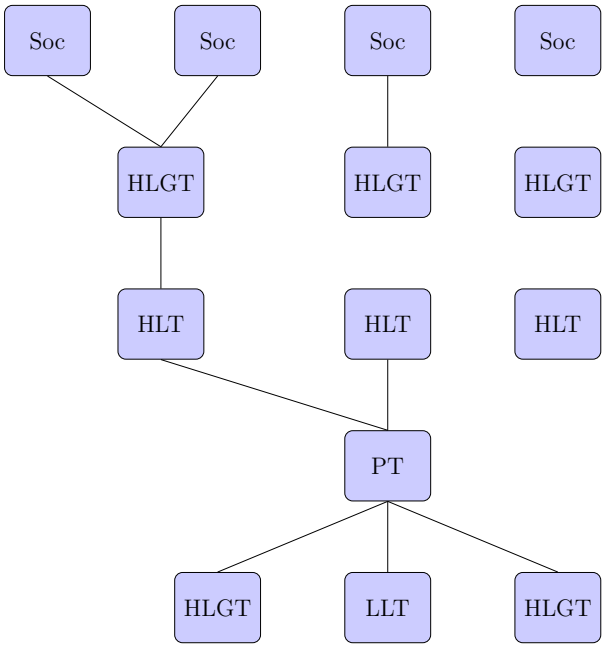

FIG. 3.3 – Schéma récapitulatif de la hiérarchie MedDRA

Catégorie des systèmes ou organes : ce sont des groupes de termes d'effets indésirables relatifs au système du même organe.

Termes de haut niveau : ce sont des termes permettant de grouper des termes similaires sous certaines conditions. Par exemple, le terme de haut niveau « thrombophlébits » regroupe les deux termes (termes préférés) « jambe thrombophlébits » et « bras thrombophlébits ».

Termes préférés : ce sont les termes principaux utilisés pour le codage des effets indésirables. Il ont les mêmes propriétés que peuvent avoir des termes préférés dans une terminologie médicale.

Termes inclus : représentent le dernier niveau de terme dans la terminologie WHO-ART. Ce sont les termes synonymes des termes préférés. Utilisés pour aider à trouver les bons termes préférés dans le processus de codage.

Le tableau 3.3 dresse le nombre ainsi que quelques exemples de termes utilisés[3] dans chaque niveau de la terminologie WHO-ART.

La figure 3.4 montre une portion de la hiérarchie WHO-ART pour les termes de l'exemple du tableau 3.3.

[3]les termes de WHO-ART sont tous en majuscules non accentués

FIG. 3.4 – Portion de la hiérarchie WHO-ART pour la catégorie « Système vasculaire extra-cardiaque »

Type de terme	**Exemple de terme**	**Nombre de termes dans WHO-ART**
Catégorie des systèmes ou organes	SYSTEME VASCULAIRE EXTRA-CARDIAQUE	32
Termes de haut niveau	ANGEITE	162
Termes préférés	ARTERITE	1 532
Termes inclus	ENDARTERIE MALADIE DE HORTON PANARTERITE	19 991

TAB. 3.3 – Exemples et nombre de termes WHO-ART suivant chaque type de terme

CISP2 (la Classification Internationale des Soins Primaires, deuxième version) :

CISP2 Jamoulle *et al.* (2000) est la version française de l'International Classification of Primary Care (ICPC), développée par l'Organisation internationale des médecins généralistes. Elle appartient à la famille des classifications de l'OMS comme classification associée à la Classification Internationale des Maladies (CIM).
Depuis la création de la première version de l'ICPC (ICPC-1) en 1987, elle a été traduite en plus d'une vingtaine de langues. Elle a été publiée en langue française dans sa première version (CISP-1) en 1992, puis dans sa deuxième version (CISP-2) en 2000. Elle est aussi disponible en format électronique (CISP-2-E), permettant son intégration dans les dossiers médicaux informatisés. Elle a été développée initialement pour le recueil manuel et l'analyse épidémiologique des données de consultation en médecine générale. Dans le cadre du dossier médical informatisé, elle peut être utilisée avec des systèmes d'aide à la décision (diagnostique ou thérapeutique), d'assurance qualité des soins, de surveillance épidémiologique et de recherche scientifique en soins primaires. La CISP est une classification bi-axiale, dont le premier axe est composé de 17 chapitres désignant chacun un appareil corporel (incluant les chapitres psychologique et social) et le second axe de 7 composants (symptômes et plaintes, procédures diagnostiques et préventives, procédures thérapeutiques, résultats d'examens complémentaires, procédures administratives, références et autres motifs de contact, diagnostics et maladies).

SNOMED International (Systematized Nomenclature Of MEDicine - Nomenclature systématique de médecine) :

SNOMED est une terminologie clinique développée à l'origine par le Collège des Pathologistes Américains (CAP) en 1955. La première nomenclature publiée étais SNOP (Systematized Nomenclature Of Pathology), nomenclature fonctionnelle pour les pathologies. En 1973, le Dr Côté fait évoluer la SNOP vers la SNOMED (Systematized Nomenclature of Medecine) Côté (1972) qui devient en 1993 Côté *et al.* (1993) la SNOMED version 3.5, nommée aussi SNOMED International. SNOMED International est une nomenclature pluri-axiale couvrant tous les champs de la médecine et de la dentisterie humaine, ainsi que la médecine vétérinaire. Elle est traduite en 11 langues (français, espagnol, japonais, turc, ...). La version française a été réalisée par l'équipe du Centre de Recherche en Diagnostic Médical Informatisé (CRDMI), et qui s'est achevée en 2006 en partie grâce au projet VUMeF Darmoni *et al.* (2003b). La SNOMED International est multi-axiale sur 11 axes. Dans chaque axe, les concepts sont représentés par une série de termes au sein de laquelle on peut distinguer une formulation préférée référencée par des codes alphanumériques uniques et des synonymes de diverses natures syntaxiques. Chaque axe recense les termes d'un sous-domaine de la médecine. Par exemple : D pour Diagnostic, T pour Topographie, ... Par ailleurs, chaque axe est hiérarchisé en fonction de la spécialisation des concepts, qui sont reliés par des relations d'hyponymie et de méronymie. Notons qu'il existe aussi des relations transversales plus complexes (entre concepts appartenant à des axes différents). Le tableau 3.4 liste tous les axes que comporte la SNOMED International.

La Classification Internationale des Maladies : version 10

L'appellation complète de la Classification Internationale des Maladies est « Classification statistique internationale des maladies et des problèmes de santé connexes » (en anglais : International Statistical Classification of Diseases and Related Health Problems). La désignation usuelle abrégée de « Classification internationale des maladies » est à l'origine du sigle couramment utilisé pour la désigner : « la CIM » (en anglais : ICD). La CIM permet le codage des maladies, des traumatismes et de l'ensemble des motifs de recours aux services de santé. La CIM-10 est une classification mono-axiale, elle a été publiée en 1993 par l'Organisation Mondiale de Santé (OMS)[4] OMS (1993) et est utilisée à travers le monde pour enregistrer les causes de morbidité et de mortalité, à des fins diverses parmi lesquelles le financement et l'organisation des services de santé qui ont pris, ces dernières années, une part croissante.

[4] http://www.who.int/classifications/icd/en/

Axe	**Nom de l'axe**	**Nombre de termes**
T	Topographie	13 528
M	Morphologie	6 171
F	Fonctions	20 587
A	Artefacts, activités physiques	1 686
L	êtres vivants	26 325
C	produits chimiques	15 940
J	Métiers	2 303
S	Contexte social	1 110
D	Diagnostic	42 492
P	Actes	31 980
G	Qualificatifs	1 595
X		363
Total		164 180

Tab. 3.4 – Les axes de la SNOMED International

Elle a été conçue pour permettre l'analyse systématique, l'interprétation et la comparaison des données de mortalité et de morbidité recueillies dans différents pays ou régions à différentes époques. Son histoire a commencé avec la Classification des causes de décès de Jacques Bertillon (1893). Cette classification connut cinq révisions décennales jusqu'en 1938. À sa création en 1945, l'OMS se vit confier l'évolution de la classification de Bertillon qui devint en 1948, avec la sixième révision, la « Classification statistique internationale des maladies, traumatismes et causes de décès » : elle cessait en effet de ne répertorier que les causes de décès pour s'intéresser de façon plus générale à la morbidité, alors que la CIM10 permet le recueil de diagnostics à des fins de santé publique ou d'évaluation de l'activité hospitalière pour le codage médico-économique des dossiers patients à des fins statistiques et budgétaires. La CIM10 est ordonnée en une hiérarchie à héritage simple. La hiérarchie de la CIM10 a été un processus complexe réalisé par l'OMS, mais a le mérite d'être clair. Cela signifie que toute entité hiérarchique ou rubrique possède un et un seul père (sauf les entités du niveau 1 au sommet de la pyramide qui n'ont pas de père). À tout moment, il est possible pour toute entité hiérarchique donnée de reconstituer la liste exhaustive de tous ses ancêtres. La hiérarchie de la CIM10 a jusqu'à 6 niveaux, bien que plusieurs chapitres n'en aient que 5. La CIM10 est divisée en 21 chapitres couvrant l'éventail complet des états morbides classés par appareil fonctionnel et associés à une lettre (exemple : F : « Troubles mentaux et du comportement »).

- Chapitre 5 : Troubles mentaux et du comportement (F00-F99)
  - Groupe : Troubles de l'humeur (F30-F39)
    - Catégorie : Épisode maniaque (F30)

      ...Hyponamie (F30.0)

      ...Manie sans symptôme psychotique (F30.1)

      ...Manie avec symptôme psychotique (F30.2)

      ...Autres épisodes maniaques (F30.8)

      ...Épisodes maniaques, sans précision (30.9)
    - Catégorie : Trouble affectif bipolaire (F31)

      .....

FIG. 3.5 – Extrait de la classification CIM10

Les chapitres sont toujours au niveau le plus élevé de la hiérarchie de la CIM10, ils sont divisés en groupes (ou blocs), eux-mêmes divisés en sous-groupes (ou sous-blocs), ce qui est facultatif (utilisé seulement dans les chapitres 2, 13, 19 et 20) composés de catégories à 3 caractères et de sous-catégories à 4 caractères, englobant le contenu des termes CIM10. Les catégories à 3 caractères représentent l'unité diagnostique de base, signifiante et présentée comme le niveau minimum de codification entrant dans les comparaisons internationales. Toutefois, de nombreux pays exigent le niveau suivant à 4 caractères comme niveau minimum de codification (c'est le cas de la Suisse par exemple). Alors que pour les sous-catégories à 4 caractères, il y a une spécialisation des catégories en 10 parties au maximum numérotées de 0 à 9. Et finalement des subdivisions ou descripteurs peuvent apparaître de manière facultative dans certains chapitres, permettant d'introduire un axe classificatoire systématique supplémentaire là ou cela s'avère nécessaire, du fait que le niveau du ou des blocs n'appartient pas à tous les chapitres.

3.2 Unified Medical Language System (UMLS)

En 1986, la NLM (National Library of Medecine) a lancé un programme de développement sur plusieurs années, nommé « Unified Medical Language System » (UMLS) Lindberg *et al.* (1993). Ce projet associait plusieurs équipes de recherche et compagnies commerciales de différentes disciplines médicales ou informatiques. Le but du projet UMLS est de fournir une assistance automatisée en établissant des liens conceptuels

à partir de l'expression de l'utilisateur qui a besoin de l'information (question, problème) jusqu'à l'obtention d'une requête directement exploitable sur des ressources biomédicales. Ce qui permettra, de lever toutes les ambiguïtés et toutes les barrières à l'application de l'informatique au domaine médical.
Une des caractéristiques de cette automatisation est de fournir un lien entre différentes terminologies biomédicales de plusieurs sources de données terminologiques. Par conséquent, l'un des objectifs de l'UMLS et de fournir une plate-forme permettant de regrouper tous les thésaurus, nomenclatures, et classifications existantes dans le domaine médical Bodenreider (2004).
Le « métathésaurus » est la partie de l'UMLS permettant de regrouper le plus grand nombre possible de terminologies médicales disponibles. Deux autres parties composent l'UMLS : le réseau sémantique et Specialist Lexicon (dont une version a été développée par le projet UMLF et poursuivie pour InterSTIS).

Le métathésaurus

Considéré comme la plus grande base de données terminologiques, le métathésaurus constitue la base unifiée des concepts médicaux. Il comprend des synonymes, des variations lexicales et des concepts associés. La première version du métathésaurus « Meta-1 », comprenait déjà 30 000 concepts avec plus de 60 000 termes et 100 000 relations. Actuellement, la version 2009AA du métathésaurus, contient plus de 2 millions de concepts avec plus de 7 millions de termes de 140 terminologies biomédicales (dont le MeSH, la SNOMED CT, et 3.5, CIM9, CIM10, ...).
Il a fallu pour regrouper toutes les terminologies dans le métathésaurus suivre un certain nombre de règles :

1. regrouper sous un même concept les différents termes qui l'expriment. Chaque concept ajouté dans le métathésaurus recevra un unique identifiant et il sera placé dans la structure du métathésaurus. Cette structure est composée de quatre niveaux (voir tableau 3.5) :
   - Concept Unique Identifiers (CUI) : il regroupe tous les termes qui partagent le même sens. Par exemple, les termes « Froid (Cold) » (MeSH), « température froide (cold temperature) » (CSP) appartenant à différentes terminologies doivent être regroupés dans un même concept UMLS.
   - Lexical Unique Identifiers (LUI) : il regroupe toutes les variations lexicales pour un terme donné. Cependant, ce regroupement est appliqué seulement pour les termes en anglais. Par exemple, les deux termes « Headaches » et « Headaches » (céphalée) ont le même LUI.

- String Unique Identifiers (SUI) : chaque nom de concept ou terme dans chaque langue est associé à un identifiant unique SUI. De plus, chaque variation dans le nombre de caractères, la ponctuation... est considéré comme des termes différents ce qui implique des SUI différents. Par exemple, les deux termes « Adrenal Gland Diseases » (maladies de la glande surénale) et « Disease of adrenal gland » ont des SUI différents. Alors que, les termes « Cold » du MeSH et « Cold » de la SNOMED ont un même SUI.
- Atom Unique Identifiers (AUI) : chaque occurrence d'un terme dans chaque terminologie est associée à un unique identifiant AUI. Par exemple, les deux mêmes termes « Cold » du MeSH et « Cold » de la SNOMED ont des AUI différents.

2. si les mêmes concepts appartiennent à différents contextes hiérarchiques, alors toutes les hiérarchies doivent être incluses dans le métathésaurus.
3. les différentes relations entre concepts de différentes terminologies doivent être aussi incluses.

« *En d'autres termes, le métathésaurus ne représente ni une ontologie biomédical propre à la NLM, ni une seule vision du domaine biomédicale. Le métathésaurus préserve toutes les visions actuelles présentes dans toutes les terminologies* »[5].
Les principaux composants du métathésaurus sont : « concepts », « termes » et « relations ». Chacun de ces composants a un identificateur unique dans le métathésaurus. Les concepts UMLS représentent un et un seul sens distinct. À chaque concept correspond : une définition, un terme préférentiel, éventuellement des synonymes, des variantes lexicales, un ou plusieurs types sémantiques et un identifiant unique « Concept Unique Identifier » (CUI).
Plusieurs relations existent entre différents concepts. Ces sont des relations qui proviennent des terminologies d'origine et des développeurs de la NLM durant la construction du métathésaurus.

[5]UMLS Reference Manual: http://www.ncbi.nlm.nih.gov/bookshelf/br.fcgi?book=nlmumls

Concepts (CUI)	Termes (LUI)	Strings (SUI)	Atomes (AUI)
C0009264 **cold temperature**	**L0215040** cold temperature	**S0288775** cold temperature	**A0318651** cold temperature (from CSP)
	L0009264 Cold Cold	**S0007170** Cold	**A0016032** Cold (from MTH)
		S0026353 Cold	**A0040712** Cold (from MeSH)

Tab. 3.5 – Les concepts de l'UMLS

3.3 Serveur Multi Terminologique de Santé (SMTS)

3.3.1 Définition

« Le serveur multi-terminologique de santé » francophone (SMTS) Darmoni *et al.* (2009a); Joubert *et al.* (2009b) est un exemple d'outil qui permet de regrouper plusieurs terminologies mais exclusivement francophones. Trois partenaires se sont associés pour réaliser le SMTS : Le LERTIM, CISMeF et la société MONDECA. Cette dernière est spécialisée dans la gestion des terminologies et des ontologies ainsi que du Web sémantique. Le but principal du SMTS est l'intégration de plusieurs terminologies dans un même et unique serveur pour les exploiter simultanément. Actuellement, le SMTS intègre plus de 11 terminologies médicales francophones.

Outre la gestion des terminologies de santé francophones, le SMTS va permettre aux professionnels de santé ainsi qu'aux applications un accès en temps réel à toutes les terminologies francophones. La figure 3.6 proposée par MONDECA représente une vision de l'architecture en trois parties du système. Le premier niveau de ce schéma représente tous les outils permettant la gestion des terminologies (les plate-formes d'intégration, les outils de mise en relation...). Le deuxième niveau représente tout ce que nous pouvons attendre comme services du SMTS ; il regroupe tous les services web qui peuvent être développés sur le serveur, les API, ... Le dernier niveau du schéma représente toutes les applications qui peuvent utiliser les services proposés par le SMTS (moteurs de recherche, outils de codage, ...).

Afin de permettre l'intégration de plusieurs terminologies, il nous a fallu dans un

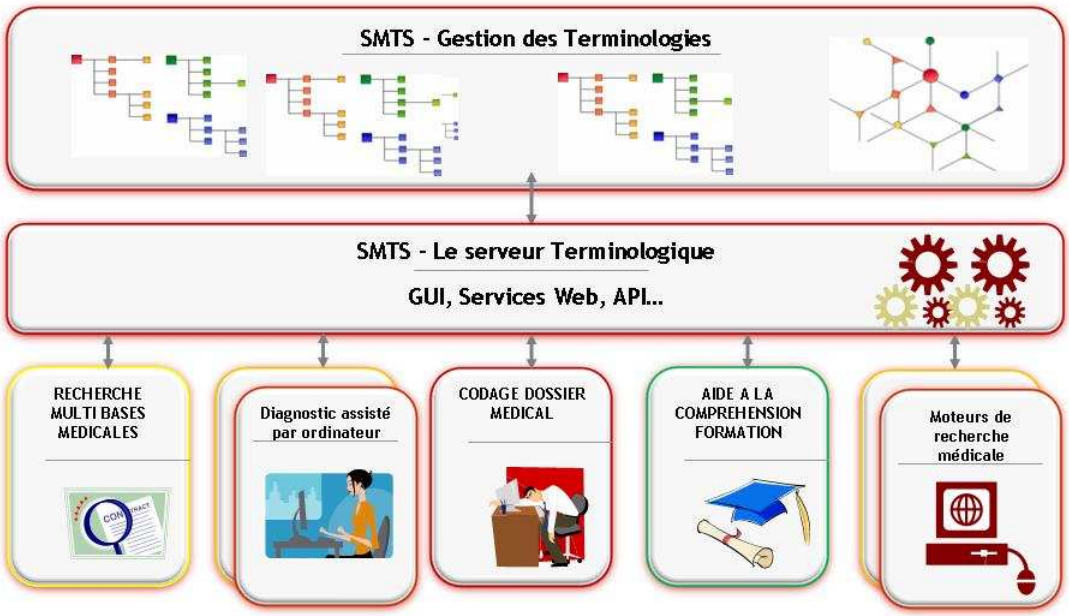

FIG. 3.6 – Architecture trois parties du SMTS

premier temps concevoir un modèle générique pouvant représenter toutes les terminologies et qui soit compatible avec la plate-forme d'intégration d'« ITM® » (Intelligent Topic Manager) de MONDECA. La deuxième étape de l'intégration des terminologies consiste à développer un « analyseur » [6] pour chaque terminologie. Le travail de modélisation a été réalisé en deux étapes :

1. La première consistait à modéliser chaque terminologie d'une manière individuelle.
2. La deuxième étape porte sur l'élaboration du modèle général. Chaque terminologie unitaire modélisée dans l'étape précédente représente une spécialisation du modèle général.

[6] Une fonction capable de transformer le modèle de représentation original de la terminologie vers un autre modèle de représentation.

3.3.2 Modélisation des terminologies médicales

La modélisation est définie comme l'approche permettant de créer une représentation simplifiée d'un problème **modèle**. Pour la modélisation des terminologies médicales, une approche fondée sur la structure des terminologies a été utilisée. Cette méthode a été explorée dans un stage de master recherche dans l'équipe TIBS (Siwar Rekik) Rekik (2007) en 2007 pour modéliser un certain nombre de terminologies. La thèse de S. Pereira Pereira (2007) détaille l'approche et propose dans son manuscrit plusieurs modélisations utilisées pour la plupart pour SMTS. UML (Unified Modeling Language) « langage de modélisation unifié » Booch *et al.* (2000), a été le langage de représentation utilisé pour la modélisation des terminologies. Les modèles ont été tous réalisés par des diagrammes de classes. Les figures 3.7 et 3.8 montrent respectivement les diagrammes de classes des terminologies CIM10 et SNOMED International.

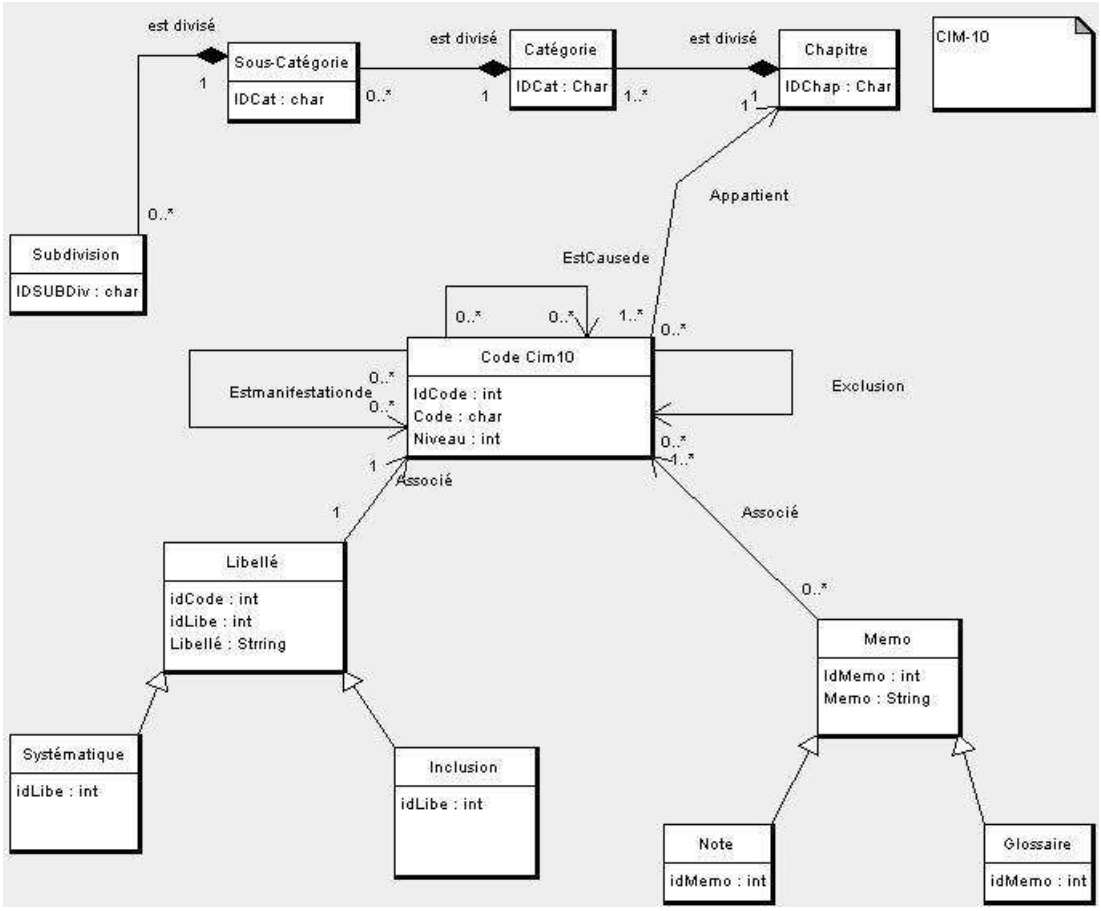

Fig. 3.7 – Modéle UML de la classification CIM10

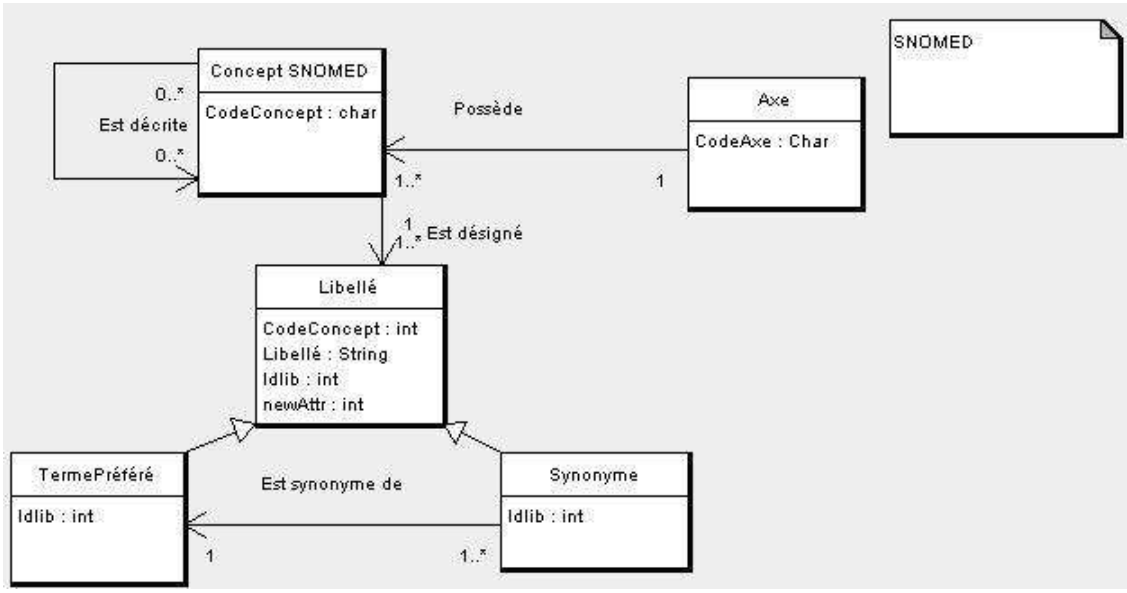

FIG. 3.8 – Modèle UML de la nomenclature SNOMED International

3.3.3 Modèle générique du SMTS

Le modèle général doit être le plus générique possible pour faciliter l'intégration des terminologies (actuelles ou nouvelles) et les stocker dans un format standard : le Web Ontology Language (OWL) Bechhofer *et al.* (2004). Notre modèle[7] définit au moins :

- les « Classes », éléments représentant les « concepts » et dont les instances sont les individus constitutifs de la terminologie. Nous distinguons les classes d'associations (censées héberger les relations entre « concepts (descripteurs) ») des classes de « concepts » elle mêmes.
- les « DatatypeProperty », qui sont les différents attributs de classes d'un type donné. En effet, chaque DatatypeProperty possède un type (numérique, texte, etc.) défini dans l'ontologie « publishing » qu'il est nécessaire de préciser.
- des « ObjectProperty » également attributs de classes mais faisant référence à un objet et non à un type. Il s'agit généralement des rôles dans les relations mais ils peuvent également pointer vers d'autres « concepts » au sein d'un « concept (descripteur) ».

De même, à un niveau d'abstraction supérieur, le méta-modèle a besoin d'être clairement défini par un méta-méta-modèle. Afin d'éviter une décomposition infinie de niveaux d'abstraction, le plus haut niveau s'auto-définit. Conformément au standard MOF (Meta Object Facility)[8] OMG (2002), notre modélisation est organisée en plu-

[7] Ce modèle a été réalisé en collaboration avec P.Y. Vandenbussche doctorant chez MONDECA, et les ingénieurs de CISMeF : B. Dahamna et I. Kergoulay.

[8] MOF est un standard de l'OMG dédié à la représentation des métamodèles.

sieurs niveaux d'abstraction (du plus haut niveau au plus bas):

- le méta-modèle : UMV2 (Unified Metamodel of Vocabularies 2) et les modèles UMV1 (Unified Metamodel of Vocabularies 1) (extension et spécialisation de UMV2 pour chaque terminologie) ;
- les instances : correspondent au contenu d'une terminologie qui se conforme au modèle défini par UMV2 et UMV1.

UMV2 rassemble les éléments communs à toutes les terminologies : classes d'associations communes (comme les relations hiérarchiques), attributs communs (comme l'attribut multi-valué UF), classe mère « Concept » (et ses attributs), ...
UMV1 quant à lui est contenu dans UMV2 et l'étend aux spécificités de chaque terminologie (exemples : Concept MeSH, Notion SNOMED, Exclusion CIM-10, etc.). Il y a donc un UMV1 par terminologie.

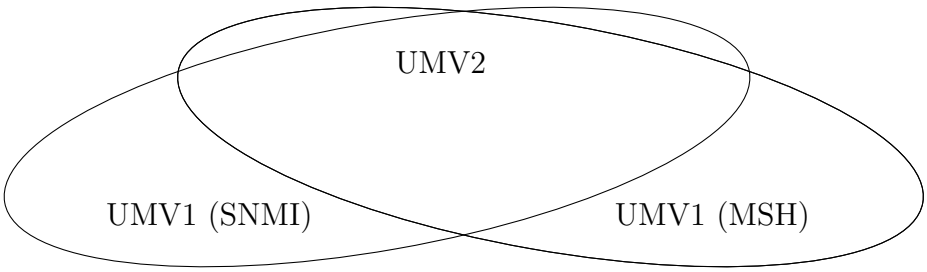

FIG. 3.9 – Relations entre les UMV1 (terminologies) et le méta-modèle UMV2

Le méta-modèle proposé est centré sur la notion de « Concept » (voir figure 3.10), qui définit les attributs communs aux différentes terminologies. À cette classe, seront attachées (au sens « héritage » orienté objet) toutes les classes représentatives des modèles des autres terminologies. Par exemple, si nous reprenons l'exemple de la figure 3.7, la classe « Code CIM10 » sera attachée à la classe « concept » du modèle général comme montré dans la figure 3.11.

Un partie du méta-modèle très importante pour nos travaux de recherche est la partie alignement (voir partie alignement dans la figure 3.10). La capacité de représenter les relations entre les terminologies est une caractéristique très importante d'un modèle multi-terminologique. Dans notre méta-modèle, quatre types de relations inter-terminologiques sont représentées. Ces types de relations sont inspirés des définitions SKOS (Simple Knowledge Organization System) pour les propriétés des matchings W3C (2004). SKOS permet de gérer des informations de mapping en indiquant le degré de recouvrement sémantique entre deux concepts issus de différents thésaurus.

- ExactMatch : elle correspond à la relation « skos:exactMatch » de SKOS, définie pour identifier les termes de différentes terminologies qui sont exactement les mêmes.

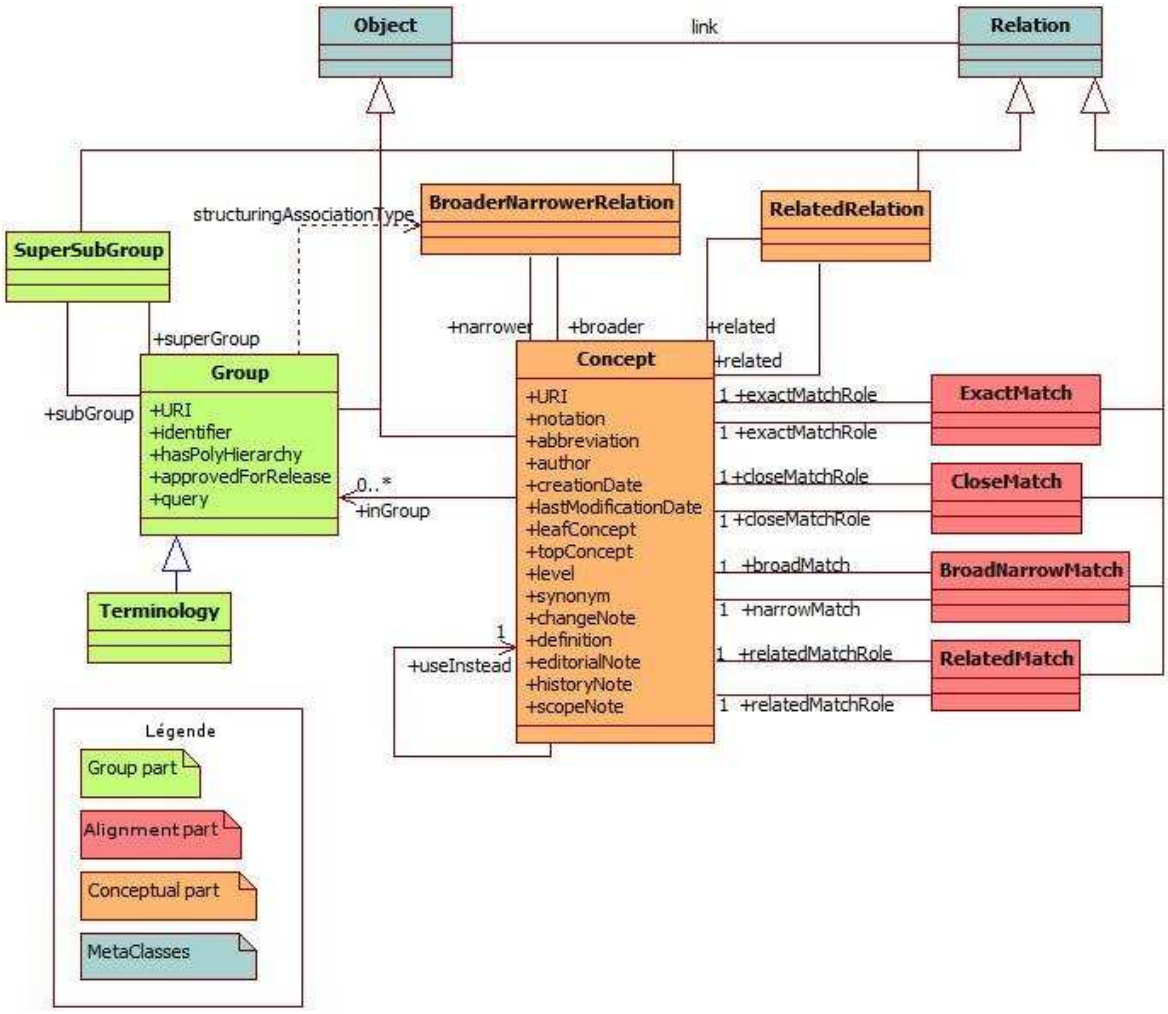

FIG. 3.10 – Modèle UML représentant le méta-modèle UMV2

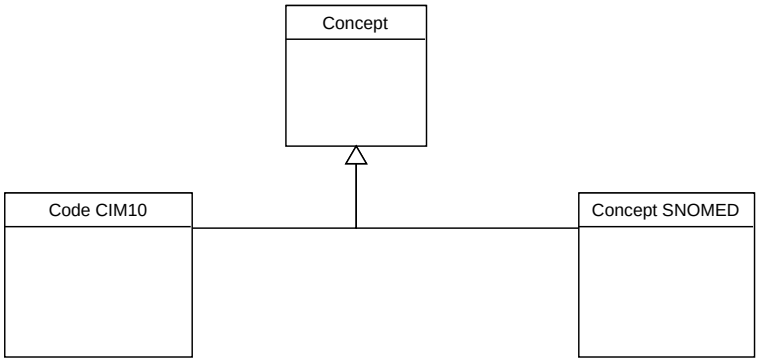

FIG. 3.11 – Héritage de la classe Concept vers les modèles des terminologies

- CloseMatch : elle correspond à la relation « skos:closeMatch », définie pour mettre en correspondance des termes suffisamment similaires pour être utilisés de façon interchangeable. L'équivalence désignée par cette relation est différente de l'équivalence de la relation ExactMatch puisqu'elle n'est pas exacte, contrairement à l'équivalence de la relation ExactMatch qui l'est, mais elle peut être acceptée pour répondre aux besoins d'une application déterminée.
- BroadNarrowMatch : elle correspond aux deux relations « skos:broadMatch » et « skos:narrowMatch », utilisées pour mettre en correspondance des termes de niveaux hiérarchiques différents à travers différentes terminologies. Nous présentons deux types d'alignements qui peuvent être classées dans ce type de correspondance dans la section 4.2.4.
- RelatedMatch : utilisée pour créer des liens associatifs entre des termes de différentes terminologies. Cela veut dire que toute relation définie entre des termes qui n'est pas incluse dans les trois types de relations « ExactMatch », « CloseMatch » et « BroadNarrowerMatch » sera représentée comme une relation inter terminologique de type « RelatedMatch ».

3.3.4 Intégration des terminologies dans le SMTS

L'intégration des terminologies peut être définie comme le processus d' « immersion » de toutes les entités composant ces terminologies « Termes, relations, ... », dans une même plate-forme et suivant un même langage, afin de permettre l'interopérabilité entre elles.
En se basant sur le modèle général décrit dans la section précédente, la plate-forme « ITM » (Intelligent Topic Manager) Amardeilh et Francart (2004) a été choisie comme « plate-forme d'intégration » des terminologies du SMTS. ITM est une plate-forme logicielle pour la gestion de connaissances et l'exploitation d'ontologies. ITM intègre un portail sémantique fournissant quatre fonctions clefs : l'édition, la recherche, la navigation et la publication. « ITM »offre de nombreux mécanismes de gestion et d'exploitation de contenu terminologique et ontologique Amardeilh *et al.* (2005).

Cependant, toutes les terminologies qui seront intégrées dans « ITM » devront être conformes au méta-modèle qu'ITM utilise. En d'autres termes, le modèle général décrit dans la section précédente devra être adapté suivant un langage de représentation de connaissances pour qu'il soit intégré dans le méta-modèle.
Le langage OWL a été choisi comme le langage avec lequel seront représentées les données de chaque terminologie. Nous avons réalisé pour chaque terminologie un « analyseur » pour permettre leurs représentations suivant le langage « OWL ». Dans ce cas

aussi, il s'agit de spécialiser une classe « analyseur » existant en fonction des spécificités de chaque terminologie. Ce qui impose un traitement individuel de chaque terminologie. Cependant, les données ne sont pas dans un format standard (XML typiquement), ce qui constitue une difficulté supplémentaire dans la conception des analyseurs.
Différents types de analyseurs ont été réalisés : des analyseurs de fichiers (héritant de « analyseurFichier ») ou des parseurs de base de données (héritant de « parseurBD »). La figure 3.12 illustre l'organisation générale des différents parseurs.

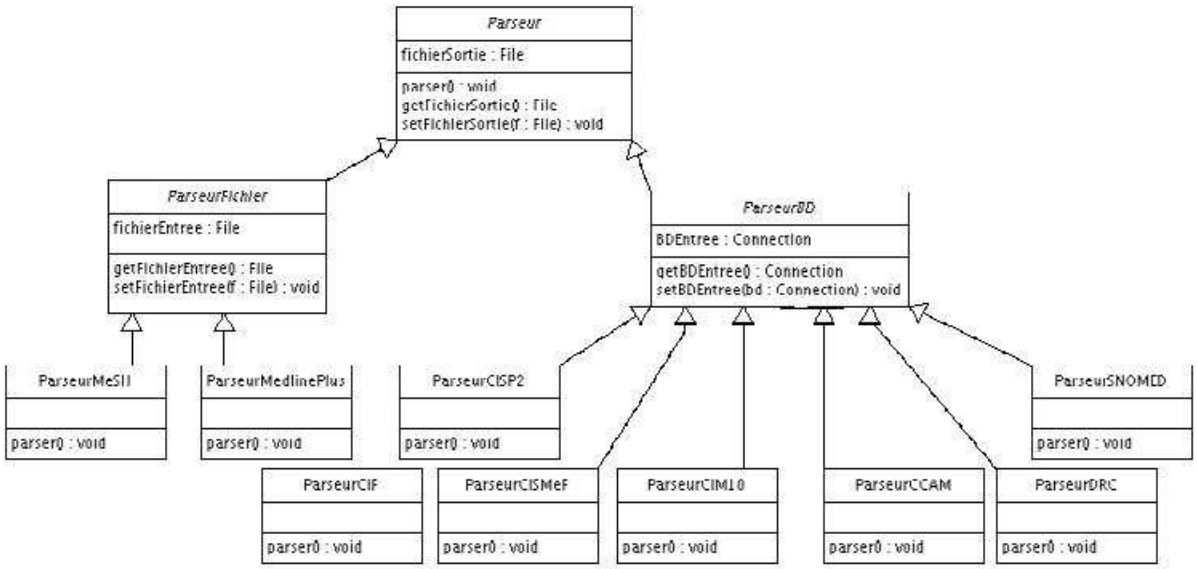

FIG. 3.12 – Organisation générale des parseurs

Actuellement, plusieurs terminologies médicales de santé francophones sont intégrées ou en cours d'intégration au sein du SMTS :

- MeSH (Extension CISMeF), SNOMED, CCAM, CIM10, CISP2 (Classification Internationale Des Soins Primaires version 2) Jamoulle *et al.* (2000), TUV (Thésaurus Unifié de Vidal) Pereira (2007) (InterSTIS) ;
- DRC (Dictionnaire des Résultats de Consultation) publié par la SFMG (Société Française de Médecine Générale) Ferru et Kandel (2003), MEDLINEPlus thésaurus patient développé par la NLM en deux langues anglais et espagnol Miller *et al.* (2000), MedDRA, la classification ATC (Anatomique, Thérapeutique et Chimique) Skrbo *et al.* (2004), CIF (Classification Internationale du Fonctionnement, du handicap et de la santé - OMS) OMS (2001), WHO-ART (ASIP Santé) ;
- LOINC (Logical Observation Identifier Names and Code) McDonald *et al.* (2003), ORPHANET (thésaurus des maladies rares) (voir section 4.2.2).

3.4 Interopérabilité Sémantique Inter et Intra Terminologique

Définition de l'interopérabilité

L'interopérabilité est définie comme la capacité à échanger de l'information et l'utiliser entre différentes sources de données distribuées (hétérogènes) Wegner (1996). L'interopérabilité peut être traitée suivant deux niveaux : technique (syntaxique) et sémantique. L'interopérabilité syntaxique permet essentiellement, aux sources de données distribuées, d'échanger de l'information en prenant en compte l'hétérogénéité syntaxique et structurelle existante entre ces différentes sources. Le fait qu'une ressource soit exprimée sous un format standardisé permet une interopérabilité syntaxique. Cependant, même avec ce type d'interopérabilité plusieurs « mauvaises interprétations » des données peuvent se faire, impliquant des « malentendus » entre utilisateurs, des « erreurs de calcul » ou bien même des « défaillances » au niveau des systèmes. À un niveau supérieur, l'objectif de l'interopérabilité sémantique est d'éviter ces problèmes et d'assurer que les échanges qui s'effectuent entre les sources de données conservent leur sens en prenant en compte la sémantique associée à chaque donnée Dougoulet *et al.* (1997) (figure 3.13). Pour résoudre cette incompatibilité entre les différentes terminologies, la recherche s'est concentrée initialement sur la possibilité d'unifier les terminologies. L'UMLS décrit dans les sections précédentes, regroupe actuellement plus de 140 terminologies biomédicales dans un seul métathésaurus. Actuellement, l'UMLS est probablement le plus grand projet de regroupement entre terminologies jamais réalisé.

3.5 Méthodes pour la mise en relations entre terminologies

3.5.1 Terminologies

Récemment, plusieurs travaux ont été menés par diverses équipes de recherches pour la création d'outils et/ou de systèmes permettant la transition (automatique, semi-automatique, manuelle) d'une terminologie à une autre Wang *et al.* (2008); Rocha

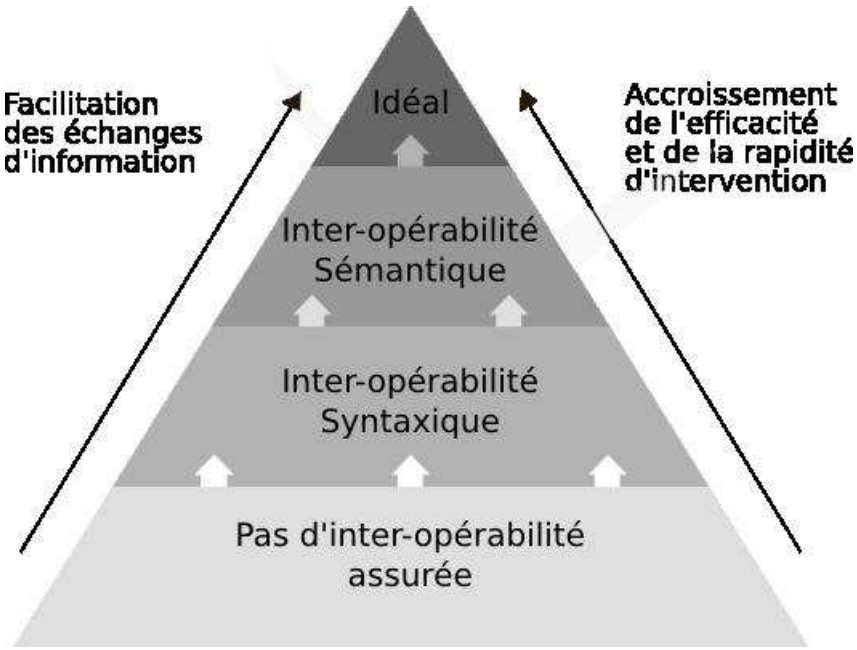

FIG. 3.13 – Pyramide d'interopérabilité

et al. (1994); Cimino et Barnett (1990); Fung et Bodenreider (2005); Bodenreider *et al.* (1998).
Plusieurs termes sont utilisés par différents auteurs pour définir le mécanisme de mise en correspondance entre les termes de différentes terminologies. Euzenat et Shvaiko (2007), ont proposé un certain nombre de définitions pour décrire ces processus entre les ontologies.

Dans le reste de ce manuscrit, nous utiliserons le terme « alignement » pour décrire la méthode permettant de déterminer les correspondances entre les terminologies. Formellement, un alignement est définit comme une fonction qui prend en entrée deux terminologies/ontologies T1 et T2 avec un ensemble d'entités (e.g., tables, élements XML...), et qui retourne un ensemble de correspondances T'. L'ensemble des éléments de T1 et T2 (e.g., équivalences, subsumption) (Figure 3.14).

Cependant, il s'avère très difficile d'automatiser l'alignement entre les terminologies. Cela est du principalement à l'hétérogénéité des terminologies médicales. En effet, la structure et le contenu de chaque terminologie sont créés en fonction de l'utilisation qui doit en être faite. Elles sont généralement créées pour des tâches bien précises. Plusieurs travaux ont été menés pour permettre l'interopérabilité entre terminologies en utilisant des méthodes algorithmiques d'alignement. En général, ces méthodes peuvent

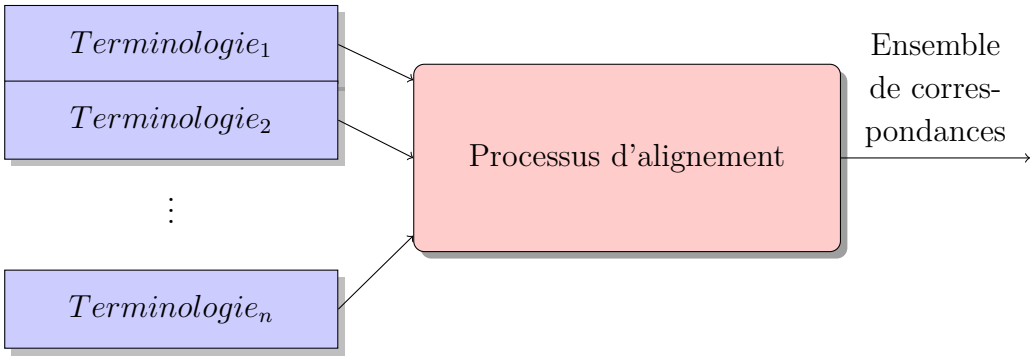

FIG. 3.14 – Le processus d'alignement

être classées en des méthodes lexicales ou des méthodes sémantiques.

3.5.2 Méthodes lexicales

Ce sont les méthodes qui s'appuient sur des propriétés lexicales des termes de chaque terminologie. Des outils de normalisation sont utilisés dans ces méthodes pour réduire les termes sous un format commun de représentation. Les méthodes lexicales représentent la façon la plus triviale d'identifier les correspondances entre termes. L'utilisation de telles méthodes dans le domaine de la médecine pour réaliser mettre en relation les terminologies est motivée par le fait que la plupart des terminologies partagent un grand nombre de termes similaires. Le développement d'outils pour le traitement automatique de la langue en médecine a fortement contribué à l'amélioration de ces méthodes.

Les méthodes fondées sur les chaînes de caractères

Dans ces méthodes, les termes ou les libellés sont considérés comme des séquences de caractères dans un alphabet donné. Elles définissent des distances entre chaînes de caractères pour en déduire une similarité. En fonction du modèle de similarité utilisé, une chaîne de caractères A est modélisée de différentes façons :

- une séquence de caractères notée $A = (\mathrm{a}_1\,;\,\mathrm{a}_2\,;\ldots;\,\mathrm{a}_n)$ où les a_i sont des lettres (ou symboles) ;
- une séquence de sous-chaînes de caractères (mots) séparées par des délimiteurs (espaces, tirets, ponctuations,...). Dans ce cas, une chaîne de caractères sera représentée par une séquence de mots notée $\mathrm{A} = (\mathrm{A}_1\,;\,\mathrm{A}_2\,;\ldots;\mathrm{A}_m)$ où les A_i sont des chaînes de caractères.

Ces distances peuvent ignorer l'ordre d'apparition des caractères dans la séquence. Dans ce cas, la comparaison consiste à utiliser une mesure de similarité ensembliste. Parmi lesquelles nous pouvons citer :

Distance de Hamming Hamming (1950) : définie pour les chaînes de caractères de longueur égales. Pour deux mots A et B, la distance de Hamming $d_{hamming}(A, B)$ représente le nombre de positions en lesquelles les deux mots possèdent des lettres différentes.

Distance de Jaccard Jaccard (1901) : définie par le nombre des objets en commun (les caractères) divisé par le nombre total des objets :

$$d_{Jaccard}(A, B) = 1 - \frac{|A \cap B|}{|A \cup B|}$$

Distance de Dice Salton et McGill (1983); van Rijsbergen (1979) : une variante de la distance de Jaccard définie par la formule suivante :

$$d_{Dice}(A, B) = 1 - \frac{2\,|A \cap B|}{|A| + |B|}$$

D'autre part, une famille de mesures adaptées existe appelée *distance d'édition*, qui prennent en compte l'ordre d'apparition des caractères :

Distance d'édition : elle est définie par le nombre d'opérations qui permettent de transformer une chaîne de caractères en une autre. Dans ce type de distance, les opérations *Oper* utilisées appelées ***opérations d'édition*** sont :

1. $Ajout(a)$, l'insertion d'une lettre a ;
2. $Subst(a, b)$, la substitution de la lettre a par la lettre b ;
3. $Supp(a)$, la suppression d'une lettre a.

Un coût de valeur entière positive est associé à chacune des opérations, noté
$coût : Oper \rightarrow \Re$.
Soit $E_{A \rightarrow B}$, l'ensemble des séquences d'opérations $SOp_x = (op_1 ; \ldots ; op_n)$ (avec $op_i \in Oper$) permettant de passer d'une chaîne A à une chaîne B. La distance d'édition δ entre les chaînes A et B est définie par :

$$\delta(A, B) = min_{SOp_x \in E_{A \rightarrow B}} \sum_{op_i \in SOp_x} coût(op_i)$$

La distance de Levenshtein représente le nombre minimum d'insertions, de suppressions et de substitutions de caractères nécessaire pour transformer une chaîne de caractère en une autre Levenshtein (1966). C'est une distance d'édition avec un coût des opérations égale à 1. Un autre exemple de distance est SMOA Stoilos *et al.* (2005). Cette distance

est dépendante de la longueur des sous-mots en communs et non commun. La valeur de la similarité calculée par cette distance est comprise entre -1 et 1.

D'autres variantes des distances présentées ci-dessus ont été proposés, en considérant les chaînes de caractères comme des mots. Ces mesures sont souvent appelées distances n-gramme Kondrak (2005). Ces distances calculent le nombre en commun des n-grams (i.e., les séquences de n caractères) entre chaînes de caractères pour en déduire une similarité entre eux. Le modèle vectoriel est un moyen statistique de comparaison ensembliste entre mots Salton et McGill (1983). Il est fondé sur un calcul du poids des mots utilisant la mesure de TF-IDF (Term Frequency - Inverse Document Frequency). Ce modèle définit l'importance d'un mot suivant sa fréquence. La similarité entre deux séquences est calculée comme un cosinus entre deux vecteurs représentatifs des séquences Salton et Buckley (1988).

Les méthodes fondées sur les chaînes de caractères sont plus adaptées pour comparer les termes simples (libellés, identifiants). Plusieurs outils ont été développée autour de ces méthodes pour permettre l'alignement entre les ontologies dans plusieurs domaine Euzenat et Shvaiko (2007) (voir table 3.6). Dans le domaine médical, l'outil OnAGUI[9] Mazuel et Charlet (2009) est un exemple d'outil permettant l'alignement entre ontologies médicales (voir figure 3.15). Cet outil se base essentiellement sur deux distances : Levenshtein Levenshtein (1966) et SMOA Stoilos *et al.* (2005) pour mesurer la similarité entre concepts.

Simetrics	AlignAPI	SimPack
Levenshtein	Levenshtein	Levenshtein
Jaccard		Jaccard
Dice		Dice
	TF-IDF	TF-IDF
	SMOA	

Tab. 3.6 – Quelques outils d'alignement utilisant des mesures de similarité

L'utilisation de mesures fondées sur le modèle vectoriel est pertinente pour une comparaison entre des données textuelles. Cette mesure a été utilisé dans Merabti *et al.* (2008) (voir section 2.3.2) pour quantifier les ressemblances entre les titres et résumés des ressources CISMeF. L'algorithme « Related Articles » de PubMed Kim *et al.* (2001) est fondée sur ce modèle vectoriel pour mesurer la similarité entre les articles indexés dans MEDLINE.
Cependant, l'utilisation de ce type de mesures permet seulement de quantifier la ressemblance entre libellés ou concepts. Ainsi, ces méthodes donnent des similarités faibles

[9] http://sourceforge.net/projects/onagui/

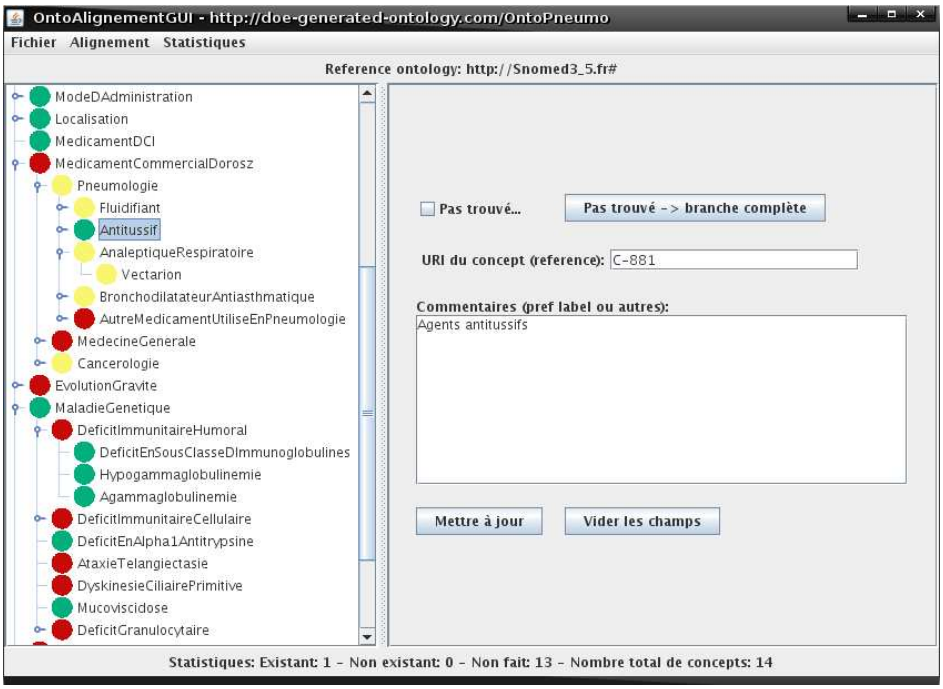

FIG. 3.15 – Aperçu de l'interface OnAGUI

pour des synonymes avec des formes différentes. Par exemple, les deux termes « douleur » et « algie » sont des termes synonymes mais les différentes mesures décrites dans cette section ne peuvent par détecter les liens entre eux. À l'inverse, ces méthodes trouveront des similarités importantes entre des termes qui ne sont pas les mêmes (faux positifs). Par exemple, les deux termes « Vitamine A » et « Vitamine U ».

Les méthodes fondées sur le langage

Ces méthodes considèrent les termes comme étant des mots dans un langage naturel. Elles se basent principalement sur les outils de Traitement Automatique de la Langue (TAL). Ces outils exploitent les propriétés morphologiques des mots traités. Nous distinguons deux classes de méthodes, celles qui s'appuient sur des algorithmes et celles qui utilisent en plus des ressources externes comme les dictionnaires.

Les méthodes intrinsèques

Ces méthodes se base sur des outils de traitement automatique de la langue pour normaliser les termes sous des formes standard qui peuvent être facilement reconnais-

sables. Trois types de variation peuvent être distinguées sur les termes Maynard et Ananiadou (2001) :

Morphologique Elle concerne la manière dont les termes sont constitués à partir d'unités minimales signifiantes. Les variations morphologiques prennent trois types de formes : flexionnelles, dérivationnelles et compositionnelles. Le tableau 3.7 montre un exemple pour chacune de ces variations sur le mot « membrane ».

Syntaxique Elle décrit la manière dont les mots se combinent en phrase syntaxiquement correcte.

Sémantique Elle concerne les sens des mots et la manière dont ils se combinent.

Flexion	Carcinomes (+s)
Dérivation	Carcinomateux (+ateux)
Composition	hépatocarcinome (+hépat)

TAB. 3.7 – Exemples de variation morphologiques sur le mot « membrane »

Tokenisation Elle consiste à segmenter une séquence (mot) en unités atomiques appelées « tokens ». Cette segmentation consiste à éliminer les ponctuations, les caractères blancs...

Déssufixation Les chaînes de caractères représentées par des « tokens » sont analysées pour les réduire sous une forme de base normalisée. L'analyse morphologique permet de retrouver toutes les flexions et les dérivations à partir de la racine du mot. La déssufixation est l'analyse qui cherche à rassembler les différentes variantes d'un mot autour d'un stème (la forme canonique).

Élimination des mots vides Les mots vides sont des mots non significatifs. Ces mots sont généralement générateurs de bruit pour un but donné. Il est donc fortement recommandé de les éliminer.

Plusieurs outils fondés sur des outils TAL ont été utilisés pour mettre en correspondance les terminologies médicales. Nous citons à titre d'exemple, le travail élaboré dans Wang *et al.* (2008) où les auteurs utilisent les techniques de tokenisation et de déssufixation pour aligner CISP-2 et SNOMED CT. C'est le cas aussi des principales techniques lexicales proposées par la NLM dans l'API UMLSKS[10]. L'outil MetaMAP Aronson (2001) est un outil qui fait partie des outils lexicaux de la NLM. Il permet de détecter les termes médicaux à partir du texte (documents, phrases et termes) en anglais et de déterminer les concepts du Métathésaurus de l'UMLS correspondants. Le texte passe par une série de traitements. Il subit une analyse syntaxique, le décomposant en phrases, expressions et mots clés. À partir des différentes variantes générées, des concepts UMLS candidats sont proposés. Cependant, comme décrit dans la figure 3.16 qui résume les

[10] http://umlsks.nlm.nih.gov/

différentes étapes utilisées, une recherche lexicale des différents mots qui compose le terme en entrée est faite dans un « Lexique Spécialiste »(Specialist Lexicon McCray *et al.* (1994)). Ainsi, cet outil peut être classé dans les méthodes extrinsèques qui seront définies dans la section suivante. Dans Johnson *et al.* (2006), les auteurs utilisent la bibliothèque Lucene[11] pour la recherche d'information Cutting *et al.* (2004) pour repérer des relations entre le Gene Ontology [12] et trois autres ontologies biomédicales. Ces outils ont été utilisés aussi dans plusieurs travaux dans CISMeF : recherche d'informations Soualmia (2004), l'indexation automatique Névéol (2005) et l'indexation multi-terminologique Pereira (2007). L'algorithme de sac de mots décrit dans Pereira (2007) utilise plusieurs méthodes TAL pour permettre l'appariement des termes issus d'une ou plusieurs terminologies à une phrase. Certaines techniques utilisées dans cet algorithme seront détaillées dans la section 4.2.4 lorsque nous introduisons notre méthode d'alignement entre terminologies.
L'avantage des méthodes fondées sur les outils TAL réside principalement dans leurs simplicité d'implémentation. En effet, l'utilisation des techniques telles que la désuffixation permet de limiter le nombre de ressources utilisées.

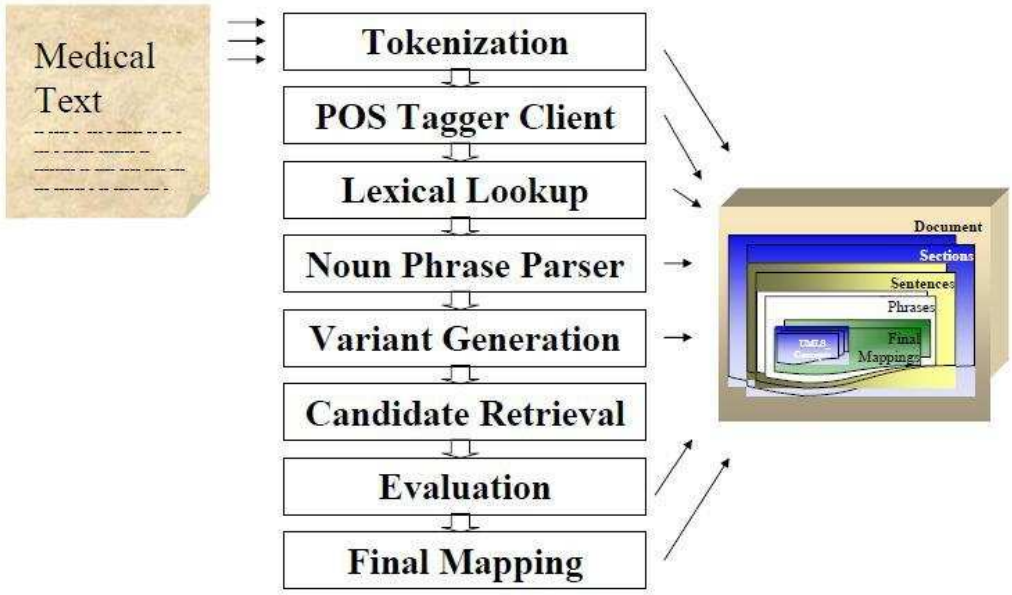

FIG. 3.16 – Étapes suivies par MetaMap

[11] http://lucene.apache.org/
[12] Gene Ontology : http://www.geneontology.org

Les méthodes extrinsèques

Ces méthodes utilisent des ressources externes, comme des dictionnaires ou bien des lexiques. Plusieurs sources linguistiques peuvent être utilisées pour trouver des correspondances éventuelles entre les termes, par exemple : les dictionnaires, les lexiques, les thésaurus...
Elles sont utilisées pour trouver des correspondances fondées sur des relations de synonymie, hyponemie... Ces méthodes constituent les bases des outils lexicaux proposées dans l'API UMLSKS. Elles sont combinées avec les synonymes des termes et des concepts dans différentes ressources externes pour optimiser les alignements vers les concepts UMLS. Fung et Bodenreider (2005) utilisent l'UMLS comme ressource externe pour produire un alignement inter-terminologie. Une autre ressource externe largement utilisées dans d'autres domaines que le domaine biomédical est WordNet[13] Fellbaum (1998). WordNet est une base de données lexicale électronique développé depuis 1985 à l'université de Princeton par une équipe de psycholinguistes et de linguistes du laboratoire des sciences cognitives, sous la direction de Georges A. Miller. L'avantage de WordNet réside dans la diversité des informations qu'elle contient (grande couverture de la langue anglaise, définition de chacun des sens, ensembles de synonymes, diverses relations sémantiques). Les éléments de base de WordNet sont des ensembles de termes synonymes appelés « synsets ». Chaque « synset » est associé à une définition et à un ensemble de synsets avec lesquels il est en relation.

En outre, WordNet est librement et gratuitement utilisable[14]. Dans certains travaux de recherches Leroy et Chen (2001), WordNet a été utilisée en combinaison avec UMLS pour produire des alignements inter-terminologies.
Dans nos algorithmes d'alignements, en plus des méthodes lexicales fondées sur les outils TAL, nous utilisons les relations de synonymie provenant des différentes terminologies pour optimiser nos alignements. Nous utilisons UMLS comme ressource externe pour tirer avantage des alignements conceptuels existants entre termes de différentes terminologies (voir section 4.2.4).
L'avantage de ces méthodes est que nous avons une plus grande couverture syntaxique et sémantique. En effet, l'application de ressources externes permettra d'améliorer considérablement les alignements entre terminologies. Dans le reste de nos travaux nous présentons l'avantage de l'utilisation de l'UMLS comme ressource externe pour trouver plus d'alignements entre différents terminologies biomédicales. Cependant, l'utilisation de ces ressources pourra affecté la qualité des alignements. En effet, des termes synonymes dans une terminologies n'implique pas que ces termes le soient dans une autre.

[13] http://wordnet.princeton.edu/

[14] À noter que Dominique Dutoit (HDR, membre associé du LITIS) a fortement developpé WordNet en français. Il a également collaboré avec CISMeF dans les projets VODEL et InterSTIS.

3.5.3 Méthodes structurelles (sémantiques)

Ces méthodes utilisent les propriétés structurelles de chaque terminologie pour établir des correspondances vers des termes ou des concepts d'autres terminologies. Les techniques utilisées considèrent les terminologies (thésaurus, classifications...) comme des graphes où les nœuds représentent les termes de la terminologie et les arêtes représentent les relations entre les termes dans la terminologie. La plupart des terminologies médicales peuvent être représentées avec des graphes. Ces techniques sont généralement utilisées en combinaisons avec des méthodes lexicales. Dans ce cas de figure, le travail de Bodenreider *et al.* (1998) est un bon exemple illustrant l'utilisation des relations entre terminologies pour aligner les termes qui n'ont pas été alignés par les techniques lexicales. L'algorithme de Bodenreider *et al.* (1998) utilise les relations sémantiques d'UMLS pour trouver les correspondances inter-terminologies. En effet, pour aligner un terme dont les outils lexicaux n'ont pas trouvé de correspondant dans le MeSH, l'algorithme commence par la construction d'un graphe avec comme terme source le terme non aligné et dont les nœuds représentent les parents du terme source (relation hiérarchique de l'UMLS). À partir de ce graphe, le terme le plus proche dans la hiérarchie qui a une correspondance vers le MeSH est sélectionné. La figure 3.17 reprend l'exemple présenté dans Bodenreider *et al.* (1998) permettant d'aligner le terme « veine du cou » vers MeSH. Dans cet exemple, les termes MeSH sélectionnés sont encadrés en double : « Cou » et « Veine ».

Dans Bodenreider *et al.* (1998), le terme le plus proche est calculé par rapport aux nombres d'arêtes séparant le terme source des autres termes dans la terminologie. Cependant, plusieurs mesures de similarité sur les structures hiérarchiques ont été proposées. La plus commune consiste à calculer le nombre d'arêtes entre les termes pour déterminer la distance entre eux. La distance la plus connue est la similarité de Wu-Palmer Wu et Palmer (1994). Cette similarité est définie par rapport à la distance qui sépare deux termes dans la hiérarchie et également par leur position par rapport à la racine. Pour deux termes T_1 et T_2 la similarité est :

$$SIM(T_1, T_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3}$$

où N_3 est le nombre d'arêtes qui séparent le plus petit parent commun de la racine, N_1 et N_2 représentent le nombre d'arêtes qui séparent les termes T_1 et T_2 de la racine respectivement. Cette mesure a l'avantage d'être simple à implémenter.

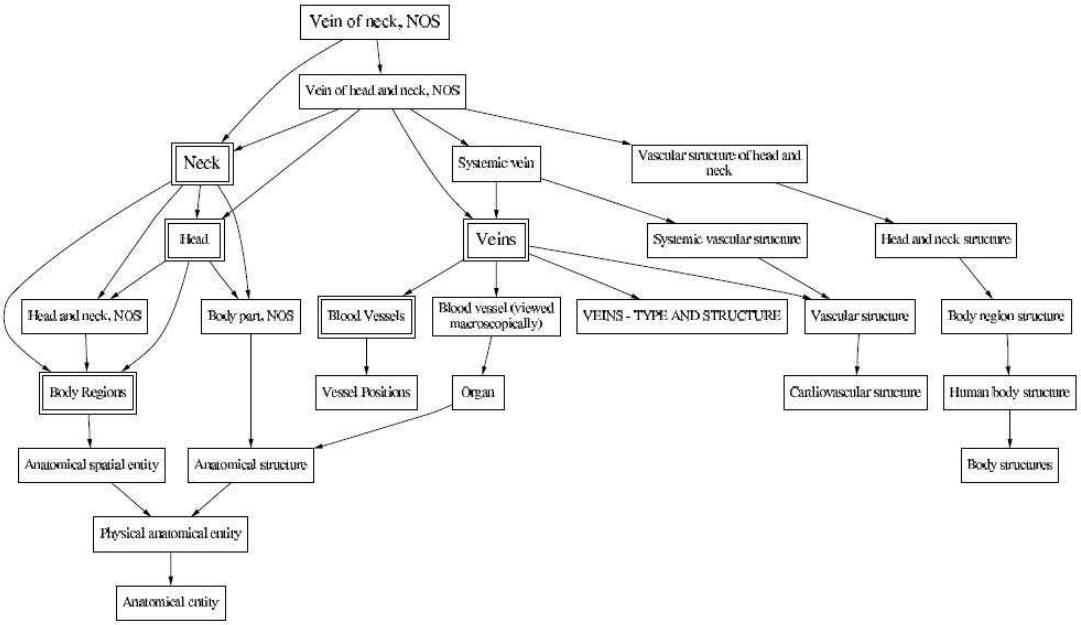

FIG. 3.17 – Graphe représentant les parents du terme « veine du cou » dans UMLS

La notion de Contenu Informationnel (CI) a été introduite la première fois par Resnik (1995). Elle combine positions et corpus. Le contenu informationnel d'un terme (concept) traduit la pertinence d'un concept dans le corpus en tenant compte de sa spécificité ou généralité. La fréquence du terme dans le corpus est généralement utilisée pour calculer le contenu informationnel. Elle est calculée en fonction de la fréquence d'apparition du terme ainsi que les parents de ce terme.

$$CI(terme) = -log(P(terme))$$

où $P(terme)$ est la probabilité de retrouver une instance du terme. Ces probabilités sont calculées par : $fréquence(terme)/N$ où N est le nombre total des termes. Resnik (1995) définit une mesure de similarité entre deux concepts par la quantité d'information qu'ils partagent. Cette similarité est calculée par :

$$SIM(T_1, T_2) = CI(pcs(T_1, T_2))$$

où $pcs(T_1,T_2)$ est le parent commun le plus spécifique de T_1 et T_2 dans la hiérarchie. Lin (1998) propose une définition théorique de la similarité applicable à partir du moment où l'on dispose d'un modèle de probabilité. Cette similarité est définie comme le rapport des informations partagées par T_1 et T_2 sur les informations nécessaires pour décrire d'une façon complète T_1 et T_2. Cette mesure est définie par :

$$SIM(T_1, T_2) = \frac{2 \times max_{commun(T_1,T_2)}[log(p(commun(T_1, T_2)))]}{log(p(T_1)) + p(T_2))}$$

où *commun*(T_1,T_2) est l'ensemble des termes parents partagés par T_1 et T_2 et $p(T)$ représente la probabilité de trouver T ou un de ses fils dans le corpus. Elle génère une similarité normalisée entre 0 et 1. Cette mesure de similarité a été démontrée comme la plus adaptée quand elle est appliqué à Gene Ontology Lord *et al.* (2003). Cette mesure de similarité a été utilisée dans Névéol *et al.* (2006) pour mesurer la similarité entre les termes de la terminologie MeSH.
Cette dernière mesure Lin (1998) a été aussi utilisée dans « CISMeF related resources » Merabti *et al.* (2008) (voir section 2.3.2) combinée à une distance syntaxique (vectorielle) pour calculer la similarité entre les ressources CISMeF.

3.6 Synthèse

Nous avons exposé dans ce chapitre les principales terminologies médicales utilisées dans le cadre de cette thèse. Nous avons vu que ces terminologies ont des visées et des formats de représentation très différents. La nomenclature SNOMED 3.5 pour le codage d'informations cliniques, les classifications CIM10 et CCAM pour le codage épidémiologique puis médico-économique, le thésaurus MeSH pour la bibliographie... Nous avons aussi présenté l'UMLS (Unified Medical Language System) et plus précisément le Métathésaurus la partie de l'UMLS utilisée dans cette thèse. Puis nous avons abordé la problématique liée à l'intégration des différentes terminologies au sein d'un même serveur, une solution a été détaillée celle du SMTS (Serveur Multi-Terminologique de Santé).
Dans la deuxième partie de ce chapitre, après avoir définie la notion d'interopérabilité et d'alignement entre terminologies, nous avons exposé différentes méthodes existantes pour mettre en correspondance différents concepts (termes) de différentes terminologies. Nous avons distingué deux catégories de méthodes :

- les méthodes lexicales : elles utilisent les propriétés lexicales des termes pour définir une distance syntaxique entre eux. Elles représentent la façon la plus triviale d'identifier des correspondances entre termes ;
- les méthodes structurelles : elles sont fondées sur les structures hiérarchiques des terminologies pour identifier les correspondances. Ces méthodes sont souvent combinées avec les méthodes lexicales pour une large couverture.

Pour les deux catégories, nous avons aussi montré que ces méthodes peuvent être combinées avec plusieurs ressources linguistiques et terminologiques. Nous avons aussi exposé les différents travaux réalisés et outils développés dans le domaine médicale.
Les méthodes et les outils développés dans le cadre de cette thèse appartiennent pour la plupart à la catégorie des méthodes lexicales. Cependant, nous présentons dans un

de nos travaux une approche structurelle combinée avec une méthode lexicale. Nous utilisons aussi dans le cadre de cette thèse, le principal outil lexical « MetaMap » de la NLM pour aligner les termes en anglais des différentes terminologies utilisées. À chaque fois, nous comparons les résultats de nos méthodes en français avec celles obtenues par MetaMap. Toutes ces méthodes seront décrites dans le chapitre suivant.

Chapitre 4

Alignement des terminologies francophones avec UMLS (F_UMLS)

Ce chapitre est consacré aux différentes méthodes utilisées et implémentées dans le cadre de cette thèse, nous détaillerons notre algorithme d'alignement lexical lorsque nous aborderons l'alignement du thésaurus ORPHANET avec F_UMLS (les terminologies francophones de l'UMLS). Nous introduisons dans cette partie une approche mixte fondée sur les outils TAL et les relations hiérarchiques pour aligner les termes ORPHANET avec F_UMLS. L'utilisation de l'UMLS a deux avantages, d'une part, d'avoir une large couverture sur toutes les autres terminologies non francophones comme la SNOMED CT, d'autre part, l'utilisation de l'alignement conceptuel (voir section 4.2.4) de l'UMLS permettra de trouver plus d'alignements non repérés par nos méthodes lexicales.
La deuxième partie de ce chapitre est consacrée à l'alignement de la classification ATC vers UMLS. Dans cette partie, en plus de nos méthodes et outils, nous utilisons l'outil MetaMap pour aligner les termes en anglais de l'ATC vers UMLS puis comparer les résultats des deux méthodes. Nous terminerons ce chapitre en proposant une méthodologie permettant d'aligner la classification CCAM vers UMLS. La méthode proposée dans cette partie est assez différente des autres méthodes car nous nous basons sur la structure des codes de la CCAM pour appliquer notre alignement.

4.1 Positionnement de nos méthodes d'alignement

Dans le cadre de cette thèse, nous utilisons principalement trois types d'alignement parmis ceux définies dans la section 3.5. Le premier type d'alignement est fondé sur le langage. Plusieurs outils de traitement automatique de la langue (TAL) sont utilisés dans ce premier type d'alignement. Dans la section 4.2.4, nous présentons ces principaux outils TAL tels que : la désuffixation et l'élimination des mots vides. En plus de ce type d'alignement, nous utilisons aussi le métathésaurus de l'UMLS comme ressource externe pour trouver plus de correspondances entre les différentes terminologies étudiées.
Enfin, nous avons utilisé des alignements structurels fondés sur les structures hiérarchiques de quelques terminologies. Cependant, dans tous nos alignements aucune distance sémantique définie dans la section 3.5 est utilisée. L'utilisation de ces distances implique une complexité supplémentaire due à la taille des terminologies et au nombre de relations multiples existantes. Nous utiliserons aussi dans le cadre de deux études, l'outil MetaMap Aronson (2001) sur les libellés en anglais, afin d'y comparer notre méthode fondée sur le français, pour aligner avec UMLS.

4.2 Alignement du thésaurus Orphanet avec F_UMLS

4.2.1 Contexte de travail

Ce travail entre dans le cadre d'un projet européen DG SANCO en partenariat[1] avec ORPHANET France, financé par l'agence exécutive pour la santé et les consommateurs (AESC) et en collaboration avec l'équipe Bio-Health Informatics Group du Professeur Alan Rector de l'université de Manchester[2]. Parmi ses objectifs figurent :

– la « mise en correspondance » de la version 10 de la Classification Internationale des Maladies (CIM10) OMS (1993) avec toutes les maladies rares ;

[1] À noter qu'à la marge du projet, une coopération est mise en place entre Orphanet et CISMeF.

– Orphanet met à disposition son thésaurus Orphanet pour permettre une indexation et une recherche d'information avec ce thésaurus ;

– En contrepartie, CISMeF assiste Orphanet dans le développement de son parseur pour intégrer la terminologie Orphanet dans le PTS.

CISMeF met également à disposition d'Orphanet « CISMeF InfoRoute » (outil de connaissance contextuelle) pour accéder à PubMed à partir d'Orphanet.

[2] http://intranet.cs.man.ac.uk/bhig/

– proposer des changements pour améliorer la classification CIM10 dans l'optique d'adopter la version 11 de la CIM ;

– la « mise en correspondance » d'ORPHANET avec d'autres classifications et terminologies comme MedDRA, MeSH, SNOMED CT...

4.2.2 Le Portail ORPHANET

ORPHANET est un portail européen d'information sur les maladies rares et les médicaments orphelins, accessible pour tous publics[3] et disponible en six langues européennes (anglais, français, espagnol, allemand, italien et portugais). L'objectif principal d'ORPHANET est d'optimiser l'utilisation des informations disponibles pour améliorer le diagnostic, le traitement et la prise en charge des malades et faire progresser la recherche dans le domaine des maladies rares et des médicaments orphelins. Sa mission est reconnue comme un axe prioritaire du plan national des maladies rares. Parmi les services proposés par ORPHANET :

– Une base de données sur plus de 5 000 maladies rares. Il propose une encyclopédie qui contient des informations détaillées sous forme de résumés, d'articles de synthèse ou de fiches à destination du grand public. La plus complète actuellement dépassant OMIM (Online Mendelian Inheritance in Man) McKusick (2004).

– Un annuaire de service qui contient des informations sur les consultations spécialisées, les laboratoires de diagnostics, les projets de recherche en cours, les registres, les essais cliniques et les associations des malades en liens avec les maladies rares dans 20 pays européens.

– Un service d'aide au diagnostic, permettant la recherche par signes cliniques.

– Une base de données de médicaments pour les maladies rares, un nombre de 342 médicaments sont ainsi répertoriés. Certains d'entre eux ont le statut de médicament orphelin[4].

– OrphanetXchange, un service de mise en relation des chercheurs et des industriels pour les aider à développer des solutions diagnostiques et thérapeutiques dans le domaine des maladies rares.

ORPHANET est devenu le site de référence mondial pour la documentation et l'information sur les maladies rares et les médicaments orphelins. Le site reçoit près de 20 000 utilisateurs chaque jour, provenant de plus de 150 pays.

[3] http://www.orphanet.net

[4] Les médicaments orphelins sont destinés au traitement de maladies qui sont si rares que les promoteurs sont peu disposées à les développer dans les conditions de commercialisation habituelles.

4.2.3 Le thésaurus ORPHANET

ORPHANET a développé un thésaurus multi-hiérarchique pour les maladies rares. Le thésaurus ORPHANET compte 7 428 termes préférés. Chaque terme appartenant au thésaurus est identifié par un numéro unique « Numéro ORPHANET », un code CIM10 et le code MIM (Mendelian Inheritance in Man) McKusick (2004). Le code MIM est un nombre à six chiffres attribué aux maladies génétiques (rares).
Un nombre total de 4 268 synonymes existe dans le thésaurus ORPHANET. Ces synonymes correspondent à des formulations alternatives des termes préférés. Par exemple, les termes « syndrome de Williams-Beuren », « Monosomie 7q11.23 » et « Délétion 7q11.23 » sont des synonymes du terme préféré « syndrome de Williams ». La figure 4.1 représente un extrait de la fiche descriptive sur le site d'ORPHANET pour la maladie « syndrome de Williams » avec toutes les informations techniques (numéro ORPHANET, code CIM10 et code MIM) ainsi qu'un résumé descriptif de la maladie.

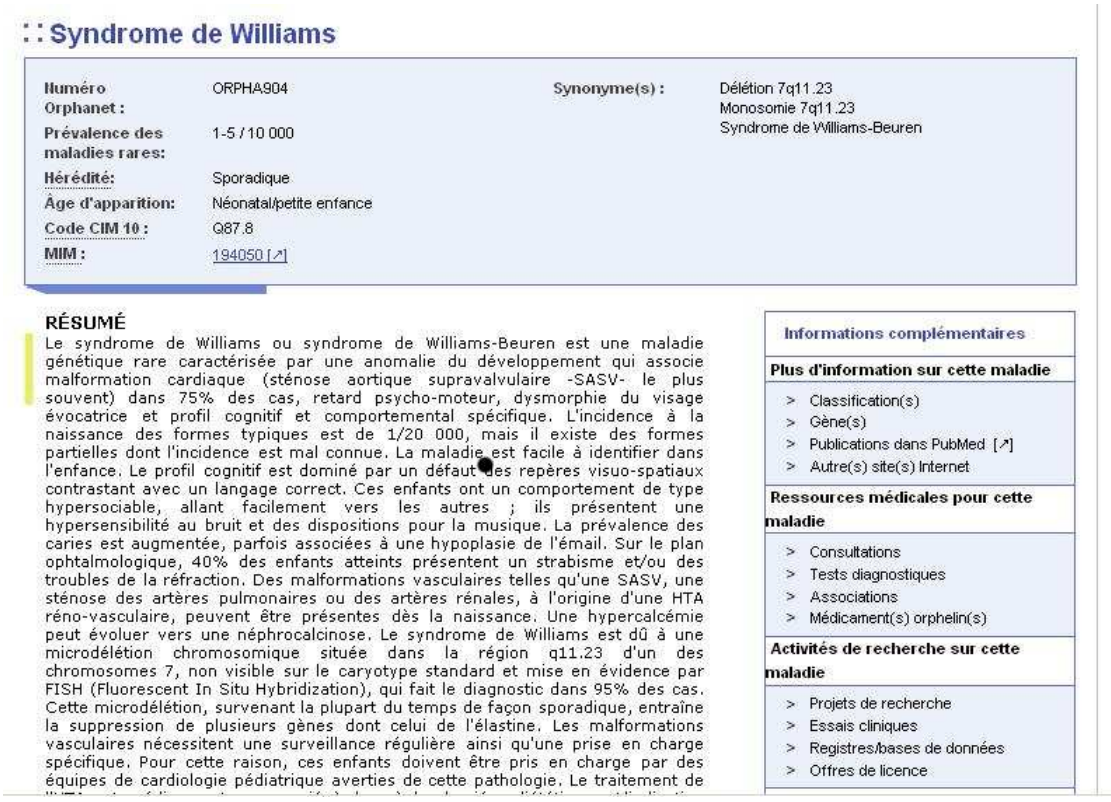

FIG. 4.1 – Exemple d'une fiche descriptive pour la maladie « syndrome de Williams »

Les termes ORPHANET sont organisés hiérarchiquement suivant des groupes de maladies dans un système de classification clinique. Environ 1 409 termes correspondant aux classifications ont été créés par l'équipe ORPHANET. Ces classifications sont organisées suivant la spécialité médicale ou chirurgicale spécifique à chaque maladie rare.

Les maladies ont été classées selon des critères cliniques ou des critères étiologiques. Les classifications sont réalisées à partir d'articles scientifiques et/ou d'avis d'experts. Elles sont régulièrement mises à jour, et de nouvelles classifications sont ajoutées. Un extrait de la classification ORPHANET pour les maladies génétiques est représenté dans la figure 4.2.

FIG. 4.2 – Extrait de la classification ORPHANET des maladies génétiques

4.2.4 Méthodes d'alignements

Pour aligner les termes ORPHANET vers les terminologies francophones de l'UMLS (F_UMLS), nous avons procédé en trois étapes : la première est l'utilisation d'un alignement manuel entre une partie des termes ORPHANET et les codes de la classification CIM10 incluses dans UMLS. La deuxième est l'utilisation d'un algorithme lexical pour aligner les termes ORPHANET et les termes francophones de l'UMLS. La dernière étape utilise une méthode structurelle qui exploite les caractéristiques de la classification ORPHANET pour aligner les termes ORPHANET qui n'ont pas été alignés lors de la deuxième étape. Les méthodes d'alignement utilisées produisent toutes un alignement entre termes et concepts UMLS. Cela à pour but d'utiliser l'alignement conceptuel produit par l'UMLS. **Un alignement conceptuel** par l'UMLS existe entre deux termes de différentes terminologies, si les deux termes partagent le même concept UMLS (voir section 3.2 pour la définition d'un concept UMLS). Par exemple, un alignement conceptuel existe entre le terme MeSH « syndrome WAGR » et le terme SNOMED International « syndrome de monosomie partielle 11p ». Par conséquent, un alignement vers un de ces deux termes impliquera un alignement vers l'autre terme. Le tableau 4.1, donne le nombre des alignements conceptuels entre les termes de chaque terminologies francophones utilisées *via* UMLS.

	CIM10	**MeSH**	**MedDRA**	**SNMI**	**WHO-ART**
CIM10		1 910	2 305	1 228	2 782
MeSH	1 910		3 490	17 907	4 149
MedDRA	2 305	3 490		7 688	6 501
SNMI	1 228	17 907	7 688		8 068
WHO-ART	2 728	4 149	6 501	8 068	

TAB. 4.1 – Nombre des alignements conceptuels *via* UMLS entre les termes de chaque terminologie francophone

Utilisation d'un alignement manuel vers CIM10 pour trouver les concepts UMLS

Cette méthode est fondée sur un alignement manuel entre les termes ORPHANET et les codes de la classification CIM10. Cet alignement a été réalisé totalement par l'équipe ORPHANET. Un nombre de 2 083 termes ORPHANET est aligné manuellement vers au moins un code CIM10 (28% de tous les termes ORPHANET). Dans cette approche, nous nous sommes limités aux concepts UMLS qui contiennent des termes MeSH. Choisir la terminologie MeSH nous permettra, entre autre, de comparer les résultats obtenus avec les méthodes automatiques et d'évaluer en plus les résultats obtenus par notre équipe (spécialiste en MeSH) Merabti *et al.* (2010a). Pour trouver les termes MeSH à partir de chaque alignement manuel, nous avons utilisé principalement l'alignement conceptuel de L'UMLS (voir paragraphe précédent). Un alignement conceptuel existe entre un code CIM10 et un terme MeSH s'ils partagent le même concept UMLS (même CUI). Par exemple, il existe un alignement conceptuel entre le code CIM10 (Code : E24) « Syndrome de Crushing » et le terme MeSH du même nom. En effet, ces deux termes partagent le même concept UMLS (CUI : C0010481).

Méthode d'alignement lexical

Dans cette méthode, nous utilisons les outils de Traitement Automatique de la Langue (TAL) pour aligner les termes ORPHANET vers les termes francophones de l'UMLS. Cette méthode permet à partir d'un terme ORPHANET de trouver les termes en français dans l'UMLS qui sont similaires lexicalement à ce terme ORPHANET. Un prétraitement sur trois étapes est appliqué sur tous les termes de toutes les terminologies utilisées (source et destination). La première étape de ce prétraitement est la tokenisation (segmentation en mots), elle consiste à découper les termes utilisés en mots. Un mot est défini comme une suite de caractères graphiques formant une unité sémantique et pouvant être distingué par un séparateur (un espace). Cependant, cette

définition reste très sommaire, en effet, plusieurs éléments sont à prendre en compte. Les règles que nous avons adoptées ont été déjà définies dans le cadre de la thèse de S. Pereira Pereira (2007) pour la segmentation des phrases en mots:

- Un mot peut être composé, accentué, il peut être un sigle ou un nom propre. De plus, un mot peut être séparé d'un autre mot par un espace ou une apostrophe.
- Les ponctuations ne constituent pas des mots mais sont de bons indicateurs de séparation de mots. Elles seront éliminées en deux temps, excepté pour les tirets (-).
- Un nombre est considéré comme un mot. De ce fait, les espaces séparant le chiffre des milliers d'autres chiffres sont à éliminer. Par contre, les virgules ou les points qui font partie intégrante du nombre ne seront pas éliminés.

La deuxième étape de notre prétraitement consiste à éliminer les mots vides (filtrage des mots). Un mot vide est un mot non significatif figurant dans le terme. Nous disposons, dans l'équipe, d'une liste de mots vides obtenus à partir de Lexique[5], créée lors de la thèse de LF. Soualmia. Cette liste a été entièrement retravaillée par S. Pereira afin d'y ajouter des mots vides. En plus des mots vides, il existe des expressions vides. Par exemple, l'expression « tout d'abord ». Une liste d'expressions vides est aussi utilisée et ajoutée à la liste des mots vides. La dernière étape de notre prétraitement applique une normalisation sur tous les termes restants composant chaque mot. Un algorithme de désuffixation en français est utilisé. La désuffixation cherche à rassembler les différentes variantes d'un mot autour d'une racine (stème). Par exemple, « antiasthmatique », « asthme », « asthmatique » ont le même stème « asthm ». Les algorithmes de désuffixation reposent généralement sur des listes de suffixes et de règles de désuffixation construites *a priori* qui permettent de trouver la racine « stème » de n'importe quel mot. Dans la thèse de S. Pereira, trois méthodes de désuffixation ont été comparées :

- L'algorithme de CISMeF développé en interne par B. Dahamna, cet algorithme traite à tour de rôle des suffixes d'une liste de 63 suffixes. Le traitement consiste à éliminer ou remplacer les suffixes rencontrés dans certaines conditions (taille du mot, le suffixe ou le mot).
- L'algorithme de Paternostre *et al.* (2002), développé par M. Paternostre dans le cadre du projet de recherche GALILEI[6] en 2002, constitue une adaptation française de l'algorithme de Porter qui traite les mots de la langue anglaise Porter (1980). L'algorithme traite les suffixes à tour de rôle, en utilisant des règles et des conditions comme l'algorithme précédent (482 règles). Les principales différences avec le précédent algorithme, outre le nombre de règles appliquées, sont

[5]Lexique fournit une base de données lexicales avec des estimations de fréquences et des formes fléchies accessibles *via* http://www.lexique.org

[6]Generic Analyser and Listner for Indexed and Linguistics Entities of Information

les conditions prises en compte. D'après Paternostre *et al.* (2002), chaque mot du français peut être réduit à cette formule : [C] $(VC)^m$ où (VC) est répété m fois (C = consonne, V = voyelle, les crochets marquent des événements optionnels).

– Le troisième algorithme est le FrenchStemmer de Lucene[7] Cutting *et al.* (2004), réalisé par P. Talbot, l'algorithme s'inspire aussi des travaux de Porter. Cet algorithme se déroule en 6 étapes : élimination des suffixes standards, traitement des suffixes verbaux, traitement des suffixes résiduels, traitement des formes particulières, traitement des caractères doubles et des accents. Pour chaque étape, une liste de règles est appliquée dépendant d'une ou plusieurs conditions, là aussi, très particulières. Ainsi, trois régions sont prises en compte pour chaque mot : RV, R1 et R2. RV est le mot. R1 est la région après la première non-voyelle suivie d'une voyelle ou la fin du mot. R2 est l'équivalent de R1 sur R1. Par exemple, pour le mot « fameusement » : RV = « fameusement », R1 = « eusement » et R2 = « ement ». Les conditions portent sur les régions, sur leurs présences ou les caractères les précédant ou les suivant.
L'évaluation effectuée par S. Pereira est détaillée dans Pereira (2007). Cette évaluation a montré que l'algorithme de Lucene est meilleur en moyenne avec un rappel de 74,7% et une précision de 81,4% par rapport à l'algorithme de carry (un rappel de 76,3% et une précision de 59,3%) et l'algorithme de CISMeF(un rappel de 69,4% et une précision de 70,9%). De ce fait, dans le reste dans nos méthodes nous utilisons l'algorithme de Lucene pour la partie désuffixation. En terme de performance il est à noter que l'algorithme de CISMeF est plus rapide par rapport aux deux autres algorithmes de désuffixation. Enfin, il faut aussi noter que l'algorithme de désuffixation de Lucene sera l'algorithme utilisé dans Doc'CISMeF à la place de l'algorithme actuel.

Le processus d'alignement utilisé dans cette méthode, est appliqué à tous les termes ORPHANET (termes préférés et synonymes) ainsi qu'à tous les termes en français de l'UMLS (termes préférés et synonymes). Il existe un alignement entre deux termes préférés si :

– Il existe un alignement entre les deux termes préférés.

– Il existe au moins un alignement entre un terme synonyme et un terme préféré des deux terminologies.

– Il existe au moins un alignement entre deux termes synonymes des deux terminologies.

Trois types d'alignement (correspondances) peuvent exister entre les termes : « Alignement Exact », « Alignement par combinaison » et « Alignement Partiel ». La figure 4.3 montre l'organigramme de l'algorithme d'alignementainsi que la précédence

[7]Lucene est un moteur de recherche libre (Open source) de la société Apache écrit en Java qui permet d'indexer et de rechercher du texte, voir : http://lucene.apache.org/

entre les types d'alignement.

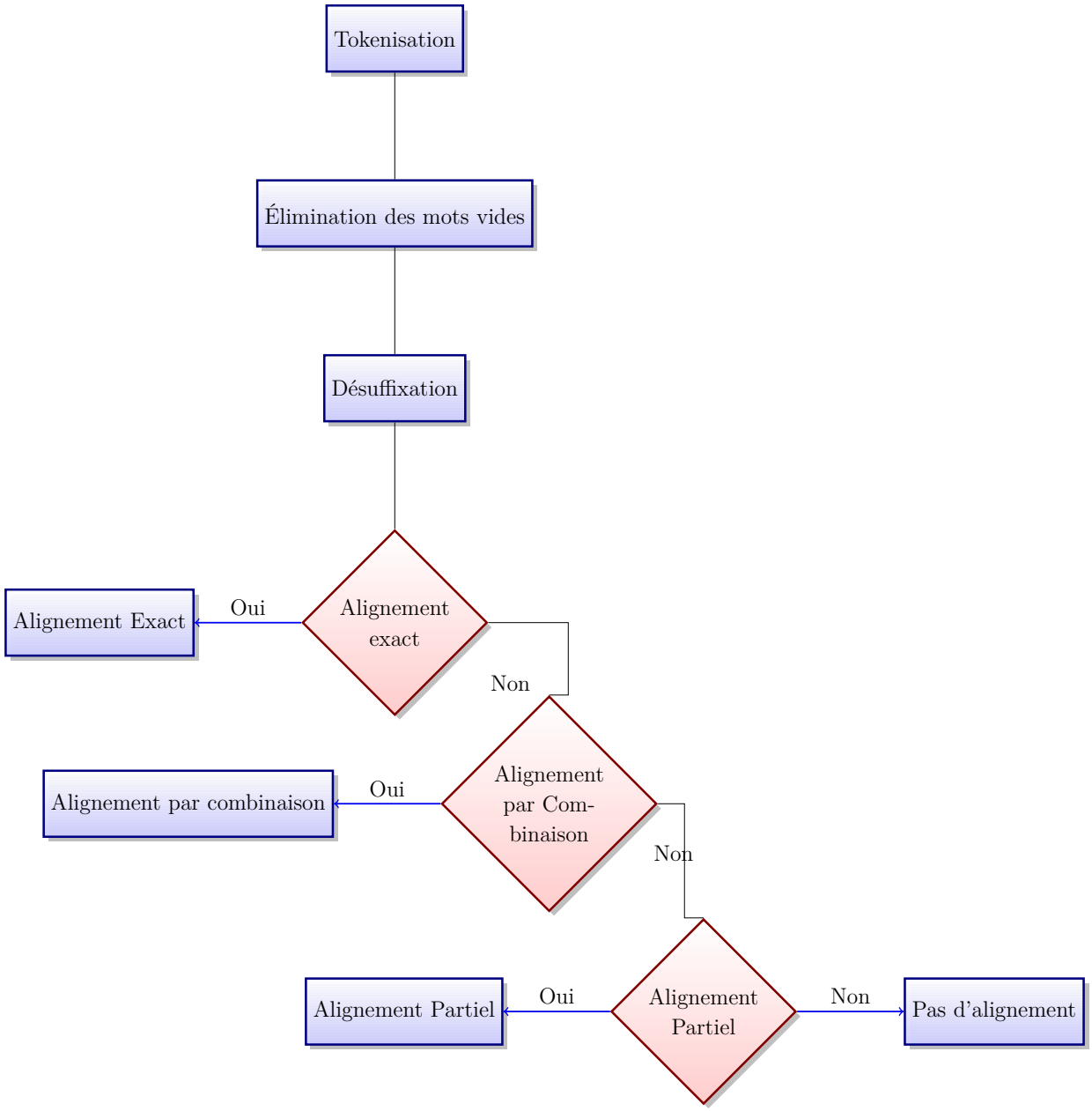

FIG. 4.3 – Organigramme de l'algorithme d'alignement

Alignement exact : Un terme ORPHANET est en relation « d'alignement exact » avec un terme d'une autre terminologie, si tous les mots composants les deux termes sont exactement similaires. La table 4.2 donne des exemples d' « alignement exact » entre les termes ORPHANET et des termes d'autres terminologies.

L'exemple de la figure 4.4 détaille le processus permettant d'obtenir un alignement exact entre le terme ORPHANET « Glycogénose de Type 2 » et le terme MeSH « Glycogénose de Type II ». Dans cet exemple, le terme ORPHANET correspond exactement à un terme synonyme dans le MeSH (MeSH SY). Le résultat

Terme ORPHANET	**Termes correspondants**	**Terminologies**
Alexandre, maladie	maladie d'Alexandre	MeSH SNOMED International
West, syndrome de	spasmes infantiles/syndrome de West	MeSH SNOMED International MedDRA
Ankylose des dents	ankylose dentaire/ankylose des dents	MeSH MedDRA SNOMED International

TAB. 4.2 – Exemples d' « alignement exact » entre termes ORPHANET et termes d'autres terminologies

de l'alignement sera entre le terme préféré (MeSH MH) du terme synonyme et le terme ORPHANET.

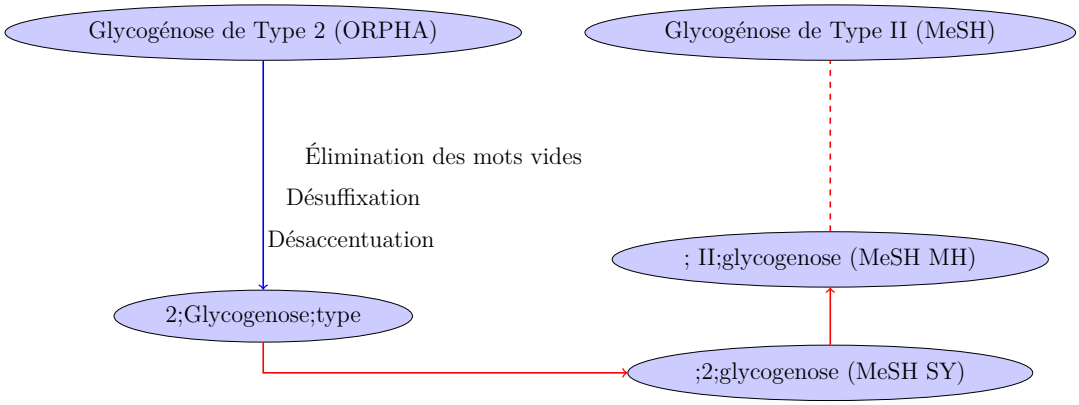

FIG. 4.4 – Exemple détaillé du processus d'alignement (Alignement exact)

Alignement par combinaison (1 à N) : Un terme ORPHANET est en relation d'« Alignement 1 à N » avec au moins deux termes d'une autre terminologie si :

- Le terme n'est pas en relation d'« Alignement exact » avec ce terme.
- La combinaison de deux ou plusieurs termes d'autres terminologies correspond exactement au terme ORPHANET.

La table 4.3 montre des exemples d' « alignement par combinaison » entre les termes ORPHANET et des termes d'autres terminologies.

Terme ORPHANET	**Termes correspondants**
Albinisme surdité	Albinisme (MeSH, MedDRA, SNOMED International) et (+) Surdité (MeSH, MedDRA, SNOMED International, WhoART)
Embryopathie diabétique	Embryopathie (MeSH, MedDRA, WHOART) et (+) Diabète (MeSH, MedDRA, WHOART)

TAB. 4.3 – Exemples d' « alignement par combinaison » entre termes ORPHANET et termes d'autres terminologies

La figure 4.5 détaille le processus permettant d'obtenir un alignement par combinaison entre le terme ORPHANET « Embryopathie diabétique » et les deux termes MeSH « Diabète » et « Maladie foetale ». Dans cet exemple, aucun terme ne correspond exactement au terme ORPHANET. Par contre la combinaison des deux termes MeSH correspond exactement au terme ORPHANET.

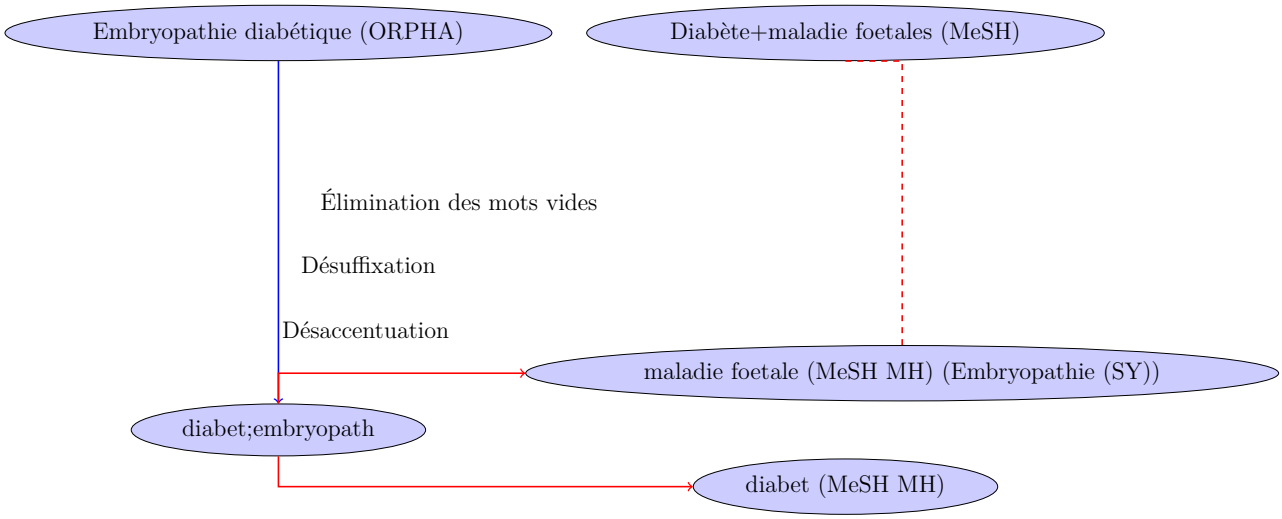

FIG. 4.5 – Exemple détaillé du processus d'alignement (Alignement par Combinaison)

Alignement partiel : Dans ce type d'alignement, une « partie » du terme ORPHANET correspond exactement ou partiellement à d'autres termes d'autres terminologies. Cet alignement est considéré par l'équipe CISMeF comme le moins pertinent par rapport aux deux autres types d'alignement. En particulier dans des tâches d'indexation automatique ou de Recherche d'information.

La table 4.4 donne des exemples d' « alignement partiels » entre les termes ORPHANET et des termes d'autres terminologies.

Terme ORPHANET	**Termes correspondants**
Chromosome 14 en anneau	Chromosome 14 (MeSH, SNOMED International) ou Chromosome en anneau (MeSH, SNOMED International)
Pseudohyperkaliémie familiale, type 1	pseudohyperkaliémie (MedDRA)

TAB. 4.4 – Exemples d' « alignement partiels » entre termes ORPHANET et termes d'autres terminologies

Méthode d'alignement structurel hiérarchique

Nous avons appliqué un alignement structurel fondé sur les relations hiérarchiques de la classification ORPHANET. Cette méthode a été utilisée pour aligner les termes ORPHANET qui ne sont pas en alignement exact avec les termes des terminologies de F_UMLS. Deux types d'alignement fondés sur la structure hiérarchique d'ORPHANET ont été utilisés :

Alignement en BT (Broader Than alignement): Un terme ORPHANET est en alignement en BT si :

- Il n'existe pas un alignement exact entre ce terme ORPHANET et un autre terme.
- Le terme ORPHANET possède au moins un parent (relation hiérarchique BT) qui est en relation en alignement exact avec au moins un autre terme de F_UMLS.

Il est essentiel de différencier des alignements en BT des alignements exacts, notamment, pour éviter le bruit dans un contexte d'indexation automatique ou de recherche d'information. Ainsi, il est donc essentiel de conserver le type de relation entre deux termes alignés. La figure 4.6 montre un exemple détaillant le processus d'alignement en BT entre un terme ORPHANET « Arthrogrypose par dystrophie musculaire » et un terme MeSH « Arthrogrypose ».

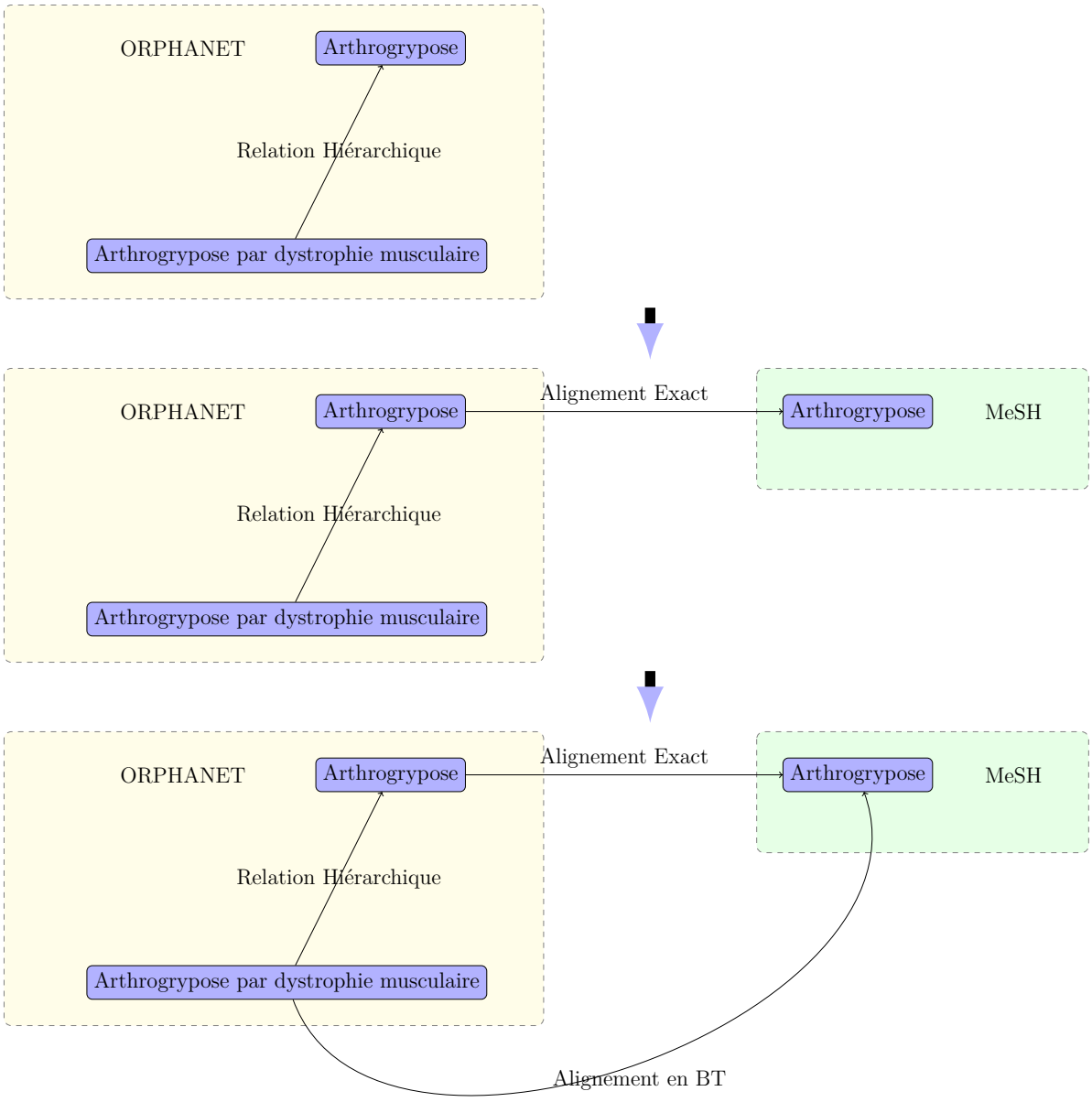

FIG. 4.6 – Exemple détaillé d'alignement structurel hiérarchique en BT

Alignement en NT (Narrower Than alignement): Un terme ORPHANET est en alignement en NT si :

- Il n'existe pas un alignement exact entre ce terme ORPHANET et un autre terme.
- Le terme ORPHANET a au moins un fils (relation hiérarchique NT) qui est en relation en alignement exact avec au moins un autre terme.

La figure 4.7 montre un exemple détaillant le processus d'alignement en NT entre un terme ORPHANET « Hétérotopie neuronale nodulaire » et un terme MeSH « Hétérotopie nodulaire périventriculaire ».

Dans les deux types d'alignement structurel présentés ci-dessus, une différence existe entre les niveaux d'alignement de chaque terme ORPHANET. Un niveau d'alignement (BT ou NT) entre un terme ORPHANET et un autre terme de F_UMLS est défini comme le nombre de relations hiérarchiques (en BT ou NT) qui existe entre le terme ORPHANET et le terme qui est en relation d'alignement exact avec le terme de F_UMLS. Par exemple, une relation d'alignement en BT de *niveau 1* existe entre un terme ORPHANET *Orpha* et un terme de F_UMLS $Terme_{FR}$ si est seulement si:

Il existe un terme *P_Orpha* tel que :

- *P_Orpha* est en relation d'alignement exact avec le terme $Terme_{FR}$;
- le nombre de relations hiérarchiques BT entre *Orpha* et *P_Orpha* est égal à 1.

D'une manière générale, une relation d'alignement en (BT ou NT) de *niveau* i existe entre un terme ORPHANET et un terme de F_UMLS $Terme_{FR}$ si et seulement si il existe un terme *P_Orpha* (ou *F_Orpha* dans le cas du NT) tel que :

- *P_Orpha* (ou *F_Orpha*) est en relation d'alignement exact avec le terme $Terme_{FR}$
- Le nombre de relations hiérarchiques BT (ou NT) entre *Orpha* et *P_Orpha* (*F_Orpha* dans le cas du NT) est égal à i.

Pour les des documentalistes de l'équipe CISMeF, plus i grandit, plus l'alignementen BT (ou NT) perd de son intérêt. Par exemple, pour la majorité des maladies du thésaurus Orphanet, on pourrait retrouver en alignement en BT un lien vers une tête d'arborescence maladie dans une autre terminologie.

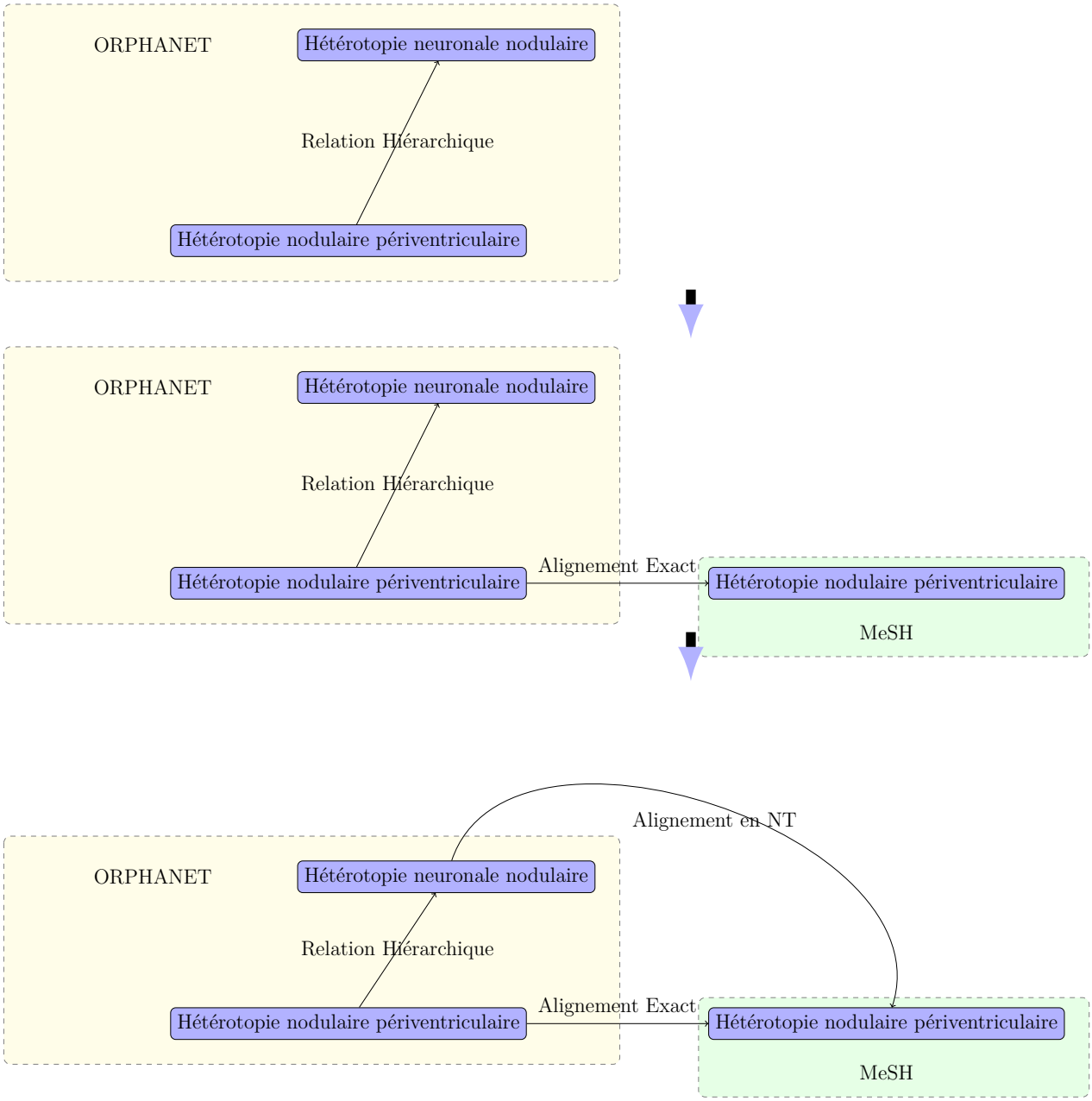

FIG. 4.7 – Exemple détaillé d'alignement structurel hiérarchique en NT

4.2.5 Critère d'évaluation et comparaison

Alignement lexical (alignement exact)

Nous avons évalué la quantité et la qualité des alignements lexicaux entre les termes ORPHANET et les termes de F_UMLS. Pour l'évaluation quantitative, nous avons mesuré principalement l'apport de l'alignement conceptuel de l'UMLS. Pour mesurer cet apport, nous avons en premier lieu appliqué notre algorithme (lexical) limité au type d'alignement exact sans utiliser l'alignementconceptuel de l'UMLS. Après, nous avons comparé ces résultats avec les résultats de l'alignement exact appliqué en utilisant l'alignementconceptuel de l'UMLS. D'un autre côté, nous montrons aussi dans le cadre de cette étude, l'apport des synonymes ajoutés par CISMeF dans la terminologie MeSH pour trouver plus de termes et de concepts UMLS. De la même façon, cet apport est mesuré par la différence de termes obtenus avec et sans utilisation des synonymes CISMeF ajoutés au MeSH. L'évaluation qualitative a été réalisée par un membre de l'équipe ORPHANET (Bertrand Bellet) pour mesurer la qualité de l'alignement entre les termes ORPHANET et les termes des autres terminologies. Cette évaluation a été faite sur un nombre de 250 alignements exacts (15%). La qualité de l'alignement entre les termes est mesurée selon deux critères :

- Pertinent : si l'alignemententre terme ORPHANET et terme de F_UMLS est jugé pertinent. C'est à dire que le terme de F_UMLS obtenu correspond exactement au terme ORPHANET. Cependant, l'évaluateur n'a pas essayé de chercher s'il existe un terme de F_UMLS plus précis que le terme proposé.
- Non pertinent : dans le cas contraire.

Nous avons aussi comparé les deux méthodes d'alignements manuelles et lexicales (alignement exact) sur les mêmes termes ORPHANET (28% alignés manuellement). Pour les comparer nous avons appliqué notre méthode d'alignement exact aux termes ORPHANET alignés manuellement. Après, nous avons limité les résultats obtenus par ce dernier alignement sur les termes MeSH. Au final, nous avons construit quatre ensembles :

Le premier ensemble : contient tous les alignements entre les termes ORPHANET et les termes MeSH obtenus uniquement en passant par l'alignementmanuel.

Le deuxième ensemble : contient tous les alignements entre les termes ORPHANET et les termes MeSH obtenus uniquement par l'approche lexicale.

Le troisième ensemble : contient tous les alignements différents obtenus par les deux approches mais pour les mêmes termes ORPHANET. Par exemple, pour le même terme ORPHANET « maladie de tangier », deux termes MeSH différents sont obtenus par les deux approches : le terme « Hyperlipémie » pour l'approche

manuel et le terme « Maladie de tangier » pour l'approche lexicale.

Le quatrième ensemble : contient tous les mêmes alignements obtenus par les deux approches.

Quatre évaluations ont été réalisées par un médecin (S.J. Darmoni). L'évaluation a été faite en aveugle sur 100 alignements de chaque ensemble choisis aléatoirement. La pertinence de chaque alignement a été évaluée selon 5 critères :

- Pertinent : si l'alignemententre le terme ORPHANET et le terme MeSH est jugé pertinent (correct) ;
- BT-NT (Broader than) : si le terme ORPHANET est jugé plus générique que le terme MeSH correspondant ;
- NT-BT (Narrower than) : si le terme ORPHANET est jugé plus spécifique que le terme MeSH correspondant. Par exemple, le terme ORPHANET « Dystrophie musculaire de Duchenne et Becker » est jugé plus spécifique que le terme MeSH « myopathie de Duchenne » ;
- Frère (du point de vue du MeSH) : si le terme ORPHANET et le terme MeSH sont jugés frères du point de vue du MeSH. Par exemple, le terme ORPHANET « Cryptophtalmie isolée » et le terme MeSH « microphtalmie » sont jugés frère (sibling). C'est-à-dire ayant un même ascendant direct sans être en relation d'alignement exacte ;
- non-pertinent.

Alignement Structurel

Une évaluation qualitative a été réalisée sur les deux types d'alignements structurels (Alignement en BT et Alignement en NT), 500 alignements en BT et 100 alignements en NT ont été évalués. Comme dans le cas de l'alignement lexical, l'évaluation a été réalisée par un membre de l'équipe ORPHANET (Ana Rath) selon trois critères qualitatifs :

- Pertinent : si l'alignementcorrespond à un alignement en BT (ou NT) (structurel)
- Exact : si l'alignementcorrespond à un alignement exact (équivalence)
- Non pertinent : si l'alignementn'est pas pertinent

4.3 Alignement de la classification ATC vers UMLS (F_UMLS)

4.3.1 La classification ATC (Anatomique, Thérapeutique et Chimique)

La classification ATC Skrbo *et al.* (2004) est utilisée pour classer les médicaments. C'est le « Collaborating Centre for Drug Statistics Methodology » de l'Organisation Mondiale de la Santé (OMS) qui la contrôle. Les médicaments sont divisés en plusieurs groupes selon l'organe ou le système sur lequel ils agissent et/ou leurs caractéristiques thérapeutiques et chimiques. Le code ATC a la forme générale LCCLLCC (où L représente une lettre et C un chiffre). Dans ce système, les médicaments sont classés en groupes à cinq niveaux différents (figure 4.8):

Le premier niveau : groupe anatomique (un caractère alphabétique).

Le deuxième niveau : groupe thérapeutique principal (deux caractères numériques).

Le troisième niveau : sous-groupe thérapeutique/pharmacologique (un caractère alphabétique).

Le quatrième niveau : sous-groupe chimique/thérapeutique/pharmacologique (un caractère alphabétique).

Le cinquième niveau : sous-groupe pour la substance chimique : le principe actif individuel ou l'association de principes actifs (deux caractères numériques).

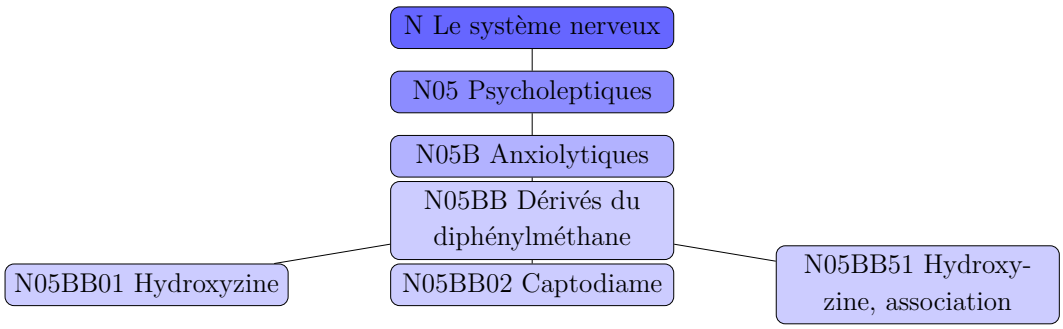

FIG. 4.8 – Les cinq niveaux différents dans ATC

À chaque stade de la classification correspond un code ATC et un libellé ATC. Le libellé du cinquième niveau correspond à la DCI (Dénominations Communes Internationales) et la substance, quand elle existe.
Ce code 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. C'est le cas pour environ 10% des médicaments qui n'ont pas le même code ATC en France et au Danemark (étude interne réalisée par la société Vidal pour le projet PSIP) Beuscart *et al.* (2009). Il a donc fallu s'adapter au contexte français et au contexte danois pour palier au problème des ATC « variants ». Cette adaptation a été rendue possible grâce à la participation de la société Vidal qui a fourni les fichiers adéquats. En 2008, l'équipe CISMeF a mis en ligne le portail PIM (Portail d'Informations sur le Médicament) Sakji *et al.* (2009b) en collaboration avec la société Vidal, spécialiste de l'information sur les médicaments. L'URL du site PIM est : `http://doccismef.chu-rouen.fr/servlets/PIM`. Ce portail a été développé pour le projet PSIP et dans le cadre de la thèse de S. Sakji. Pour des raisons de protection industrielle, ce site est protégé par un identifiant et un mot de passe. La figure 4.9 montre un exemple de recherche sur le site PIM en utilisant un code ATC.

FIG. 4.9 – Exemple de recherche utilisant un code ATC dans PIM

4.3.2 ATC vers PubMed « ATC to PubMed »

L'objectif de cette application est le développement d'un service permettant l'accès à PubMed via les codes ATC dans toutes les langues disponibles. Pour réaliser ce travail, nous avons finalisé un permier alignement manuel entre les codes ATC et les termes MeSH. Ce alignement a couvert tous les codes ATC de tous les niveaux :

1, 2, 3, 4 et 5 (N = 5 359 (97%)). Dans la majeure partie des cas, ce alignement est « 1 à N ». Par exemple, pour le code ATC « D11AX18 - diclofénac », l'alignement-MeSH est :« Diclofénac » et « produits dermatologiques » pour le diffréncier de l'autre « M01AB05 - Diclofénac » qui est aligné avec les deux termes MeSH : « Diclofénac » et « Anti-inflammatoires non stéroïdiens ». Chaque code ATC lui correspond une requête PubMed prédéfinie, qui va être lancée directement à partir du code sélectionné. Par exemple, le requête PubMed prédéfinie « desensitization, immunologic[MH] AND allergens[MH] » correspond au code ATC « VO1A » (voir figure 4.10).

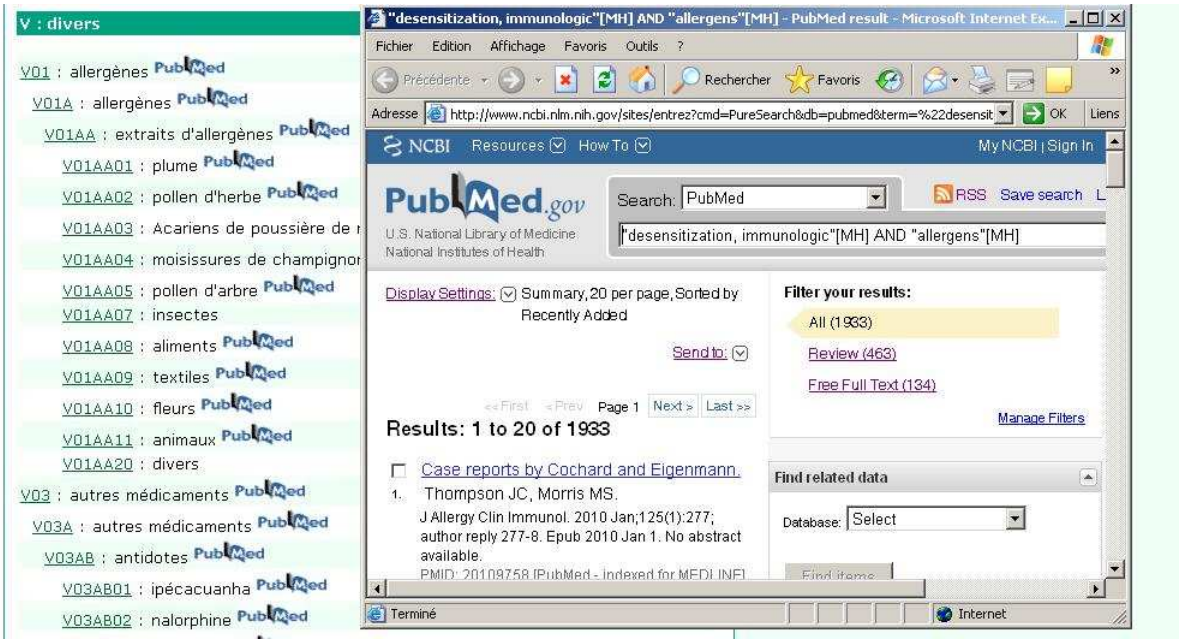

FIG. 4.10 – Capture d'écran du PIM (Partie ATC)

4.3.3 Méthodes d'alignement

Pour aligner les codes ATC avec les termes de l'UMLS nous avons utilisé deux méthodes. La première méthode utilise le même algorithme d'alignement lexical utilisé dans la section 4.2.4. La deuxième utilise une approche orientée anglais en utilisant l'outil MetaMap de la NLM Aronson (2001) effectuée par l'équipe LERTIM de Marseille. Les deux méthodes d'alignement sont appliquées aux codes ATC du cinquième niveau de la classification (4 268 codes). Ce choix est motivé par le fait que les codes du cinquième niveau correspondent aux codes les plus précis dans la classification. Les méthodes lexicales utilisées ne peuvent pas s'appliquer sur les libellés des codes de niveau supérieur (1, 2, 3 et 4). La principale cause réside dans le fait que les libellés de niveau supérieur ne sont utilisés que dans un contexte pharmaceutique, ce qui rend leur

traitement difficile avec des outils TAL. Par exemple, le code ATC « Système digestif » correspond à « Traitements médicamenteux des maladies du système digestif ». En plus, dans plusieurs cas, les libellés ATC qui appartiennent aux niveaux 2, 3, 4 ou 5 prennent également en compte les codes ATC de niveau N-1, N-2 ou N-3. Par exemple, le code ATC « chlorure de potassium » (A12BA01) doit être différencié du code ATC « chlorure de potassium » (B05XA01). Le premier code est indiqué dans le cas d'une « hypokaliémie », du fait que l'axe A prenne en compte « voies digestives et métabolisme ». Par conséquent, le code ATC A12BA01 correspondra à deux termes MeSH : « chlorure de potassium » et « hypokaliémie ». De plus, le mot clé « hypokaliémie » doit être précisé par le qualificatif MeSH « thérapie ». Ainsi, tous les alignements entre les codes ATC et MeSH ont été réalisés manuellement par un documentaliste et une pharmacienne de l'équipe CISMeF (Catherine Letord). De façon plus générale, l'alignementde l'ATC vers n'importe quelle terminologie médicale n'est pas facile :

- Une même substance peut avoir plusieurs codes ATC différents selon qu'elle est utilisée seule ou en association, selon les pathologies traitées et/ou selon sa voie d'administration.
- Les classifications chimiques varient d'une terminologie à une autre. Par exemple, dans l'ATC « la mécamylamine » (C02BB01) est considérée comme « une amine », alors que dans le MeSH c'est « un terpène ».
- La classification ATC n'est pas purement anatomique. Par exemple, les têtes d'arborescences H, J, L, P... qui ne correspondent pas à des axes anatomiques.

Alignement lexical basé sur les outils en français

Comme décrit dans la section 4.2.4. La méthode utilisée permet d'obtenir trois types d'alignement entre les codes ATC en français et les termes de F_UMLS. Un alignement exact, un alignement par combinaison et un alignement partiel. Les tableaux 4.5, 4.6, 4.7 donnent des exemples pour chaque type d'alignement entre codes ATC et termes de F_UMLS.

Alignement lexical orienté anglais (MetaMap)

Dans cette approche, nous alignons les libellés ATC en anglais vers les termes de l'UMLS en utilisant l'outil MetaMap MMTx v.2.6. MetaMap est un outil développé à la US NLM qui permet d'identifier les concepts UMLS à partir d'un texte. L'utilisation de MetaMap peut se faire sur du texte libre, mais également sur des listes de termes. L'approche que nous utilisons dans le cadre de cette méthode d'alignement, utilise

Libellés ATC (Code ATC)	**Termes correspondants**	**Terminologies**
Vincristine (L01CA02)	Vincristine	MeSH SNOMED International
Procaïne benzylpénicilline (J01CE09)	Pénicilline G procaïne	MeSH SNOMED International
Chlorure de strontium-89 (V10BX01)	Chlorure de strontium	MeSH (Concept Chimique Supplémentaire)

Tab. 4.5 – Exemples de « alignement exact » entre libellés ATC et termes d'autres terminologies

Libellés ATC (code ATC)	**Termes correspondants**
Lansoprazole, amoxicilline et métronidazole (A02BD03)	Lansoprazole (MeSH, SNOMED International) et (+) Amoxicilline (MeSH, SNOMED International) et + Métronidazole (MeSH,SNOMED International)
Oxalate de cérium (A04AD02)	Oxalate (MeSH) et (+) Cérium(MeSH)

Tab. 4.6 – Exemples de « alignement par combinaison » entre libellés ATC et termes d'autres terminologies

Libellés ATC (code ATC)	**Termes correspondants**
phytate de technétium 99m (V09DB07)	Technétium phytate (MeSH)
phosphore-32 phosphate chromique colloïdal (V10AX01)	phosphore-32 phosphate chromique (MeSH)

Tab. 4.7 – Exemples de « alignement partiel » entre libellés ATC et termes d'autres terminologies

MetaMap en prenant en entrée la liste des libellés ATC en anglais. MetaMap permet aussi de spécifier les terminologies de l'UMLS avec lesquelles nous voulons aligner les codes ATC. Le alignement entre les codes ATC et les concepts UMLS se fait de la façon suivante :

- une analyse syntaxique est appliquée sur les codes ATC en entrée.
- pour chaque syntagme, des variantes sont générées à l'aide de lexique à partir du

SPECIALIST Lexicon de l'UMLS.

- pour chaque variante générée, des concepts UMLS sont identifiés. Un concept UMLS est identifié si les chaînes de caractères contiennent au moins une des variantes.
- un score est attribué pour chaque concept UMLS proposé dans l'étape précédente, en fonction, notamment, du type de variation, de la couverture et du nombre de mots.

Trois types d'alignement ont été utilisés : un alignement exact, un alignement par combinaison et un alignement partiel. Les terminologies utilisées dans le cadre de ces trois alignements sont : MeSH, SNOMED CT, SNOMED International et NDFRT (National Drug File Reference Terminology) Carter *et al.* (2002), VANDF (Veterans Health Administration National Drug File) Nelson *et al.* (2002), NDDF (National Drug Data File Plus Source Vocabulary). Une deuxième étude utilise seulement les terminologies de F_UMLS.

Les tableaux 4.8, 4.9, 4.10 donnent des exemples pour chaque type d'alignement entre codes ATC et termes de F_UMLS utilisant MetaMap.

Libellés ATC (Code ATC)	**Termes correspondants**	**Terminologies**
glutamic acid hydrochloride (A09AB01)	glutamic acid hydrochloride	SNOMED International
pyridoxal phosphate (A11HA06)	pyridoxal phosphate	MeSH SNOMED International

Tab. 4.8 – Exemples de « alignement exact » entre libellés ATC et termes d'autres terminologies

Libellés ATC (code ATC)	**Termes correspondants**
magaldrate and antiflatulents (A02AF0A)	magaldrate (MeSH, SNOMED International) et (+) antiflatulents (MeSH, SNOMED International)

Tab. 4.9 – Exemples de « alignement par combinaison » entre libellés ATC et termes d'autres terminologies

Libellés ATC (code ATC)	**Termes correspondants**
picodralazine and diuretics (C02LG02)	Diuretics (MeSH, SNOMED International)

TAB. 4.10 – Exemples de « alignement partiel » entre libellés ATC et termes d'autres terminologies

4.3.4 Critères d'évaluation et comparaison

Pour les deux méthodes d'alignement, des évaluations quantitatives ont été réalisées. Pour l'alignement lexical utilisant les outils en français, nous avons mesuré le nombre de concepts UMLS couverts pour chaque type d'alignements (exact, combinaison et partiel). Nous avons aussi, comme dans le cas de l'alignement ORPHANET avec F_UMLS, mesuré l'apport de l'alignement conceptuel de l'UMLS. Nous avons aussi mesuré l'apport des synonymes CISMeF ajoutés au MeSH ainsi que les concepts chimiques supplémentaires traduits en français par l'équipe CISMeF sur l'alignementdes codes ATC vers le MeSH.
Pour l'évaluation des différents alignements (français, anglais) vers UMLS, nous avons utilisé un alignement manuel des codes ATC vers le MeSH comme « Gold Standard ». Ce dernier a été réalisé par Catherine Letord documentaliste et pharmacienne de l'équipe CISMeF, élaboré pour permettre une recherche d'information dans Doc'CISMeF par les codes ATC ou dans le cadre d'ATC vers PubMed décrit dans la section 4.3.2. La figure 4.11 montre un exemple d'interrogation de Doc'CISMeF par un code ATC « A01AA51 » (« sodium fluorure en association »). La méthode d'évalua-

FIG. 4.11 – Exmple de recherche dans Doc'CISMeF par un code ATC

tion consiste à comparer les codes ATC alignés par les deux méthodes avec les codes ATC alignés manuellement (N = 4 108). Toutefois, il demeure un problème lié à l'ali-

gnement manuel. En effet, plusieurs codes ATC (N = 1 137 (%26)) ont été alignés manuellement vers plusieurs termes MeSH. Par exemple, le code ATC « A01AA51 » à qui correspond le libellé « sodium fluorure en association » est aligné vers deux termes MeSH : « fluorure de sodium » et « association de médicaments ». Ce qui rend difficile la comparaison avec les deux alignements sur leurs ensembles de « alignement exact ». Cependant, une solution peut être envisagée en mesurant la précision en calculant le nombre de mots qui composent le terme trouvé automatiquement par rapport nombre de mots du terme aligné manuellement. Dans l'exemple précédant le résultat serait de 1/2 si l'alignementautomatique trouve le terme : « fluorure de sodium ».
D'un autre côté, nous avons filtré les deux ensembles d'alignements des deux méthodes sur les concepts UMLS contenant que les codes MeSH. Pour la méthode fondée sur les outils en français nous avons un nombre de 2 898 codes ATC alignés vers au moins un concept UMLS contenant un terme MeSH. Pour la méthode fondée sur l'outil MetaMap nous avons un nombre de 2 695 codes ATC alignés vers au moins un concept UMLS contenant un terme MeSH. Pour chaque méthode d'alignement nous avons construit trois ensembles :

- un premier ensemble correspondant à tous les alignements validés, c'est-à-dire tous les alignements obtenus manuellement et automatiquement (anglais ou français) ;
- un deuxième ensemble correspondant à tous les alignements obtenus seulement automatiquement ;
- un troisième ensemble correspondant à tous les alignements valides qui n'ont pas été obtenus automatiquement.

4.4 Alignement de la classification CCAM avec UMLS (F_UMLS)

4.4.1 La Classification Commune des Actes Médicaux (CCAM)

La CCAM Rodrigues *et al.* (2005a) est le référentiel des actes médicaux qui remplace, pour les médecins, la Nomenclature Générale des Actes Professionnels (NGAP[8]) en secteur libéral, et le Catalogue des Actes Médicaux (CDAM[9]) en secteur hospitalier français. Elle permet la tarification des actes en médecine libérale.
Élaborée par la CNAMTS (Caisse Nationale d'Assurance Maladie des Travailleurs Salariés) et l'ATIH (Agence Technique de l'Information sur l'Hospitalisation), en étroite collaboration avec les sociétés savantes, la CCAM a été créée afin d'obtenir une liste unique d'actes codés, commune aux secteurs publics et privés pour les professionnels de la santé afin de garantir la cohérence des systèmes d'information et de satisfaire les professionnels par l'utilisation d'un seul outil. Elle est destinée à décrire plus précisément chaque acte, à servir de base à la tarification en secteur libéral (cabinet clinique) et à l'allocation de ressources aux établissements publics dans le cadre de la tarification à l'activité (T_2A). La réalisation de la CCAM a associé la méthode traditionnelle de développement à dire d'expert Rodrigues *et al.* (2005b), et une représentation formelle GALEN GAL (2005). Cette représentation formelle a été réalisée dans le cadre du consortium européen GALEN (Genralised Architecture for Languages, Encyclopaedias and Nomenclatures in Medecine) Rector *et al.* (1993) composé de plusieurs universités, centres de recherche et sociétés informatiques. Ce projet des années 90 a été dédier à une terminologie clinique unifiée. L'ontologie développée, le Common Reference Model, est basée sur une approche descendante. GALEN est défini avec une logique de description GRAIL (Galen representation and Integration Language) Rector *et al.* (1997). Elle possède un lien sémantique avec la CIM10, crée par Jacques Chevallier Chevallier (2006). D'autres liens vers la CIM10 ont été créés dans Avillach *et al.* (2007) en se basant sur une matrice de co-occurrence à partir des codes CIM10 et CCAM contenus dans les dossiers médicaux du patient.
Le classement de la CCAM correspond à une logique médicale et se fait par grand appareil et non par spécialité. La CCAM est une hiérarchie à héritage simple organisée en 19 chapitres. Les 17 premiers chapitres sont scindés en deux parties : la première concerne les actes diagnostiques rangés par grande technique puis par organe, la se-

[8]La NGAP est la nomenclature de médecine ambulatoire

[9]Le CDAM, publiée en 1985, a été élaborée par des comités d'experts médicaux coordonnés par la Direction des Hôpitaux. Il répondait à deux objectifs : identifier les actes réalisés pendant le séjour du patient et mesurer la consommation en ressources humaines et matérielles pour réaliser cet acte.

conde concerne les actes thérapeutiques classés par organe puis par action ; le chapitre 18 regroupe les gestes complémentaires ; le chapitre 19 prend en compte les adaptations pour la CCAM transitoire.
La CCAM est fondée sur le principe de l'acte global : chaque libellé comprend implicitement l'ensemble des gestes nécessaires à la réalisation de l'acte. De plus les libellés sont non ambigus c'est-à-dire sans possibilité d'interprétation divergente. Elle est aussi bijective c'est-à-dire qu'à un libellé correspond un code et un seul et réciproquement (voir figure 4.12).
Dans le cadre de cette thèse nous nous sommes intéressés à la version 16 de la CCAM avec 7 926 actes, la version la plus récente est la version 19 applicable au 01/02/2010[10].
Chaque libellé de dernier niveau de la CCAM correspond un code à 7 caractères alphanumériques : les 4 premiers sont signifiants (topographie (anatomie), action, voie d'abord et/ou technique), les 3 derniers constituent un compteur séquentiel.

AA - AA - NNN
Topographie (Anatomie) Action Voie d'abord et/ou technique Compteur

- Le premier code constitue le codage du système. Par exemple, « Système respiratoire » (G).
- La deuxième lettre constitue le codage de l'organe ou de la fonction. Par exemple, « Plèvre » (GG). La tableau 4.11 montre un extrait de la table de codage de la CCAM pour la topographie et la fonction (Système respiratoire).
- La troisième lettre correspond au codage de l'action principale du libellé. Par exemple, « Évacuer » (J). La tableau 4.12 montre un extrait de la table de codage de la CCAM pour les actions.
- La quatrième lettre code le mode d'accès ou la technique utilisée. Par exemple, « Abord ouvert » (A). La tablaue 4.13 montre un extrait de la table de codage de la CCAM pour pour les modes d'accès.
- Chaque code à 4 caractères est affecté d'un compteur à 3 chiffres, pour différencier les actes ayant le même code anatomique, le même code d'action et le même code de voie d'abord ou de technique. Par exemple, « Évacuation de collection de la cavité pleurale, par thoracotomie sans résection costale » (GGJA002) et « Évacuation de collection de la cavité pleurale, par thoracotomie avec résection costale » (GGJA004).

Des caractères supplémentaires aux codes peuvent être ajoutés, ceux-ci permettent de :

[10] au moment de la rédaction

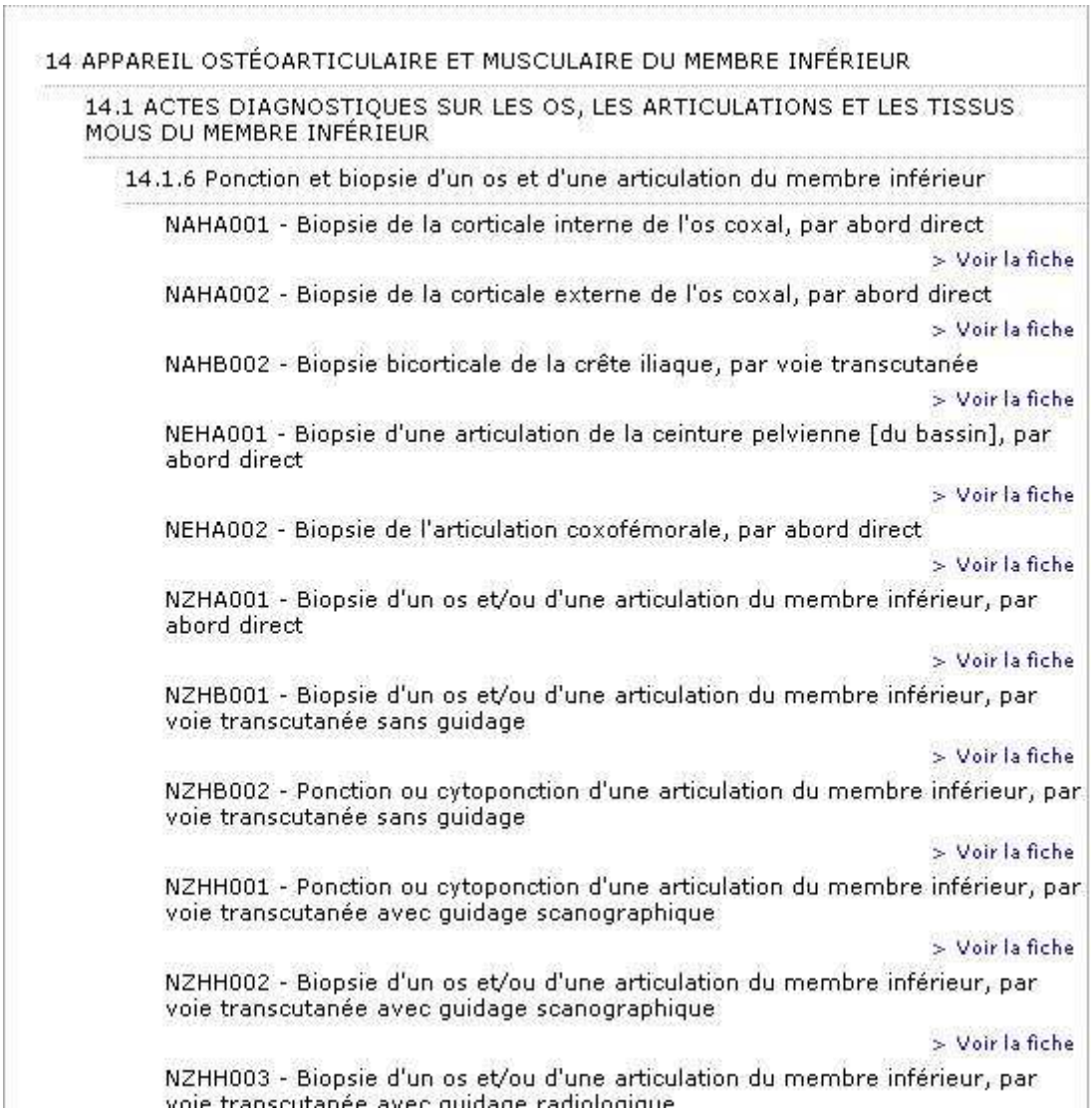

FIG. 4.12 – Extrait du chapitre 14 de la CCAM

- Décrire l'activité : permet de différencier et énumérer les gestes réalisés au cours d'un même acte par des intervenants différents (valeurs 0 à 5).
- Préciser l'extension documentaire : une lettre qui permet de donner un niveau de détail supplémentaire mais non utile à la tarification (10 valeurs possibles). Par exemple, pour le terme « dilatation intraluminale d'une branche de l'aorte abdominale à destinée digestive avec pose d'endoprothèse, par voie artérielle transcutanée » (EDAF005), nous avons entre autres les codes documentaires : « tronc iliaque » (F) et « artère gastrique gauche » (G).
- Préciser la phase de traitement : pour distinguer les différentes phases d'un acte en terme de coût et de séjour d'hospitalisation. Par exemple pour le terme « reconstruction d'un tendon de la main par transplant libre, en deux temps » (MJMA006),

G	**Système respiratoire**
GA	Nez
GB	Sinus paranasaux
GC	Rhinopharynx et fosse infratemporale
GD	Larynx et épiglotte
GE	Trachée et arbre bronchique
GF	Poumons
GG	Plèvre
GH	Espace médiastinal
GJ	Odorat
GK	Langage, phonation
GL	Respiration
GZ	Système respiratoire, sans précision

TAB. 4.11 – Extrait de la table de codage de la CCAM pour la topographie (Système respiratoire)

il existe deux phases : « reconstruction de la gaine fibreuse digitale avec pose de prothèse provisoire, par abord direct ou avec ou sans réfection des poulies » (MJMA006 1 1) et « transplant libre de tendon de la main » (MJMA006 1 2).

- Des codes influant sur la tarification peuvent être juxtaposés :
  - L'application des codes modificateurs indique les circonstances particulières de réalisation de l'acte et peut entrainer une majoration du coût du séjour.
  - Un code association qui permet de signaler des associations d'actes non prévues.
  - Un code remboursement exceptionnel.
  - Un code supplément pour un acte en cabinet (code C).

Chaque code est suivi par son tarif en euros et de précisions tarifaires, de caractéristiques générales et de précisions sur le codage et de plus de 200 autres critères divers. Plusieurs actes peuvent être associés (4 au maximum). Toutefois, il existe des associations d'actes interdites, elles sont identifiées et listées.

4.4.2 Méthodes d'alignement

Pour l'alignementdes codes CCAM vers UMLS, nous avons utilisé deux méthodes, une basée sur les outils de TAL en français Merabti *et al.* (2010b), et une deuxième

Verbe	**Définition**	**Termes utilisés**	**Lettre**
Agrandir	augmenter les dimensions (longueurs, calibre, surface ou volume) d'un élément de l'organisme	AGRANDISSEMENT ALLONGEMENT APPROFONDISSMEMENT DILATATION DISTENSION ÉLARGISSEMENT RECALIBRAGE REHAUSSEMENT	A
Appliquer	disposer d'un agent thérapeutique à visée locale ou générale à la surface de l'organisme ou d'une de ses parties, sans effraction des téguments	APPLICATION -APPLICATION	L
Combler	empiler un espace ou une cavité en y apportant un matériau biologique ou artificiel	APPOSITION COMBLEMENT CRANIALISATION ENROBAGE INTERPOSITION OBTURATION RECOUVREMENT	B

TAB. 4.12 – Extrait de la table de codage de la CCAM pour les actions

utilisant l'outil MetaMap Bousquet *et al.* (2010). Ce dernier a été effectué par l'équipe DSPIM de Saint-Étienne (Département de Santé de Public et de l'Information Médicale).
Ces deux approches ont été utilisées auparavant pour aligner les codes ATC vers les

Mode d'accès	**Définition**	**Termes utilisés**	**Lettre**
Abord ouvert	accès exposant le site opératoire, par incision des téguments et de tout autre tissu sousjacent, sans introduction d'un instrument d'optique. Par extension, concerne tout accès à travers la peau par une ouverture cutanée d'origine	à foyer ouvert par...abord... par dissection... par excision par cranitomie par sclérotomie ...	A
Accès endoscopique	accès au site opératoire par ponction ou incision minime des téguments et de tout autre tissu sousjacent, avec introduction d'un instrument d'optique	par cervicoscopie par coelioscopie par médiastinoscopie ...	C

Tab. 4.13 – Extrait de la table de codage de la CCAM pour les modes d'accès

concepts UMLS. Contrairement à la classification ATC, la longueur des libellés de la CCAM rend impossible l'alignementexact vers des concepts UMLS. En effet, plus de 85% des libellés de la CCAM possèdent plus de 5 mots. Contrairement à une terminologie comme le MeSH où 5% seulement de ces libellés contiennent plus de 5 mots. Ainsi, une approche alternative basée sur la structure du code de la CCAM (voir 4.4.1) s'impose, au lieu d'appliquer l'alignementdepuis les libellés de la CCAM. L'utilisation de la structure du code de la CCAM permet d'en extraire un certain nombre de termes élémentaires. Ces derniers correspondent à la signification de chaque axe composant un code CCAM. Au final, l'alignemententre les codes CCAM et les termes de l'UMLS

est effectué en deux étapes :

1. Une étape de prétraitement permettant d'extraire les termes correspondant aux axes composant chaque code de la CCAM.
2. L'étape d'alignement « des nouveaux termes » représentatifs vers les termes (français ou anglais) de l'UMLS en utilisant les deux méthodes d'alignements.

Prétraitement des codes CCAM

L'étape de prétraitement, commune aux deux méthodes, consiste à extraire les termes représentant les axes de chaque code CCAM. Cette méthode est fondée principalement sur la structure de tous les codes CCAM. La concaténation des nouveaux termes obtenus constituera le « nouveau libellé représentatif » du code. L'avantage de cette approche est qu'au final nous avons un libellé moins verbeux qui sera aligné vers UMLS. L'extraction a été réalisée sur les trois premiers caractères composant chaque code. Ce choix a été conseillé par l'un des principaux experts de la CCAM (J.M Rodrigues). Le quatrième caractère étant non significatif dans la construction du code n'a pas été choisi. Les deux premiers caractères représentent l'axe anatomique. Une table dont un extrait a été présenté tableau 4.11 nous a servi à faire la correspondance. Au total, 194 termes sont utilisés pour décrire l'axe topographique. Par exemple, pour le code CCAM AAGB001 « Ablation d'électrode intracérébrale, par voie transcutanée » aux deux premiers caractères « AA » corresponde le terme « Encéphale ». Le tableau 4.14 dresse quelques exemples de codes CCAM avec les termes correspondants sur l'axe anatomique.

Code CCAM	**Libellé CCAM**	**Topographie**	**termes correspondants**
ABCA001	Ventriculoventriculostomie, kystocisternostomie ou kystoventriculostomie, par craniotomie	AB	Ventricules, méninges et LCR intracrâniens
EBFA017	Thromboendartériectomie de l'artère vertébrale proximale, par cervicotomie	EB	Vaisseaux de la tête et du cou, extracrâniens ou non précisé

Tab. 4.14 – Exemples de codes CCAM avec les termes correspondant à l'axe Anatomique

L'information pour l'axe « action » est plus difficile à extraire. En effet, une lettre peut représenter plusieurs actions. Par exemple, la lettre « A » peut correspondre aux

termes suivants : « ALLONGEMENT », « AGRANDISSEMENT », « APPROFONDISSEMENT » ... (voir tableau 4.12)
Pour résoudre ce problème, nous avons utilisé le libellé du code. Pour chaque code de la CCAM, nous dressons la liste des termes susceptibles de correspondre à la lettre représentant l'axe action. Après, nous cherchons dans les mots composant le libellé CCAM le terme qui correspond à un terme de la liste. Le tableau 4.15 montre quelques exemples de codes CCAM avec la même lettre pour décrire l'action mais avec des termes correspondants différents. Au total, 331 termes sont utilisés pour décrire l'axe des actions.

Code CCAM	**Libellé CCAM**	**Action**	**termes correspondants**
AAFA001	Exérèse de tumeur intraparenchymateuse du cervelet, par craniotomie	F	exérèse
AAFA007	Excision d'une zone épileptogène, par craniotomie	F	excision
AAFA008	Résection de parenchyme cérébelleux pour infarctus expansif, par craniotomie	F	résection

TAB. 4.15 – Exemples de codes CCAM avec le même troisième caractère mais avec différentes actions

Le tableau 4.16 donne des exemples de codes CCAM avec le libellé originel et le nouveau « libellé représentatif ».

Code CCAM	**Libellé CCAM**	**termes correspondants**
AAQM00	Échographie transfontanellaire de l'encéphale	Encéphale (AA) + échographie (Q)
BDHA001	Biopsie de la cornée	Cornée (BD) + biopsie (H)
FEFF002	Prélèvement de cellules souches hématopoïétiques sanguines par cytaphérèse, pour thérapie cellulaire	Sang (FE) + prélèvement (F)

TAB. 4.16 – Exemples de codes CCAM avec nouveaux termes correspondants

Alignement lexical basé sur les outils en français

Comme pour l'alignementdes codes ATC avec F_UMLS, la méthode fondée sur les outils en français produit trois types d'alignements entre les « nouveaux libellés » représentant les codes CCAM et les termes de F_UMLS. Les tableaux 4.17, 4.18, 4.19 donnent des exemples pour chaque type d'alignement entre codes CCAM et termes de F_UMLS.

Libellés CCAM (Code CCAM)	**Nouveau terme (Topographie +Action)**	**Termes correspondants**
Abrasion mécanique de l'épithélium de la cornée avec laser (BDNP003)	Cornée + abrasion	Abrasion de la cornée
Adaptation bilatérale de verre scléral obtenu par moulage (BZMP001)	œil, sans précision + adaptation	Adaptation de l'oeil (MeSH + SNOMED International)

Tab. 4.17 – Exemples de « alignement exact » entre codes CCAM et termes de F_UMLS

Libellés CCAM (Code CCAM)	**Nouveau terme (Topographie +Action)**	**Termes correspondants**
Exérèse de lésion du tronc cérébral, par craniotomie (AAFA003)	Encéphale + Exérèse	Encéphale (MeSH + SNOMED International)+ Exérèse SAI (MedDRA)
Ventilation mécanique discontinue au masque facial ou par embout buccal pour kinésithérapie, par 24 heures (GLLD002)	Respiration + Ventilation	Respiration (MeSH + SNOMED International) + Ventilation (MeSH)

Tab. 4.18 – Exemples de « alignement par combinaison » entre codes CCAM et termes de F_UMLS

Libellés CCAM (Code CCAM)	**Nouveau terme (Topographie +Action)**	**Termes correspondants**
Extraction unilatérale ou bilatérale de bouchon de cérumen ou de corps étranger du méat acoustique externe (CAGD001)	Oreille externe + extraction	Oreille externe (MeSH + SNOMED International)
Résection de parenchyme cérébral pour infarctus expansif, par craniotomie (AAFA006)	Encéphale + Résection	Encéphale (MeSH + SNOMED International)

Tab. 4.19 – Exemples de « alignement partiels » entre codes CCAM et termes de F_UMLS

Alignement lexical basé sur MetaMap

Le même principe est utilisé pour l'alignementdes codes CCAM vers UMLS que celui utilisé pour aligner les codes ATC vers UMLS. L'outil MetaMap prend en entrée les termes correspondant aux trois premiers caractères de chaque code CCAM. Un score de similarité est associé entre les termes à aligner et les termes équivalents (concepts équivalents). Dans le cadre de cette étude, l'équipe DSPIM[11] n'a pris en considération que le « alignement exact » Bousquet *et al.* (2010), les alignements avec un score de 100%. L'exemple de la figure 4.13 montre un alignement d'un code CCAM vers UMLS en utilisant MetaMap. Les trois types d'alignement qui peuvent exister entre un code CCAM et un terme de UMLS sont :

- Un alignement exact : s'il existe au moins deux termes dans UMLS qui sont équivalents aux termes correspondant aux axes anatomique et action.
- Un alignement sur l'axe anatomique : s'il existe au moins un terme dans UMLS équivalent au terme correspondant à l'axe anatomique.
- Un alignement sur l'axe action : s'il existe au moins un terme dans UMLS équivalent au terme correspondant à l'axe action.

Les tableaux 4.20, 4.21 donnent des exemples pour chaque type d'alignement entre codes CCAM et termes d'UMLS en utilisant MetaMap.

[11] www.univ-st-etienne.fr/dspim/

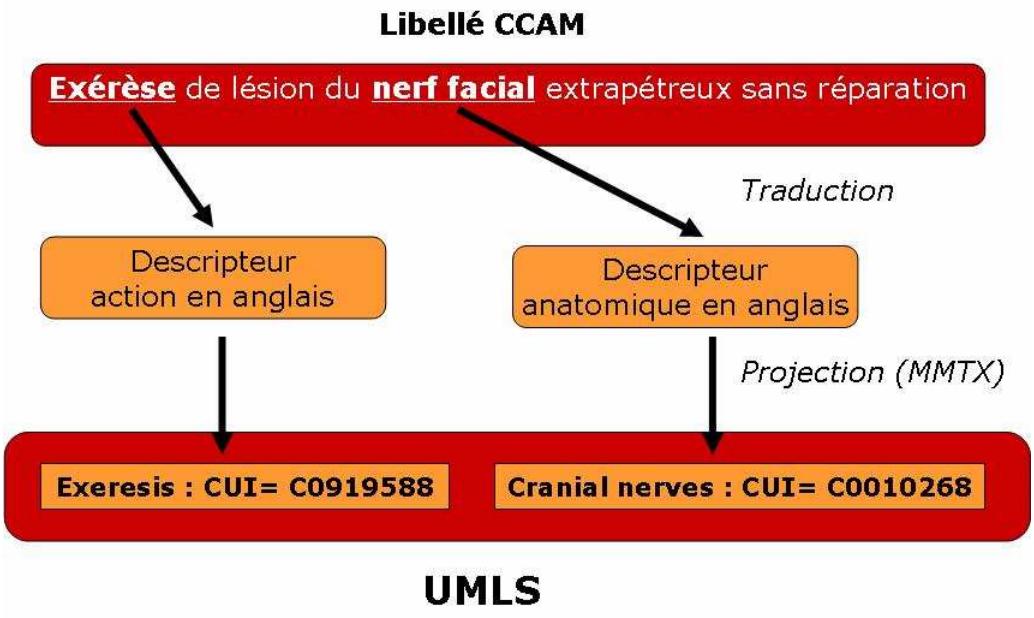

FIG. 4.13 – Exemple d'alignement de code CCAM vers UMLS utilisant MetaMap

Libellés CCAM (Code CCAM)	**Nouveau terme (Topographie + Action)**	**Termes correspondants**
Surveillance électroencéphalographique continue sans enregistrement vidéo, par 24 heures (AAQP003)	Brain (Encéphale) Surveillance (Surveillance)+	Brain: Entire brain (MeSH, SNOMED International, SNOMED CT ...) + Medical Surveillance (MeSH, SNOMED CT...)

TAB. 4.20 – Exemples de « alignement sur les deux axes » entre codes CCAM et termes de l'UMLS en utilisant MetaMap

4.4.3 Critères d'évaluation et comparaison

Pour les deux méthodes d'alignement, des évaluations quantitatives ont été réalisées.
Pour l'alignementlexical utilisant les outils en français, une étude qualitative a été faite sur tous les « alignements exacts » et sur un nombre de 100 alignements de type « par combinaison ». Le dernier type d'alignement « partiel » a été jugé non pertinent. Pour le deuxième type d'alignement (par combinaison) nous n'avons choisi que 100 alignements car plusieurs codes partagent les mêmes axes (topographiques et actions). Ainsi, les nouveaux termes correspondants après l'étape de prétraitement sont les mêmes. Par exemple, les codes CCAM HLHH003 et HLHH004 partagent les mêmes axes topogra-

Libellés CCAM (Code CCAM)	**Nouveau terme (Topographie + Action)**	**Termes correspondants**
Réparation de plaie non linéaire perforante du bulbe [globe] oculaire intéressant au moins 3 des structures suivantes : cornée, sclère, iris, corps ciliaire, cristallin, vitré, rétine (BHMA002)	Eyeball (bulbe oculaire) + reparation (Réparation)	Eye (MeSH, SNOMED International, SNOMED CT...)
Exérèse et/ou fermeture de méningoencéphalocèle de la base du crâne (ABFA001)	Intracranial ventricles, meninges and cerebrospinal fluid (Ventricules, méninges et liquide céphalorachidien intercrâniens) + exeresis (Exérèse)	Exeresis (MedDRA)

TAB. 4.21 – Exemples de « alignement sur un axe » entre codes CCAM et termes de l'UMLS en utilisant MetaMap

phique et action. Par conséquent, les termes correspondants (concepts UMLS) sont les mêmes. L'étude qualitative a été faite par un médecin spécialiste dans le codage en CCAM (Philippe Massari). La pertinence de chaque alignement a été évaluée suivant cinq critères :

- Équivalent : si le terme dans UMLS correspond exactement aux libellés CCAM.
- BT-NT (Broader than) : si le libellé CCAM est jugé comme plus générique que le terme dans F_UMLS correspondant.
- NT-BT (Narrower than) : si le libellé CCAM est jugé comme plus spécifique que le terme dans F_UMLS correspondant.
- Incomplet : si le terme correspondant dans F_UMLS ne reflète qu'une partie du libellé de la CCAM.
- Non-pertinent : si l'alignementest incorrect.

4.5 Synthèse

Nous avons décrit dans ce chapitre nos méthodes d'alignements entre terminologies médicales. Nous avons appliqué nos méthodes pour aligner trois terminologies médicales : ORPHANET, ATC et CCAM. Dans tous les travaux d'alignement que nous avons présentés, nous avons décrit des méthodologies appliquées différemment pour aligner les terminologies sources vers UMLS (F_UMLS). Ainsi, pour ORPHANET, en plus de l'algorithme d'alignement lexical, nous avons combiné les résultats obtenus avec une méthode d'alignement structurel utilisant les relations hiérarchiques du thésaurus ORPHANET. Dans le cas de la classification ATC, nous avons utilisé conjointement notre méthode d'alignement fondée sur le français avec l'outil MetaMap de la NLM fondé sur l'anglais. Pour le dernier cas de la CCAM, l'utilisation de nos méthodes d'alignements directement sur les libellés de la CCAM était impossible à cause de leurs tailles en terme de nombre de mots. Ainsi, nous avons développé une méthode permettant d'extraire à partir de la structure du code de la CCAM un nouveau « libellé représentatif » moins verbeux. Comme pour la classification ATC, nous avons aussi dans le cas de la CCAM utilisé conjointement nos méthodes avec l'outil MetaMap sur les nouveaux libellés représentatifs obtenus.
Dans le chapitre suivant, nous exposons les résultats et les évaluation pour pour les différents alignements réalisés.

Chapitre 5

Résultats et évaluations : Alignement des terminologies francophones vers UMLS (F_UMLS)

Dans ce chapitre nous présentons les différents résultats obtenus pour les alignements appliqués dans le chapitre précédente. Pour chaque terminologie alignée, nous exposons les résultats obtenus en terme de quantité et de qualité.

5.1 Alignement du thésaurus ORPHANET avec F_UMLS

5.1.1 Résultats

Alignement manuel

Sur les 2 083 termes ORPHANET alignés manuellement vers au moins un code CIM10, 619 alignements possibles sont obtenus vers au moins un terme MeSH en passant par l'UMLS (30% des 2 083).

Alignement lexical

Sur les 7 424 termes ORPHANET du thésaurus, 1 671 (22%) sont en relation alignement exact avec au moins un terme de F_UMLS couvrant 2 802 concepts UMLS. D'autre part, 2 084 (27%) termes ORPHANET sont en relation alignement par combinaison avec au moins deux termes de F_UMLS couvrant 4 397 concepts UMLS. Le tableau 5.1 montre le nombre de termes ORPHANET en relation et le nombre de concepts UMLS couverts suivant chaque type d'alignement utilisé.

Nombre de termes ORPHANET	**Nombre de concepts UMLS couverts**	**Type d'alignement**
1 671 (22%)	2 802	Alignement exact
2 048 (27%)	4 397	Alignement par combinaison
3 483 (46%)	4 529	Alignement partiel

TAB. 5.1 – Nombre de termes ORPHANET en correspondance pour chaque type d'alignement

Les tableaux 5.2, 5.3 et 5.4 fournissent le nombre de termes de chaque terminologie qui sont en relation avec les termes ORPHANET pour chaque type d'alignement.

Type d'alignement	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
Alignement exact	2 084	1 125	MeSH
	2 093	1 848	SNOMED International
	3 721	3 721	MedDRA
	409	249	WHO-ART
	47	47	CISP2

TAB. 5.2 – Nombre de termes de chaque terminologie en relation alignement exact

L'équipe CISMeF dans son ensemble a été très surprise des résultats du tableau 5.2. Elle savait le MeSH assez peu précis sur les maladies rares ; elle s'attendait à une plus forte couverture de la part de la SNOMED mais surtout les résultats avec MedDRA (thésaurus sur les effets secondaires des médicaments) ont été totalement inattendus.

Type d'alignement	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
Alignement par combinaison	2 875	1 585	MeSH
	3 201	2 705	SNOMED International
	4 919	4 919	MedDRA
	884	449	WHO-ART

TAB. 5.3 – Nombre de termes de chaque terminologie en relation alignement par combinaison

Type d'alignement	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
Alignement partiel	2 946	1 713	MeSH
	3 292	2 808	SNOMED International
	4 815	4 815	MedDRA
	851	448	WHO-ART

TAB. 5.4 – Nombre de termes de chaque terminologie en relation alignement partiel

Le tableau 5.5 donne les chiffres des termes ORPHANET en relation alignement exact avec les quatre terminologies sans passer par l'UMLS.

Comme le montre le tableau 5.6, l'application de l'algorithme d'alignement sur chaque terminologie indépendamment ne donne pas les mêmes chiffres quel'application de l'algorithme sur F_UMLS. En effet, plusieurs termes qui ne sont pas trouvés par

Nombre de termes ORPHANET	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
1 165	1 934	983	MeSH
958	1 285	1 085	SNOMED International
812	2 125	2 125	MedDRA
141	306	152	WHO-ART

TAB. 5.5 – Nombre de termes ORPHANET en correspondance en alignement exact sans utiliser l'alignementconceptuel de l'UMLS

l'alignementexact, le sont en passant par les concepts UMLS. Par exemple, le terme « WAGR syndrome » est en relation alignement exact avec le terme MeSH « syndome WAGR ». L'utilisation d'UMLS permet aussi d'aligner le terme ORPHANET avec le terme SNOMED International « syndrome de monosomie partielle 11p ».

	MeSH	SNOMED International	MedDRA	WHO-ART
Terminologie toute seule	983	1 085	2 215	152
F_UMLS	1 125 (+14%)	1 848 (+70%)	3 721 (+67%)	249 (+63%)

TAB. 5.6 – Comparaison des chiffres trouvés de l'application de l'algorithme sur chaque terminologie à part versus F_UMLS

Nous avons aussi mesuré l'apport des synonymes CISMeF ajoutés au MeSH (N = 12 293) et les concepts supplémentaires chimiques (CSC) traduits en français (substances chimiques) par l'équipe CISMeF (N = 7 200) pour l'alignementexact du MeSH. Cet apport est mesuré en comparant le nombre de termes ORPHANET alignés en appliquant l'algorithme sur les termes MeSH avec et sans synonymes CISMeF et CSC (tableau 5.7).

	MeSH	MeSH + Synonymes CISMeF + CSC
Nombre de termes ORPHANET	1 165	1 212 (+4%)

TAB. 5.7 – L'apport de l'ajout des synonymes CISMeF et les concepts supplémentaires chimiques traduits sur l'alignementexact des termes ORPHANET

La table 5.8 donne les résultats de l'évaluation de l'alignement lexical restreint au type d'alignement exact entre les termes ORPHANET et les termes de F_UMLS.

Nombre d'alignement	Qualité de l'alignement
247 (98%)	Pertinent
3 (1,2%)	Non pertinent

TAB. 5.8 – Qualité de l'alignement lexical exact entre les termes ORPHANET et les termes de F_UMLS

La tableau 5.8 montre que 98% des alignements exacts ont été évalués comme pertinents. Cependant, les trois résultats jugés comme non pertinents correspondent à des alignements en BT-NT, c'est-à-dire que le terme ORPHANET est plus générique que le terme correspondant. Par exemple, le terme ORPHANET « Cystinurie, type A » aligné vers le terme « Cystinurie ». Le dernier exemple est due au fait que « type » et « A » sont considérés comme des mots vides.

5.1.2 Comparaison entre l'alignementmanuel et l'alignemen-texact

Sur les 2 083 termes ORPHANET alignés manuellement vers au moins un code CIM10, un nombre de 593 alignements exacts est obtenu entre les termes ORPHANET et les termes MeSH. D'un autre côté nous avons (voir section 5.1.1), un nombre de 619 alignements vers au moins un terme MeSH en passant par l'alignementmanuel et en utilisant l'UMLS. Selon les résultats obtenus par chaque type d'alignement nous avons :

Premier ensemble : 327 alignements sont obtenus par l'alignementmanuel en passant par l'UMLS et non pas obtenus par l'approche lexicale.

Deuxième ensemble : 306 alignements ne sont obtenus que par l'approche lexicale.

troisième ensemble : 75 alignements différents sont obtenus par les deux approches pour les mêmes termes ORPHANET.

Quatrième ensemble : 211 mêmes alignements sont obtenus par les deux méthodes.

Le tableau 5.9 donne les résultats d'évaluations des alignements obtenus par chaque approche indépendamment (deux échantillons de 100 alignements) (alignement manuel CIM10 en passant par l'UMLS versus alignement lexical). 85 alignements obtenus par l'approche lexicale sont jugés comme pertinents. Par contre, 21 alignements seulement obtenus par l'approche manuelle + UMLS ont été jugés comme pertinents.

D'après les résultats du tableau 5.9, notre approche donne de meilleurs résultats sur cet échantillon relativement restreint (100 alignements).

	Pertinent	BT-NT	NT-BT	Frère	Non pertinent
Alignement manuel CIM10 + UMLS	21	2	32	0	45
Alignement lexical	85	0	15	0	0

TAB. 5.9 – Résultats d'évaluation des deux ensembles d'alignements obtenus par chaque approche indépendamment

Le tableau 5.10 donne les résultats d'évaluation du troisième ensemble qui correspond aux différents alignements obtenus par les deux approches pour les mêmes termes ORPHANET. Pour la première approche (manuelle), 39 alignements sont évalués comme BT-NT contre seulement 6 alignements qui sont évalués comme pertinents. Pour la deuxième approche (lexicale), 62 alignements ont été évalués comme pertinents et 8 alignements évalués comme BT-NT.

	Pertinent	BT-NT	NT-BT	Frère	Non pertinent
Approche manuelle	6	7	39	2	21
Approche lexicale	62	1	8	2	2

TAB. 5.10 – Résultats d'évaluation du troisième ensemble d'alignements (même terme ORPHANET différents termes correspondants)

Là encore l'approche lexicale semble donner de meilleurs résultats que l'approche manuelle en passant par l'UMLS.
L'évaluation du dernier ensemble qui correspond aux mêmes alignements obtenus par chaque approche donne 98% des alignements obtenus jugés comme pertinents contre seulement 2% comme non pertinents.
Le tableau 5.11 montre un exemple pour chaque type d'évaluation réalisée. Nous avons aussi extrapolé la précision des deux méthodes. Nous avons calculé le nombre d'alignements pertinents obtenus dans chaque ensemble pour les deux méthodes. Au final, l'extrapolation pour chaque méthode est le rapport des alignements pertinents obtenus sur le nombre total des alignements. Pour la méthode d'alignement manuelle CIM10 en passant par l'UMLS l'extrapolation de la précision est de 46%. Cependant, elle est beaucoup plus importante pour la méthode lexicale 81%.

Alignement structurel hiérarchique

Alignement BT (Broader Than Matching) : sur les 5 753 termes ORPHANET qui n'ont pas pu être définis par l'alignementexact, un nombre de 4 672 (62%) est en alignement BT avec au moins un terme de F_UMLS couvrants 857 concepts UMLS. Nous distinguons 7 niveaux hiérarchiques d'alignement entre termes ORPHANET et termes de F_UMLS. Le tableau 5.12 donne le nombre d'alignements

Type d'évaluation	Terme ORPHANET	Terme MeSH correspondant
Pertinent	Nocardiose	Infection à nocardia
BT-NT	Hémophilie	Hémophilie A
NT-BT	Dystrophie musculaire de Duchenne et Becker	myopathie de Duchenne
Frère	Cryptophtalmie isolée	microphtalmie
Non pertinent	Anomalie de développement sexuel	Pseudohermaphrodisme

TAB. 5.11 – Exemple de chaque type d'évaluation réalisé

en BT suivant chaque niveau hiérarchique.

Niveau hiérarchique	**Nombre d'alignements**
Niveau 1	1 555 (20,9%)
Niveau 2	951 (12,8%)
Niveau 3	1 103 (15,23%)
Niveau 4	859 (11,5%)
Niveau 5	161 (2,1%)
Niveau 6	39 (0,5%)
Niveau 7	4 (0,0005%)

TAB. 5.12 – Nombre de termes ORPHANET en alignement BT pour chaque niveau hiérarchique

Le tableau 5.13 donne le nombre de termes de chaque terminologie en relation d'alignement BT avec les 4 672 termes ORPHANET. Le tableau 5.14 donne

Type d'alignement	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
Alignement BT	679	332	MeSH
	603	531	SNOMED International
	1 189	1 189	MedDRA
	163	91	WHO-ART

TAB. 5.13 – Nombre de termes de chaque terminologie en relation alignement BT

les résultats de l'évaluation des alignements BT entre les termes ORPHANET (N=500) et les termes de F_UMLS.

Sur les 500 alignements BT évalués, 482 alignements ont été jugés comme des alignements BT, c'est-à-dire que le terme ORPHANET correspond effectivement

Nombre d'alignement	Qualité de l'alignement
482	Pertinent
2	Exact
16	Non pertinent

TAB. 5.14 – Qualité de l'alignement BT entre les termes ORPHANET et les termes de F_UMLS

à un fils du terme aligné. Les résultats de l'évaluation montrent aussi que deux alignements ont été évalués comme exacts : l'alignementdu terme ORPHANET « Déficit intellectuel - hypsarrhytmie » avec le terme « syndrome de West » ou l'alignementdu terme « Contractures du pied - atrophie musculaire - apraxie oculomotrice » avec le terme « Wieacker-Wolff, syndrome de ».

Alignement NT (Narrower Than Matching) : sur les 5 753 termes ORPHANET qui ne sont pas en alignement exact, un nombre de 734 (9% du nombre total des termes ORPHANET) sont en alignement NT avec au moins un terme de F_UMLS couvrants 2 359 concepts UMLS. Nous distinguons quatre niveaux hiérarchiques d'alignement entre termes ORPHANET et termes de F_UMLS. Le tableau 5.15 donne le nombre d'alignements en NT suivant chaque niveau hiérarchique.

Niveau hiérarchique	**Nombre d'alignements**	**Nombre de termes ORPHANET non alignés en BT**
Niveau 1	613 (8,2%)	45
Niveau 2	108 (1,45%)	4
Niveau 3	12 (0%)	0
Niveau 4	1 (0%)	0

TAB. 5.15 – Nombre de termes ORPHANET en alignement NT pour chaque niveau hiérarchique

Le tableau 5.16 donne le nombre de termes de chaque terminologies en relation d'alignement NT avec les 734 termes ORPHANET.

Type d'alignement	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
Alignement NT	1 783	931	MeSH
	1 764	1559	SNOMED International
	3 122	3 122	MedDRA
	328	207	WHO-ART

TAB. 5.16 – Nombre de termes de chaque terminologie en relation alignement NT

Le tableau 5.17 donne les résultats de l'évaluation des alignements NT entre les termes ORPHANET (N=100) et les termes de F_UMLS.

Nombre d'alignements	Qualité de l'alignement
87	Pertinent
1	Exact
12	Non pertinent

TAB. 5.17 – Qualité de l'alignement NT entre les termes ORPHANET et les termes de F_UMLS

Sur les 100 alignements NT évalués, 87 alignements ont été jugés comme des alignements NT, c'est-à-dire que le terme ORPHANET correspond effectivement à un parent du terme aligné. Les résultats de l'évaluation montrent aussi qu'un seul alignement a été évalué comme exact : l'alignementdu terme ORPHANET « Anomalies congénitales multiples/syndrome dysmorphique » avec le terme « anomalies congenitales multiples ».

Cependant, une chose très importante est que plus les alignements structurels sont effectués avec un niveau important, plus leur pertinence « métier » diminue. En poussant le raisonnement jusqu'au bout on pourrait retrouver pour chaque terme ORPHANET qu'il mappe avec le terme MeSH « Maladie ». Cette information n'a aucune valeur pour les utilisateurs du thésaurus ORPHANET.

5.2 Alignement de la classification ATC avec les terminologies francophones

5.2.1 Résultats

Alignement lexical fondé sur les outils en français

Sur les 4 268 codes ATC de niveau 5, un nombre de 2 992 codes ATC sont en relation d'alignement exact avec un moins un terme de F_UMLS couvrant 8 697 concepts. D'autre part, 668 codes ATC sont en relation alignement par combinaison avec au moins deux termes de F_UMLS couvrant 2 626 concepts UMLS. Le tableau 5.18 montre le nombre de codes ATC en relation et le nombre de concepts UMLS couverts suivant chaque type d'alignement utilisé.

Nombre de codes ATC	**Nombre de concepts UMLS couverts**	**Type d'alignement**
2 992 (70%)	8 546	Alignement exact
668 (15%)	2 626	Alignement par combinaison
350 (8%)	675	Alignement partiel

TAB. 5.18 – Nombre de codes ATC en correspondance pour chaque type d'alignement

Les tableaux 5.19, 5.20 et 5.21 montrent le nombre de termes de chaque terminologie qui sont en relation avec les codes ATC pour chaque type d'alignement.

Type d'alignement	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
Alignement exact	8 454	2 499	MeSH
	1 839	1 728	SNOMED International
	81	81	MedDRA
	18	18	CIM10
	4	4	WHO-ART

TAB. 5.19 – Nombre de termes de chaque terminologie en relation alignement exact

Type d'alignement	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
Alignement par combinaison	2 489	614	MeSH
	827	746	SNOMED International
	218	218	MedDRA
	25	15	WHO-ART

TAB. 5.20 – Nombre de termes de chaque terminologie en relation alignement par combinaison

Type d'alignement	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
Alignement partiel	566	207	MeSH
	306	256	SNOMED International
	138	138	MedDRA

TAB. 5.21 – Nombre de termes de chaque terminologie en relation alignement partiel

Le tableau 5.22 donne les chiffres des codes ATC en relation alignement exact avec les termes des quatre terminologies sans passer par l'UMLS.

Nombre de codes ATC	**Nombre de termes**	**Nombre de termes préférés**	**Terminologie**
1 566	7 112	1 175	MeSH
1 328	1 196	1 108	SNOMED Int
51	52	52	MedDRA

TAB. 5.22 – Nombre de codes ATC en correspondance et nombre des termes couverts en alignement exact sans utiliser l'alignementconceptuel de l'UMLS

Le tableau 5.23 montre que l'application de l'algorithme d'alignement sur chaque terminologie indépendamment ne donne pas les mêmes chiffres comparant à l'application de l'algorithme sur F_UMLS. Comme dans le cas de l'alignement du thésaurus ORPHANET avec F_UMLS, plusieurs termes qui ne sont pas trouvés par l'alignementexact, le sont en passant par les concepts UMLS. Nous avons aussi mesuré l'apport des synonymes CISMeF ajoutés au MeSH et les concepts supplémentaires chimiques traduits en français par l'équipe CISMeF pour l'alignementexact du MeSH. Cette apport est mesuré en comparant le nombre de codes ATC alignés en appliquant l'algorithme sur les termes MeSH avec et sans synonymes CISMeF et CSC (tableau 5.24).

La plus value des synonymes et des CSC est très importante pour l'ATC (+85%). Dans la section 5.1.1, que cet apport est beaucoup moins important : ce qui revient à dire que beaucoup de traitements sont « termino-dépendants ».

	MeSH	SNOMED Int	MedDRA	WHO-ART
Terminologie toute seule	1 175	1 108	52	2
F_UMLS	2 499 (+112%)	1 788 (+60 %)	81 (+55%)	4 (+100%)

TAB. 5.23 – Comparaison des chiffres trouvés de l'application de l'algorithme sur chaque terminologie à part versus F_UMLS

	MeSH	MeSH + Synonymes CISMeF + CSC
Nombre de codes ATC	1 566	2 898 (+85%)

TAB. 5.24 – L'apport de l'ajout des synonymes CISMeF et les concepts supplémentaires chimiques traduits sur l'alignementexact du MeSH

Alignement lexical orienté anglais (MetaMap)

Sur les 4 268 codes ATC de niveau 5, 3 170 sont en relation alignement exact avec au moins un terme de l'UMLS alors que 3 062 sont en relation alignement exact avec au moins un terme de F_UMLS. Le tableau 5.25 et 5.26 montrent le nombre de codes ATC en relation suivant chaque type d'alignement utilisé avec les termes de UMLS et F_UMLS respectivement.

Nombre de codes ATC	**Type d'alignement**
3 170 (74%)	Alignement exact
664 (16%)	Alignement par combinaison
291 (7%)	Alignement partiel

TAB. 5.25 – Nombre de codes ATC en correspondance pour chaque type d'alignement avec les termes de l'UMLS en anglais avec MetaMap

Nombre de codes ATC	**Type d'alignement**
3 062 (72%)	Alignement exact
371 (9%)	Alignement par Combinaison
567 (13%)	Alignement Partiel

TAB. 5.26 – Nombre de codes ATC en correspondance pour chaque type d'alignement avec les termes de F_UMLS en anglais avec MetaMap

5.2.2 Comparaison entre les deux méthodes d'alignement exact français et anglais

Sur les 2 992 alignements exacts entre les codes ATC et les termes de F_UMLS obtenus par la méthode d'alignement fondée sur le français et les 3 062 alignements exacts obtenus par la méthode fondée sur l'anglais, un nombre de 1 298 alignements en commun sont obtenus par les deux méthodes. Selon les résultats d'alignements obtenus par chaque méthode nous avons :

Premier ensemble : 1 298 alignements en commun sont obtenus par les deux méthodes d'alignement.

Deuxième ensemble : 342 alignements différents obtenus par les deux méthodes (136 alignements par la méthode fondée sur le français et 205 par la méthode utilisant MetaMap)

Troisième ensemble : 1 558 alignements différents sont obtenus par les deux approches pour les mêmes codes ATC. Nous remarquons aussi que dans le 1 558 alignements il existe 1 458 (93%) alignements avec au moins un terme correspondant en commun (Concept UMLS). Par exemple, pour le libellé ATC « sodium monofluorophosphate » (code ATC : A01AA02) MetaMap propose un seul terme correspondant « sodium monofluorophosphate », alors que la méthode fondée sur le français propose un autre terme (concept UMLS) que celui proposé par MetMAP qui est « fluorophosphate ».

Évaluation des deux approches

Sur les 2 898 alignements exacts (limités aux concepts UMLS contenant au moins un code MeSH) obtenus par la méthode fondée sur le français nous avons :

1. 2 582 (89%) alignements correspondent à des alignements validés (l'intersection avec l'ensemble des alignements validés est non vide),
2. 316 (11%) alignements ont été obtenus seulement automatiquement,
3. 389 alignements valides n'ont pas été obtenus automatiquement.

Sur les 3 052 alignements exacts (limités aux concepts UMLS contenant au moins un code MeSH) obtenus par l'outil MetaMap nous avons :

1. 2 695 (88%) alignements correspondent à des alignements validés (intersection avec l'ensemble des alignements validés est non vide),
2. 357 (11%) alignements ont été obtenus seulement automatiquement,
3. 276 alignements valides n'ont pas été obtenus automatiquement.

L'union des alignements valides obtenus par les deux méthodes donne un nombre de 2 798 (65% de tous les codes ATC de 5^{e}niveau) codes ATC alignés vers MeSH (concepts UMLS) validés. Parmi ces codes nous avons : 2 479 codes qui sont communs aux deux méthodes, 216 codes obtenus uniquement par MetaMap et 103 obtenus uniquement par la méthode en français. De plus, l'union des codes ATC alignés seulement en automatique est égale à 370 codes et l'union des codes ATC alignés seulement en manuel est égale à 492.

Le nombre d'alignements obtenu par l'outil MetaMap est un peu plus grand que le nombre obtenu par les outils en français (3 170 versus 2 992). Cependant, la précision des alignements pour les deux méthodes est identique et élevée (88% versus 89%) sur les termes MeSH. De plus, les codes ATC alignés seulement par les méthodes automatiques correspondent généralement aux codes ATC alignés manuellement vers au moins deux termes MeSH. Par exemple, le code ATC « hydrogène peroxyde » (code ATC : « A01AB02 ») est aligné automatiquement vers le terme MeSH : « peroxyde d'hydrogène ». Contrairement à l'alignement manuel qui lui fait correspondre deux termes MeSH : « maladie de la bouche/traitement médicamenteux » et « peroxyde d'hydrogène ». Concernant les codes ATC alignés seulement en manuel (N = 492), la plupart des termes correspondants représentent des substance chimiques ou des noms de médicaments (voir tableau 5.27), ainsi, c'est impossible pour ces cas précis pour les méthodes lexicales (anglais ou français) de trouver des correspondances.

Libellé ATC	**Terme MeSH correspondant**
mifamurtide (L03AX15)	muramylNAc-Ala-isoGln-Lys-tripeptide-phosphatidylethanolamine (CSC)
saproptérine (A16AX07)	5,6,7,8-tétrahydrodictyoptérine (CSC)

TAB. 5.27 – Exemples de codes ATC alignés seulement en manuel vers MeSH

5.3 Alignement de la classification CCAM avec les terminologies francophones

5.3.1 Résultats

Alignement lexical fondé sur les outils en français

Sur les 7 926 codes de la CCAM, nous avons 5 212 (65%) qui sont en alignement avec au moins un terme en français de l'UMLS. Le tableau 5.28 montre le nombre de codes CCAM en alignement suivant chaque type d'alignement.

Type d'alignement	**Nombre d'alignements**
Alignement exact	200 (2,5%)
Alignement par combinaison	2 010 (25%)
Alignement Partiel	3 002 (37,8%)

TAB. 5.28 – Nombre d'alignements suivant chaque type d'alignement

Sur les 5 212 alignements trouvés, un nombre de 2 210 (43%) alignements est effectué sur les deux axes des codes CCAM (topographique et action). D'un autre côté,
1 716 (32%) alignements sont effectués que sur l'axe topographique alors que 1 286 (25%) alignements le sont sur l'axe action.
Sur les 194 descripteurs utilisés pour décrire l'axe topographique, 127 (65%) sont alignés avec F_UMLS. D'un autre côté, sur les 331 descripteurs utilisés pour décrire l'axe des actions, 123 (37%) sont alignés avec F_UMLS.

Alignement lexical fondé sur MetaMap

Sur les 7 926 codes de la CCAM, nous avons 5 909 (74%) qui sont en alignement avec au moins un terme de UMLS. Sur les 5 909 codes de la CCAM, 2 100 (35%) codes ont été alignés avec MetaMap sur les deux axes topographiques et anatomiques. De plus, 1 314 (23%) d'alignements sont effectués que sur l'axe topographique et 2 495 (42%) alignements sont effectués que sur l'axe action.
Sur les 194 descripteurs utilisés pour décrire l'axe topographique, 96 (49%) sont alignés vers UMLS. D'un autre côté, sur les 331 descripteurs utilisés pour décrire l'axe des actions, 205 (62%) sont alignés vers UMLS.

Sur les 2 210 alignements exacts sur les deux axes entre les codes CCAM et les termes de F_UMLS obtenus par la méthode d'alignement fondée sur le français et les 2 100 alignements exacts sur les deux axes obtenus par l'outil MetaMap, un nombre de 620 alignements en commun (seulement) est obtenu par les deux méthodes. Pour tous les alignements exacts sur les deux axes, les alignements sont parcellaires, même si projeter sur les deux axes d'autres termes de la CCAM ne sont pas alignés. Les deux méthodes diffèrent aussi dans le nombre d'axes alignés vers UMLS. La méthode fondée sur les outils lexicaux en français, aligne beaucoup plus de descripteurs de l'axe topographique vers UMLS que la méthode fondée sur l'outils MetaMap (65% versus 49%). Le nombre très faible de descripteurs de l'axe topographique alignés par l'outil MetaMap, s'explique par le fait que la plupart de ces descripteurs ont plus de deux mots. Par exemple, en utilisant MetaMap le terme : « Lips, tongue, oral cavity as a whole » (œil, sans précision) n'est aligné vers aucun terme dans UMLS, contrairement à la méthode lexicale en français qui lui fait correspond les deux termes en français : « kyste et lèvre ».
À l'inverse, la méthode utilisant l'outil MetaMap qui aligne beaucoup plus de descripteurs de l'axe action vers UMLS que la méthode fondée sur les outils en français (62% versus 37%).

5.3.2 Évaluation de l'alignement lexical fondé sur les outils en français

Pour tous les alignement exacts (n=200), 182 (91%) des alignements entre les codes CCAM et les concepts de F_UMLS ont été évalués comme NT-BT et seulement 9 alignements ont été évalués comme équivalent (voir table 5.29).

Équivalent	**BT-NT**	**NT-BT**	**Incomplet**	**Non-pertinent**	**Total**
9 (4,5%)	0 (0%)	182 (91%)	3 (1,5%)	6 (3%)	200

TAB. 5.29 – Résultats d'évaluations pour l' « alignement exact »

Pour l'ensemble des « alignements par combinaison »(n=100), 61 et 44 des axes anatomiques et actions sont respectivement équivalents à au moins un concept UMLS. Pour ce type d'alignement, 27% des alignements entre les codes CCAM et au moins un concept UMLS ont été évalués comme exactement équivalents, contre seulement 54 alignements qui ont été évalués comme NT-BT (voir tableau 5.30). Les deux évaluations

Alignement par combinaison	**Équivalent**	**BT-NT**	**NT-BT**	**Incomplet**	**Non-pertinent**
Anatomique	61	1	29	9	0
Action	44	0	49	1	6
Code CCAM	27	0	54	10	9

TAB. 5.30 – Résultats d'évaluations pour l'« alignement par combinaison »(n=100)

effectuées sur les deux alignement donnent une première indication sur la qualité des alignements. En effet, que ce soit pour l'alignement exact ou l'alignement par combinaison, la plupart des alignements ont été évalués comme NT-BT (narrower than). Ces résultats ne sont pas étonnants vu le niveau de généralité des descripteurs représentant les codes CCAM. Ainsi, le passage du libellé CCAM vers la nouvelle représentation sur les deux axes topographique et action implique forcément une perte de précision par rapport à l'acte original. De plus, dans certains cas, avec la nouvelle représentation nous perdons plusieurs notions présentes dans le libellé original de l'acte. Par exemple, l'acte « Lithotritie extracorporelle de la vessie » correspondant au code CCAM : JDNM001. Une décomposition suivant le code produit deux termes : « Vessie » et « lithotritie ». Ainsi, nous perdons la notion d'« extracorporelle » présente dans le libellé original.

5.4 Synthèse

Dans cette partie nous avons présenté les différents résultats obtenus par nos méthodes sur les trois alignement réalisés. Les résultats obtenus dans les méthodes utilisées dépendent des terminologies sources (à aligner). Cependant, la qualité des alignements peut différer d'un objectif à un autre et suivant les terminologies utilisées et suivant le

contexte. Ces résultats seront discutés en détail dans le chapitre 8.

Chapitre 6

Projection des relations SNOMED CT entre plusieurs terminologies (Inter et Intra)

Dans ce chapitre nous proposons une méthode d'interopérabilité entre terminologies fondée sur UMLS afin de projeter de façon automatique les relations de la terminologie SNOMED CT sur trois terminologies francophones. Cette méthode va permettre de lier différentes terminologies (CIM10, SNMI et MeSH) avec des relations issues d'une autre terminologie (SNOMED CT).

L'idée de ce travail Merabti *et al.* (2009a) est de projeter des relations d'une terminologie sémantiquement riche vers une terminologie plus pauvre sémantiquement. Les concepts à l'intérieur de la SNOMED CT Spackman (2000) sont de deux types : « primitif » ou « complètement défini ». Les concepts « Complètement définis » représentent tous les concepts qui peuvent être différenciés de leurs concepts parents et frères en vertu des relations avec d'autres concepts. Autrement, tous les concepts sont « primitifs ». Il existe 261 264 (84%) concepts primitifs dans la SNOMED CT et 50 049 (16%) concepts complètements définis. De plus, ces concepts sont organisés en différentes classes. Le tableau 6.1 donne le nombre et le pourcentage des concepts dans les 14 classes les plus représentées dans la SNOMED.

SNOMED CT fournit des définitions formelles à ces concepts en utilisant les relations « is-a » et des relations attribuées. Ces relations visent principalement trois objectifs : expliciter la sémantique, automatiser la classification et permettre la post-cordination. Premièrement, les relations reflètent formellement la sémantique du concept. En effet, SNOMED CT ne donne pas une définition textuelle des concepts, mais elle

Classe	**Nombre des terme préférés**	**% des termes préférés dans la SNOMED CT**
Maladie	74 993	24,0
Procédure	50 253	16,1
Localisation	32 630	10,5
Organisme	27 643	8,9
Région du corps	25 478	8,2
Substance	22 767	7,3
Produits	18 530	6,0
Valeur qualifiée	8 583	2,8
Événement	8 415	2,7
Entité observable	7 749	2,5
Situation	4 863	1,6
Anomalie morphologique	4 746	1,5
Objet Physique	4 489	1,4
Occupation	4 084	1,3

Tab. 6.1 – Le nombre et le pourcentage des concepts par classe dans la SNOMED CT

vise plutôt à spécifier de manière formelle les propriétés du concept. Par exemple, dans la SNOMED CT, « Pneumonie » est définie comme une « maladie du poumon ». SNOMED CT est à ce jour, la terminologie de santé la plus détaillée pour décrire un dossier médical électronique. Un total de 61 relations est défini dans la SNOMED CT. Ces relations peuvent être classées en quatre type de relations :

- Les **caractéristiques définies** : l'ensemble relation « ISA » + attributs définis est considéré comme les « caractéristiques définies ». Les attributs définis relient deux concepts et établissent le type de relation entre eux. Elles sont considérées comme des « caractéristiques définies » car elles représentent, d'une façon formelle, la définition d'un concept en le liant avec d'autres concepts. La définition logique d'un concept dans la SNOMED CT inclut un ou plusieurs « concepts hiérarchiques » (modélisés avec la relation « ISA ») et un ensemble d'attributs définis pour expliciter la sémantique du concept et aider à le différencier des autres définitions possédant les mêmes concepts hiérarchiques. Par exemple, le concept « Pneumonie » est défini par :
  - ISA « Maladie des Poumons » ;
  - Finding_Site_of (Localisation) « Poumons » ;
  - Associated Morphology (Morphologiquement associé à) « Inflammation » ;
- Les **caractéristiques qualifiées** : elles sont utilisées pour créer des concepts plus complexes (post-cordination) comme, par exemple, les relations « severity » (gravité) et « laterality » (latéralité). Par exemple, le concept « bronchite aiguë » peut

être post-cordonné en utilisant le concept « bronchite » et l'attribut qualifié « Clinical_Course » avec la valeur « aiguë ».

- Les **relations historiques**: elles relient des concepts inactifs à des concepts actifs.
- Les **relations supplémentaires** (« Additional » en anglais) : elles ne sont pas définies mais retenues pour être compatibles avec SNOMED RT (Reference Terminology)[1] Spackman *et al.* (1993). Les quatre types de relations définies précédemment sont représentés dans les 61 relations.

Un terme préféré est un terme choisi pour représenter un concept au sein d'une terminologie ; par exemple, dans SNOMED International le terme « achondroplasie » est le terme préféré pour représenter la classe des termes « achondroplasie de Parrot et Marie », « achondrodystrophie foetale », « maladie de Parrot », « nanisme achondroplasique ». Les autres termes sont des synonymes du terme préféré.

Schéma d'interopérabilité pour la projection des relations SNOMED CT : Exemple de projection entre CIM10 et SNMI

La première étape de notre étude consiste à extraire tous les concepts UMLS liés au moins par une relation SNOMED CT. Par exemple, les deux concepts UMLS C0004099 et C0004096 sont rattachés respectivement aux deux termes préférentiels SNOMED CT « asthme à l'effort » et « asthme » lesquels sont liés suivant la relation SNOMED CT « ISA ». Le tableau 6.2 représente les 10 relations SNOMED CT les plus représentées dans l'UMLS. La deuxième étape consiste à projeter les relations SNOMED CT vers les termes préférés de CIM10 et de SNOMED International, cela consiste à projeter les couples de concepts UMLS trouvés dans l'étape précédente vers les deux terminologies CIM10 et SNOMED International en filtrant uniquement sur les termes préférentiels. Au final, les couples de concepts UMLS sont remplacés par des couples de codes des deux terminologies liées par des relations SNOMED CT (Figure 6.1).

Formellement, le schéma d'interopérabilité limité aux terminologies SNOMED 3.5 et CIM10 est défini ainsi :
Supposons que nous ayons quatre termes SNOMED CT A, B, C et D. Ces termes sont

[1]Ce projet est issu d'une collaboration entre le College of American Pathologists, la société Kaiser Permanente (Health Management Organization) et la Mayo Clinic

Top 10 des relations SNOMED CT	**Nombre de couples en relations**
ISA	496 784
Finding_Site_of (Localisation)	86 358
Associated morphology (Morphologiquement associé à)	80 036
Method_of (Méthode de)	54 107
Part_of (Partie de)	47 810
Compenent_of (Composant de)	9 135
Direct_procedure_site_of (Procédure directe sur site de)	34 002
Causative_agent_of (Agent causal de)	23 628
Indirect_procedure_site_of (Procédure indirecte sur site de)	8 925
Finding_method_of (Localisation de méthode de)	8 419

TAB. 6.2 – Les 10 relations SNOMED CT les plus représentées dans l'UMLS

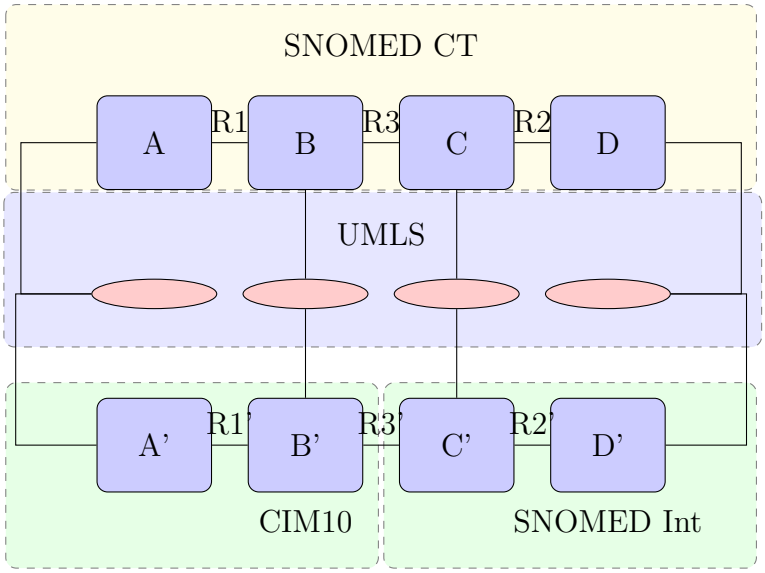

FIG. 6.1 – Schéma d'interopérabilité liant termes CIM10 et SNOMED International par des relations SNOMED CT

relié s par les relations $R1$, $R2$ et $R3$ tel que:

$$A\ R1\ B$$
$$C\ R2\ D$$
$$B\ R3\ C$$

– S'il existe deux termes CIM10 A' et B' tels que A' et B' correspondent respecti-

vement aux termes A et B de SNOMED CT,

- S'il existe deux termes SNOMED International C' et D' tels que C' et D' correspondent respectivement aux termes C et D de SNOMED CT,
- Le fait qu'une relation SNOMED CT $R1$ existe entre A et B implique que cette relation soit projetée d'une manière automatique entre les termes A' et B' de la CIM10. De la même manière, l'existence d'une relation SNOMED CT $R2$ entre les termes C et D implique automatiquement une projection de cette relation entre les termes C' et D' de SNOMED International.

L'extensibilité de schéma permet une interopérabilité intra-terminologies entre termes d'une même terminologie et une projection entre termes de différentes terminologies, on parle alors de relation inter-terminologies.

- Donc, s'il existe une autre relation SNOMED CT $R3$ entre B et C, la relation R3 sera projetée pour relier les deux termes B' et C' qui sont respectivement deux termes CIM10 et SNOMED International.

Nous avons aussi projeté les relations SNOMED CT entre les termes de la terminologie MeSH Merabti *et al.* (2009b). L'utilisation de la terminologie MeSH permettra de faire en plus de l'étude quantitative, une étude qualitative. Dans cette dernière, nous mesurons la qualité des relations SNOMED CT projetées entre les termes MeSH d'un point de vue documentaliste. C'est-à-dire avec une finalité d'indexation.
Pour projeter les relations SNOMED CT entre les termes MeSH, nous avons repris le même schéma d'interopérabilité de la figure 6.1, appliqué à une seule terminologie comme montré dans la figure 6.2.

L'étude qualitative a été appliquée sur les quatre relations SNOMED CT les plus fréquentes entre les termes MeSH : ISA, Finding_Site_of (Localisation), Causative_agent_of (Agent Causal de) et Associated morphology (Morphologiquement associé à) (voir la section 7.2 des résultats). Pour l'évaluation, nous avons préparé tous les couples de termes préférés en relations suivant les quatre relations. Pour les quatre ensembles obtenus, nous avons choisi aléatoirement 100 couples de termes. Chaque ensemble a été évalué par un expert documentaliste de l'équipe CISMeF spécialisé dans la terminologie MeSH (Catherine Letord). L'évaluation a été réalisée sur trois échelles de qualité :

- Pertinent : si la relation entre les deux termes MeSH est jugée pertinente. Du point de vue MeSH et du point de vue documentaliste.
- Moyen : si la relation entre les deux termes n'est pas mauvaise mais elle n'est pas parfaite.
- Mauvais : si la relation entre les deux termes est mauvaise.

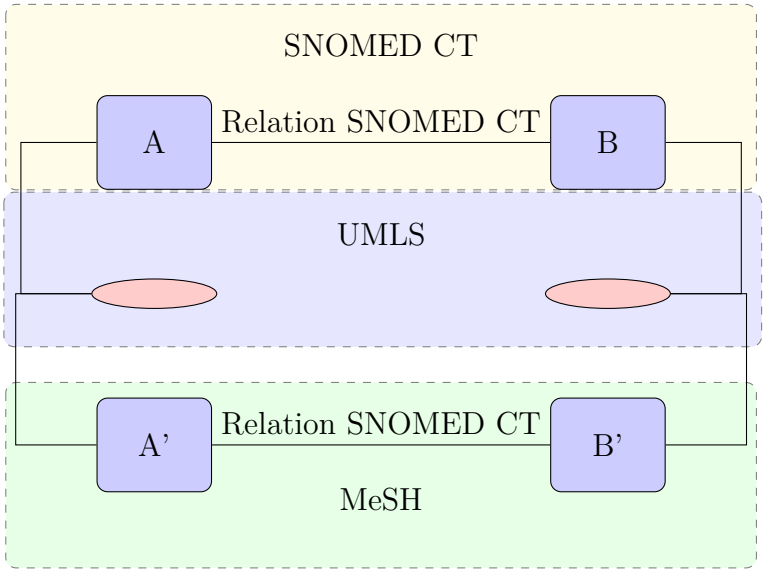

FIG. 6.2 – Schéma d'interopérabilité liant des termes MeSH par des relations SNOMED CT

Cependant, nous avons réalisé un traitement spécifique pour les couples de termes en relation « ISA » dans le but d'éliminer les relations hiérarchiques explicites qui existe dans la terminologie MeSH. Par exemple, une relation SNOMED CT « ISA » a été trouvée entre les deux termes « maladie des bronches » et « asthme à l'effort ». Cependant, dans la terminologie MeSH il n'existe pas de relation hiérarchique directe entre les deux termes. Par contre, il existe une relation hiérarchique explicite entre les deux termes, le terme « maladie des bronches » est parent du terme « asthme » qui est lui aussi parent du terme « asthme à l'effort ». Ceprendant, ce type d'évaluation doit être refait par un ou plusieurs experts professionnels de santé si l'objectif est le domaine des dossiers médicaux électronique.

Synthèse

Dans cette partie nous avons présenté une méthode d'interopérabilité entre terminologies fondée sur UMLS afin de projeter les relations de la terminologie SNOMED CT entre trois terminologies francophones. Nous avons vu que cette méthode a permis de lier différentes terminologies (CIM10, SNMI et MeSH) avec des relations issues d'une autre terminologie (SNOMED CT). Dans le chapitre suivant nous présentons les différents résultats obtenus ainsi qu'une évaluation qualitative sur la terminologies MeSH.

Chapitre 7

Résultats et évaluations : projection des relations SNOMED CT vers d'autre terminologies

Dans ce chapitre nous présentons les différents résultats obtenus concernant la projection des relations SNOMED CT vers les trois terminologies CIM10, SNMI et MeSH. Nous termions ce chapitre par une évaluation des relations SNOMED CT projetées sur la terminologie MeSH.

7.1 Projection des relations SNOMED CT entre les terminologies CIM10 et SNOMED 3.5

Un total de 1 051 085 termes de SNOMED CT est inclus dans l'UMLS, avec un nombre de 308 893 termes préférentiels. 2 437 839 couples de termes préférentiels sont en relation avec au moins une relation SNOMED CT, par conséquent chaque terme préférentiel SNOMED CT est lié à, au moins, un autre terme préférentiel SNOMED CT. De plus, 2 867 568 couples de termes non préférentiels sont en relation via, au moins, une relation SNOMED CT.
Le tableau 7.1 montre le nombre de termes préférentiels SNOMED International et CIM10 qui partagent, au moins, un concept UMLS avec un terme préférentiel SNOMED CT. Pour ces deux terminologies (CIM10 et SNOMED International), 91% de SNOMED International et 85% de CIM10 ont un équivalent dans SNOMED CT.

Terminologies	Nombre de termes préférentiels	Nombre de termes préférentiels dans SNOMED CT
SNOMED International	107 900	97 080 (91%)
CIM10	9 308	7933 (85%)

TAB. 7.1 – Le nombre des termes préférentiels de SNOMED International et de CIM10 dans la SNOMED CT

Au sein de SNOMED CT, les 136 relations sont orientées, donnant lieu à 68 couples de relations : par exemple « la relation ISA » est représentée par deux relations symétriques dans SNOMED CT, « ISA » et « inverse ISA ». Nous ne considérons, dans cette étude, que les 68 relations directes en représentant seulement les relations à caractéristiques définies (N=50). En plus de ces relations, nous représentons aussi la relation supplémentaire « Part_Of » (Partie de), très utile en anatomie.

Projection des relations SNOMED CT vers SNOMED International

Un total de 183 726 couples de termes préférentiels SNOMED International est en relation par, au moins, une relation SNOMED CT. Le nombre de couples de termes SNOMED International en relation est distribué d'une façon non-uniforme suivant le type de relation ; ainsi, nous avons 93 221 couples de termes préférentiels SNOMED International qui sont en relation, via la relation SNOMED CT « ISA » et seulement 1401 couples de termes préférentiels SNOMED International qui sont en relation, suivant la relation SNOMED CT « Part_Of ». Le tableau 7.2 montre les 10 premières relations SNOMED CT (en nombre) projetées vers les couples de termes SNOMED International. Nous avons 74% des termes préférentiels de SNOMED International qui sont représentés dans les 183 726 couples de termes liés par, au moins, une relation SNOMED CT.

Projection des relations SNOMED CT vers CIM10

Comme présenté dans la section précédente pour SNOMED International, un total de 5 890 couples de termes préférentiels CIM10 est en relation avec, au moins, une relation SNOMED CT. De même pour CIM10, le nombre de couples de termes CIM10 en relation est distribué d'une façon non-uniforme suivant le type de relation. Ainsi, nous avons 5 019 couples de termes préférentiels CIM10 qui sont en relation, via la relation SNOMED CT « ISA »; par contre, nous n'avons aucun terme en relation, via les deux relations « Method_of » et « Part_of ».

Relation SNOMED CT	**Nombre de couples de termes préférentiels de SNOMED International**
ISA	93 221
Finding_Site_of (Localisation)	24 661
Associated_Morphology_of (Morphologiquement associé à)	18 760
Direct_Procedure_Site_of (Procédure directe sur site de)	11 077
Method_of (Méthode de)	7 362
Causative_agent_of (Agent causal)	6 062
Indirect_procedure_site_of (Procédure indirecte sur site de)	2 883
Component_of (Composant de)	2 678
Direct_morphology_of (Morphologie directe de)	2 421
Active_ingredient_of (Ingrédient actif de)	2 097
Interprets (Interpréter)	2 090
Procedure_site_of (Procédure sur site de)	1 508
Part_of (Partie de)	1 401

TAB. 7.2 – Les 10 premières relations SNOMED CT projetées entre les termes de SNOMED International et le nombre de couples de termes préférentiels SNOMED international

Le tableau 7.3 montre les principales relations SNOMED CT projetées vers les couples de termes CIM10 ; 48% des termes préférentiels de CIM10 sont représentés dans les 5890 couples de termes préférentiels, liés par, au moins, une relation SNOMED CT.

Relation SNOMED CT	**Nombre de couples de terme préférentiels de CIM10**
ISA	5019
Associated_Morphology_of (Morphologiquement associé à)	834
Definitional manifestation (Manifestation définie)	109
Associated with (Associé à)	74
Due to (Causé par)	49

TAB. 7.3 – Les principales relations SNOMED CT projetées entre les termes CIM10

Projection des relations SNOMED CT vers des couples de termes CIM10 et SNOMED International

L'extensibilité de notre schéma permet de projeter les relations SNOMED CT entre deux couples de terminologies différentes, ainsi en appliquant ce schéma sur nos deux

terminologies de référence CIM10 et SNOMED International, 33 097 couples de termes CIM10 et SNOMED International sont en relation via au moins une relation SNOMED CT. Le tableau 7.4 expose les principales relations SNOMED CT projetées entre au moins un terme CIM10 et un terme SNOMED International, 6 242 des termes préférentiels de CIM10 (67%) et 14 276 des termes préférentiels de SNOMED International (13%) sont liés via au moins une relation SNOMED CT.

Relation SNOMED CT	**Nombre de couples CIM10 et SNOMED International en relation**
ISA	17 780
Associated_Morphology_of (Morphologiquement associé à)	6 518
Finding_Site_of (Localisation)	5 243
Causative agent_of (Agent causal)	1 258
Associated with (Associé à)	218
Direct_morphology_of (Morphologie directe de)	204

TAB. 7.4 – Les principales relations SNOMED CT projetées entre termes SNOMED International et CIM10

L'exemple de la figure 7.1 démontre bien la création d'une relation SNOMED CT inter-terminologique entre un terme CIM10 et deux termes SNOMED International. La figure montre l'exemple du terme CIM10 « achondroplasie » lié d'une part avec le terme SNOMED International « dysplasie congénitale » suivant la relation SNOMED CT « associated morphology » (Morphologiquement associé à) et, d'autre part, avec le terme SNOMED International « Os » via la relation SNOMED CT « Finding Site of » (Localisation). Les résultats de cette étude devraient permettre, entre autre,

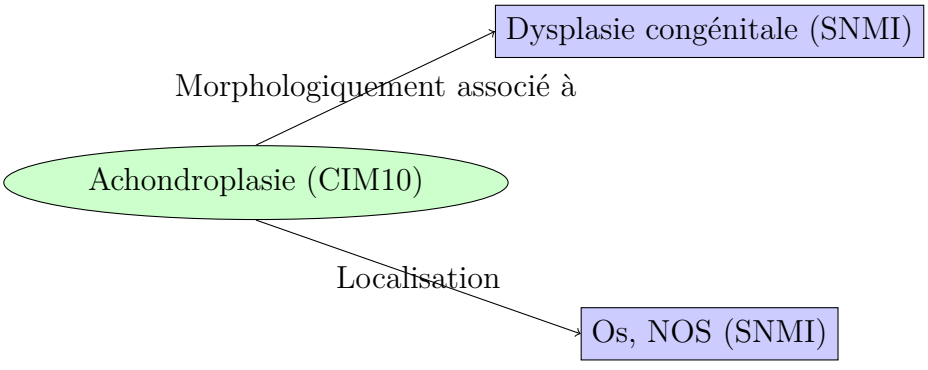

FIG. 7.1 – Exemple d'application d'une projection de relations SNOMED CT entre deux termes SNOMED International et un terme CIM10

d'optimiser l'indexation mutli-terminologique (semi-) automatique en utilisant la projection des relations SNOMED CT vers plusieurs autres terminologies. Ces relations

seront utilisées comme un poids supplémentaire dans le processus d'indexation. Ces résultats vont aussi être exploités dans le cadre de la recherche d'informations multi-terminologique afin d'étendre ou bien restreindre une requête.

7.2 Projection des relations SNOMED CT entre les termes de la terminologie MeSH

Un total de 12 030 couples de termes préférés MeSH est en relation suivant au moins une relation SNOMED CT. Le tableau 7.5 montre les cinq premières relations SNOMED CT (en nombre) projetées vers les couples de termes MeSH. Notons aussi que seulement les relations : « ISA », « Finding_Site_of » (Localisation), « Causative_agent_of » (Agent causal de) et « Associated morphology » (Morphologiquement associé à) sont projetées sur plus de 300 couples de termes MeSH chacune. L'analyse

Top 5 des relations SNOMED CT	**Nombre de couples MeSH en relations**
ISA	6 871
Finding_Site_of (Localisation)	2 512
Associated_Morphology_of (Morphologiquement associé à)	1 080
Causative agent_of (Agent causal)	328
Associated with (Associé à)	53

Tab. 7.5 – Les principales relations SNOMED CT projetées entre termes MeSH

qualitative a été réalisée sur les quatre principales relations projetées vers les termes MeSH. Le tableau 7.6 montre le résultat de cette analyse par une documentaliste CISMeF (Catherine Letord) sur une échelle de Likert avec les trois possibilités décrites dans le chapitre 6.
Pour la relation SNOMED CT « Associated morphology » (Morphologiquement associé à), plus de 90% des relations projetées vers les termes MeSH ont été jugées pertinentes. Cependant, pour la relation « Causative_agent_of » (Agent causal de) ; seulement 64% des relations projetées ont été jugées pertinentes. Le tableau 7.7 donne des exemple d'évaluations pour chaque critère de la projection de la relation « Finding_Site_of »(Localisation) vers les termes MeSH. Pour les quatre relations évaluées dans cette étude (ISA, « Finding_Site_of » (Localisation), « Associated morphology » (Morphologiquement associé à) et « Causative_agent_of » (Agent causal de) les résultats sont très encourageants. Certaines de ces relations sont implémentées

Relations SNOMED CT	**Pertinent (%)**	**Moyen (%)**	**Mauvais (%)**
ISA	75	17	8
Finding_Site_of (Localisation)	88	10	2
Associated_Morphology_of (Morphologiquement associé à)	90	2	8
Causative agent_of (Agent causal)	64	36	0
Moyenne	79,25	16,25	4,5

TAB. 7.6 – Qualité de la projection des quatre principales relations SNOMED CT vers les termes MeSH

Termes MeSH	**Termes MeSH**	**Évaluation**
Abcès abdominal	Cavité abdominale	Pertinent
Abcès du psoas	Cavité abdominale	Moyen
Afibrinogenemia	Système immunitaire	Mauvais

TAB. 7.7 – Exemples d'évaluations pour les trois critères de la projection de la relation « Finding_Site_of » (Localisation)

dans le nouveau Portail Terminologique de Santé (PTS) de CISMeF (voir figure 7.2). Elles vont être aussi utilisées pour optimiser la recherche d'informations dans le moteur de recherche. Cette optimisation va permettre de limiter ou d'étendre une requête. Par exemple, la requête « abdomen aigu » sera étendue ou limitée suivant la relation « Localisation » avec le terme « Localisation ». L'utilisateur pourra ainsi choisir d'étendre sa recherche avec « abdomen aigu **ou** abdomen » ou bien la limiter avec « abdomen aigu **et** abdomen ».

7.3 Synthèse

Dans cette partie nous avons présenté les différents résultats obtenus par la projection des relations SNOMED CT sur trois terminologies francophones. L'évaluation qualitative des relations SNOMED CT projetées sur la terminologies MeSH est très encourageante d'un point de vue documentaliste. Actuellement, toutes les projections pertinentes ont été intégrées dans le Portail Terminologique de Santé. Dans le chapitre 8, nous discutons en détail les résultats de cette méthode.

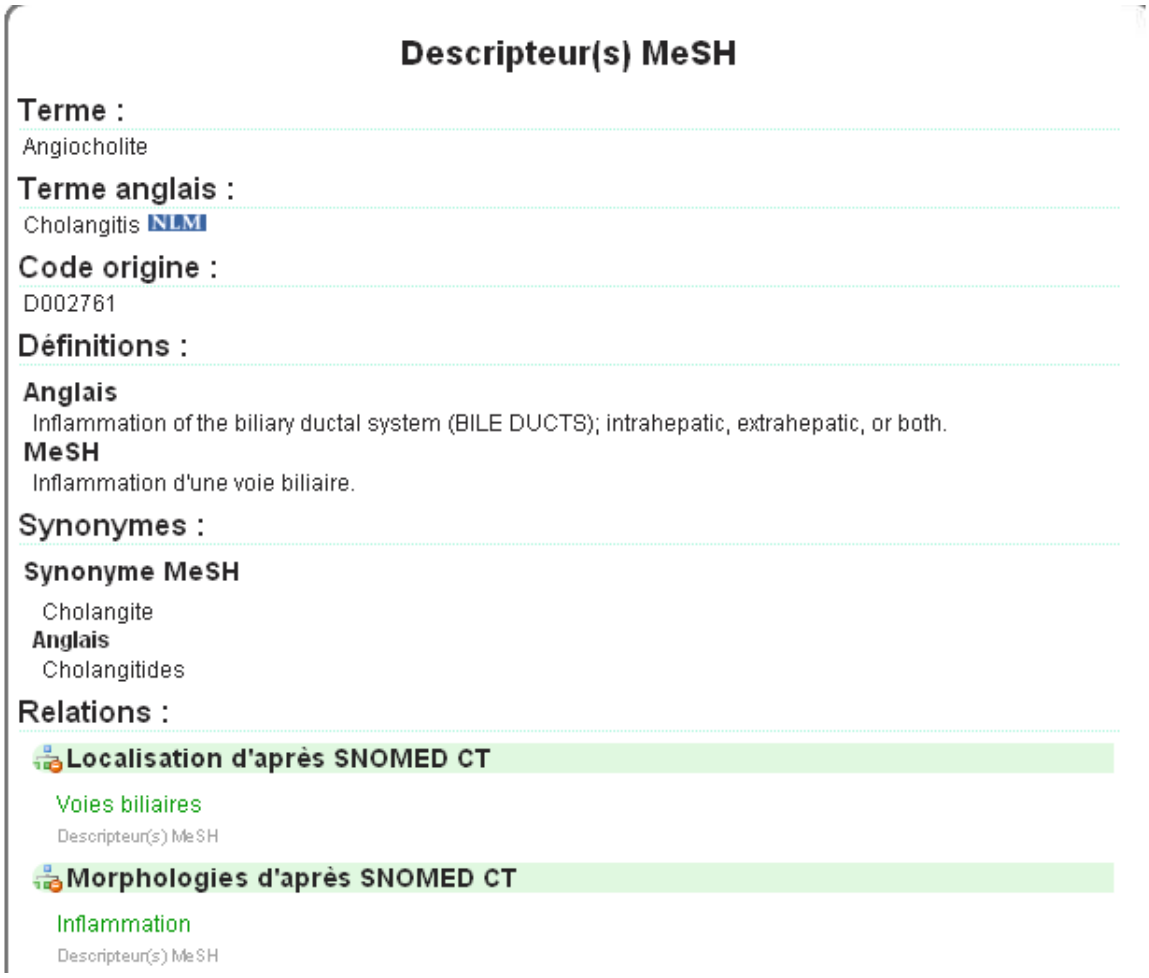

FIG. 7.2 – Exemple de deux relations SNOMED CT projetées entre termes MeSH implémentées dans PTS

Chapitre 8

Discussion

Les méthodes d'alignements et de projections présentées dans cette thèse comportent un certain nombre d'avantages et d'inconvénients et offres diverses perspectives pour les futurs travaux.

8.1 Alignements entre terminologies

Les méthodes d'alignements développées dans cette thèse ont été évaluées sur différentes terminologies de santé :

- le thésaurus ORPHANET : le thésaurus des maladies rares développé et utilisé dans le portail ORPHANET, et depuis peu dans le PTS;
- la classification ATC : une classification spécialisée et utilisée pour classer et hiérarchiser les médicaments;
- la classification CCAM : une classification utilisée dans le codage des actes médicaux qui sert de base à la tarification en secteur libéral en France.

Dans tous les travaux d'alignements que nous avons présentés, nous avons décrit des méthodologies appliquées différemment pour aligner les terminologies sources vers UMLS (F_UMLS). Ainsi, pour ORPHANET, en plus de l'algorithme d'alignement lexical, nous avons combiné les résultats obtenus avec une méthode d'alignement structurel utilisant les relations hiérarchiques du thésaurus ORPHANET. Dans le cas de la classification ATC, nous avons utilisé conjointement notre méthode d'alignement fondée sur le français avec l'outil MetaMap de la NLM fondé sur l'anglais. Pour le dernier cas de la CCAM, l'utilisation de nos méthodes d'alignements directement sur les libellés de la CCAM était impossible à cause de leurs tailles en terme de nombre de mots. Ainsi,

nous avons développé une méthode permettant d'extraire à partir de la structure du code de la CCAM un nouveau « libellé représentatif » moins verbeux. Comme pour la classification ATC, nous avons aussi dans le cas de la CCAM utilisé conjointement nos méthodes avec l'outil MetaMap sur les nouveaux libellés représentatifs obtenus.

Nous avons montré que les résultats étaient différents suivant chaque terminologie source utilisée et suivant chaque méthode appliquée (français ou anglais). Pour ORPHANET, la comparaison de notre méthode lexicale avec un alignement manuel existant a montré que notre méthode donnait plus de résultats pertinents pour le passage vers F_UMLS. Cependant, cette comparaison reste relative à la terminologie intermédiaire utilisée dans l'alignement manuel. En effet, même si la classification CIM10 est présente dans UMLS avec un nombre de codes égal à 13 505, les termes présents dans les mêmes codes CIM10 pour la plupart ne sont pas des synonymes. Par exemple, les termes « abdomen sensible », « colique », « colique infantile » et « douleurs abdominales, autres et non précisées » partagent le même code CIM10 « R10.4 ». Ainsi, un alignement manuel vers un de ces termes implique un alignement vers le code « R10.4 ». Dans l'UMLS le code « R10.4 » n'est présent que dans un seul concept UMLS (terme « douleurs abdominales, autres et non précisées »). L'exemple d'un alignement manuel impliquant le code « colique » donnera forcément le mauvais concept UMLS.
L'application des méthodes lexicales sur ORPHANET nous a permis d'aligner 22% des termes ORPHANET vers F_UMLS en relation d'alignement exact. Nous avons montré aussi que l'utilisation des méthodes lexicales conjointement avec un alignement conceptuel comme celui issu de l'UMLS avait augmenté le nombre des alignements.
Concernant le mathcing structurel, les résultats obtenus montrent que ce type d'alignement perd sont intérêt dès que le mathcing est effectué dans un niveau hiérarchique trop grand par rapport au terme aligné. De plus, il est très important de différencier ce type d'alignement par rapport à l'alignement exact afin d'éviter le bruit dans la recherche d'informations ou dans l'indexation automatique.
Pour l'alignement de la classification ATC vers F_UMLS, les résultats obtenus sont très encourageants par rapport à l'alignement manuel existant entre ATC et la terminologie MeSH. La comparaison entre la méthode lexicale fondée sur le français et la méthode fondée sur l'anglais a montré que la différence n'est pas très grande entre ces deux méthodes. En terme de couverture la méthode fondée sur l'anglais donne 3 062 alignements alors que en français nous avons 2 992 alignements, en les comparant par rapport à l'alignement manuel vers le MeSH comme gold standard, les deux méthodes sont très proches en terme de précision (89% (français) versus 88% (anglais)). Cependant, il est à noter que dans un contexte de pharmacien l'apport manuel est très important pour n'avoir que des alignements valides, en effet, les résultats obtenus ont montré que certains alignements ne peuvent pas être obtenus automatiquement

(lexicalement) non seulement parce que les libellés ATC n'existent pas dans F_UMLS, mais surtout parce que la plupart de ces libellés correspondent à des noms de familles de médicaments à utiliser dans un contexte particulier qui n'est pas présent dans les libellés mais dans sa hiérarchie ascendante que les méthodes automatiques ne peuvent pas traiter.
La dernière classification que nous avons essayé d'aligner est la CCAM. La méthodologie suivie pour aligner cette terminologie est différente par rapport aux autres terminologies définies précédemment. En effet, l'utilisation des méthodes lexicales appliquées directement sur les libellés CCAM pour trouver les concepts UMLS est quasiment impossible. Pour la simple raison, que la longueur des libellés de la CCAM ne permet pas de trouver des concepts UMLS exactement similaires. Pour cette raison, nous avons proposé une méthode fondée sur la structure des codes CCAM. Cette méthode donne des nouveaux libellés représentatifs moins verbeux. Nous avons aussi utilisé dans le cadre de la CCAM deux méthodes lexicales en français et en anglais. L'évaluation des résultats sur la méthode en français a montré que la plupart des alignements correspondent à des alignements NT-BT, c'est-à-dire que les concepts UMLS sont d'un niveau hiérarchique plus général que les libellés CCAM originels. Ces résultats ne sont pas étonnants, vu le niveau de généralité des nouveaux libellés représentatifs obtenus à partir de la structure des codes.

En résumé, les résultats obtenus dans les méthodes utilisées dépendent des terminologies sources (à aligner). Cependant, la qualité des alignements peut différer d'un objectif à un autre et suivant les terminologies utilisées et suivant le contexte, notamment celui très variable de l'utilisateur. Un alignement peut être considéré comme correct dans le cadre de la recherche d'informations et faux dans un contexte d'une indexation automatique ; correct dans un point de vue MeSH et faux dans le cas de la SNOMED. D'un autre côté, l'application des différentes méthodes lexicales a aussi montré différents problèmes que ces méthodes sont incapables de traiter :

- La gestion des acronymes entre les différentes terminologies. Par exemple, l'acronyme « CMT » qui correspond dans ORPHANET au terme : « Maladie de Charcot-Marie-Tooth » alors que le même acronyme correspond dans la terminologie MeSH au terme : « Tumeurs de la thyroïde ». D'un autre côté, les deux termes ORPHANET et MeSH « Canal atrioventiculaire » partageant le même acronyme : « Cavc ». Néanmoins, dans le backoffice CISMeF, ces acronymes sont identifiés comme ambigus et gérés comme tel en recherche d'information mais pas en terme d'alignement (à moins de proposer tous les termes ambigus).
- Un des problèmes lié à l'utilisation des méthodes lexicales concerne les termes avec des sens différents suivant chaque terminologie. Par exemple, « Gauche » et « maladie de gaucher ». Ce problème a été détaillé Weeber *et al.* (2001) où les

auteurs proposent une solution manuelle permettant de réduire les ambiguïtés des termes avec plusieurs sens. Une autre solution applicable seulement aux termes appartenant à l'UMLS consiste à filtrer les correspondances entre les termes qui partagent au moins un type sémantique. Cependant, cette solution, ne peut pas être appliqué dans le cadre de nos travaux car les terminologies à aligner ne sont pas dans l'UMLS (ORPHANET, ATC et CCAM).

- Comme le soulignent les auteurs dans Bodenreider et Zhang (2006), un des problèmes réside aussi dans la gestion des synonymes à travers les différentes terminologies, en effet, le fait que deux termes soient synonymes dans une terminologie n'implique pas que ces mêmes termes sont synonymes dans une autre terminologie. Nos méthodes n'ont pas échappé à ce problème. Dans différents alignements trouvés, les experts ont noté différentes incohérences dues principalement à la différence de ces terminologies en terme de représentation des connaissances. Par exemple, le terme ORPHANET « Syndrome de Marfan » est aligné vers le terme SNOMED « Arachnodactylie », cela s'explique par le fait que le terme MeSH « Syndrome de Marfan » a comme synonyme « Dolichostenomelie » ce dernier est un synonyme du terme SNOMED « Arachnodactylie ». Du côté de l'équipe ORPHANET cet alignement est jugé comme faux, car « Arachnodactylie » correspond à un signe clinique de la maladie dans ORPHANET. Toutefois, corriger ce type de problèmes est « politiquement » difficile. Changer la façon de représenter un terme dans une terminologie peut durer des années comme cela peut ne jamais se faire.
- Au final, il est à noter que notre algorithme lexical reste dépendent aussi de l'algorithme de désuffixation utilisé, pour ce point, nous continuons à mener un travail important à l'intérieur de l'équipe CISMeF à travers les différents projets (actuels et futurs) pour améliorer tous ces outils.

D'un autre côté, nous avons rencontré des problèmes liés à l'évaluation de nos alignements. En effet, il est nécessaire, de notre point de vue, de disposer d'alignements manuels de référence pour chaque type d'alignement pour permettre une bonne évaluation de nos méthodes. L'évaluation des résultats d'alignements permet d'estimer la qualité des alignements obtenus mais ne donne aucune indication sur le rappel. C'est-à-dire que nous n'avons aucun moyen de déterminer les alignements manquants. Dans Sun et Sun (2006), les auteurs soulignent le compromis entre la vitesse de l'automatisation des méthodes d'alignement et la précision manuelle. Bien que les évaluations manuelles soient plus coûteuses en temps, elles restent indispensables pour valider les résultats des méthodes automatiques.

8.2 Projection des relations SNOMED CT

Dans la deuxième partie de la thèse, nous avons proposé une méthode fondée sur l'UMLS permettant de projeter les relations SNOMED CT surtrois terminologies médicales : SNOMED International, CIM10 et MeSH. Dans un premier temps, nous avons projeté les relations SNOMED CT entre les termes de chaque terminologie en « Inter ». Nous avons montré l'intérêt de cette méthode pour lier des termes de différentes terminologies. Dans un autre travail, nous avons appliqué la même méthode pour projeter les relations SNOMED CT entre les termes MeSH. Les résultats obtenus nous ont permis d'évaluer la qualité des projections d'un point de vue documentaliste. Les évaluations ont montré qu'en moyenne plus de 79% des relations obtenues entre les termes MeSH ont été jugés comme pertinentes.
Les résultats de ces études devraient permettre :

- L'optimisation de l'indexation multi-terminologique (semi-) automatique. Les relations projetées seront utilisées comme un poids supplémentaire dans le processus d'indexation.
- Ces relations seront utilisées dans le cadre de la recherche d'informations multi-terminologique afin d'étendre ou restreindre les requêtes. Par exemple, si un utilisateur formule, dans CISMeF, la requête suivante « Achondroplasie », une proposition d'extension ou de limitation de sa requête lui sera faite : « Localisation (Finding_Site_Of) Os ». L'utilisateur pourra ainsi choisir d'étendre sa recherche avec « Achondroplasie ou Os » ou bien la limiter avec « Achondroplasie et Os ».

La méthode décrite dans cette partie est néanmoins dépendante des relations existantes dans la SNOMED CT. En effet, une erreur d'attribution d'une relation SNOMED CT entre deux concepts peut amener à des déductions incorrectes ou abusives dans les autres terminologies.
Prenons l'exemple de la relation « ISA », le concept SNOMED CT « tumeur de l'utérus » est subsumé par le concept « tumeur de l'abdomen », ce qui présente une classification fausse et entraînera une déduction erronée en passant vers d'autres terminologies. Beaucoup d'erreurs de ce type peuvent être trouvées dans la SNOMED CT. Ceusters *et al.* (2004) expliquent que ces erreurs sont causées par deux principaux facteurs :

- Traitement inapproprié de la négation : le concept SNOMED CT « maladie de Dupuytren sans contracture » (Dupuytren's disease of palm, nodeules with no contracture) est subsumé par le concept « rétraction de l'aponévrose palmaire ».
- Traitement inapproprié de la distinction partielle/complète : le concept SNOMED CT « extraction partielle du siège » est subsumé par le concept « extraction du siège » à son tour subsumé par le concept « extraction complète du siège ».
- Plusieurs cas peuvent exister où les relations « ISA » sont confondues avec les relations « Part_Of » Guarino (1998). Un autre problème est la surabondance

des relations « ISA » et l'utilisation peu fréquente des relations qualifiées Cornet (2008). Ces relations ont été introduites afin de faciliter la post-coordination. Par exemple, les deux termes « maladie cardiaque » et « maladie cardiaque aiguë » sont tous les deux présents dans la SNOMED CT, pour lesquelles le terme « infarctus aigu du myocarde » est relié par une relation « ISA ». On aurait pu supposer que l'infarctus du myocarde serait relié à une maladie cardiaque via la relation « ISA » qualifiée par « aigue », et, par conséquent, dans ce cas, se passer du terme « infarctus aigu du myocarde ».

En résumé, le problème dans ce cas réside dans la vérification des relations SNOMED CT Bodenreider *et al.* (2007); Jiang et Chute (2009) ce qui ne relève pas du champ d'application de nos préoccupations.

Chapitre 9

Perspectives

9.1 Amélioration des méthodes

Nous allons continuer à appliquer nos méthodes d'alignement sur les terminologies intégrées dans le PTS. Cela va permettre de réaliser une matrice (N × N), où N représente le nombre des terminologies intégrées. Nous continuons aussi à améliorer nos méthodes d'alignement grâce à l'application des distances de Levenshtein et de Stoilos (actuellement, en cours de développement par Zied Moalla[1]) et l'utilisation des relations *Inter* et *Intra* terminologiques pour trouver plus d'alignements. Pour l'alignement d'ORPHANET, le travail de l'équipe Bio-Health Informatics de l'université de Manchester[2] sur l'alignement de la version anglaise vers la SNOMED CT, nous sera utile pour évaluer les résultats obtenus dans cette thèse par rapport à une autre méthode. Concernant l'alignement de la classification CCAM, nous avons commencé, en collaboration avec l'équipe LERTIM, à travailler sur la possibilité d'utiliser conjointement le libellé et la décomposition du code de la CCAM pour avoir plus de précision. Par exemple, l'acte CCAM « Lithotritie extracorporelle de la vessie » après décomposition en utilisant son code « JDNM » qui sera représenté par les deux termes : « Lithotritie » et « vessie », ainsi, nous perdons la notion d'« extracorporelle » présente dans le libellé. Un découpage sur le libellé permettra de trouver le terme « extracorporelle ».

[1] Zied Moalla, commence sa thèse dans l'équipe CISMeF sur la problématique des questions-réponses dans les catalogues de santé

[2] http://intranet.cs.man.ac.uk/bhig/

9.2 Aide à la traduction

Nous allons poursuivre l'application de nos méthodes pour aider à traduire plusieurs terminologies (médicales ou autres). Dernièrement (en collaboration avec Louis Deléguer et Pierre Zweignbaum), nous avons proposé une traduction de MEDLINEPLUS vers le français, en utilisant des méthodes TAL Deléger (2009) et conceptuelles basées sur UMLS (article soumis à l'AMIA 2010).
Ces méthodes vont être utilisées dans un autre travail permettant la traduction de la FMA (Foundational Model of Anatomy)[3] vers le Français. L'approche envisagée est fondée sur une traduction par partie des libellés de FMA. La traductions de ces parties se fera en utilisant toutes nos terminologies intégrées dans le PTS.

9.2.1 Traduction de la SNOMED CT

Dans ce travail Joubert *et al.* (2009a), nous avons décrit une méthode fondée sur l'UMLS afin de proposer une aide à la traduction de la SNOMED CT en utilisant les alignements conceptuels de l'UMLS. Cette méthode utilise, en plus, un nombre de relations « explicites » existant dans UMLS et identifiées dans la table MRREL. Quatre terminologies en français ont été sélectionnées : MeSH, MedDRA, CIM10 et SNOMED international. La méthode d'alignement utilisée dans cette étude est la suivante :
Supposons que nous ayons deux termes t_1 et t_2 de deux terminologies T_1 et T_2 respectivement. Soient les deux concepts UMLS : CUI_1 et CUI_2, correspondant à la projection des deux termes t_1 et t_2 dans le métathésaurus de l'UMLS respectivement. Deux types d'alignements peuvent exister entre les deux termes t_1 et t_2 :

- Alignement exact (voir figure 9.1) : $CUI_1 = CUI_2$ (dans MRCONSO), ou
- Alignement partiel (voir figure 9.2) : Il existe un alignement explicite entre CUI_1 et CUI_2 dans la table (MRREL).

Si un alignement existe entre CUI_1 et CUI_2, tous les termes du concept CUI_2 sont alignés vers les termes CUI_1. Définies de cette façon, T_1 représente une des terminologies en français utilisées et T_2 représente la SNOMED CT. Cependant, un alignement explicite provenant d'une terminologie T_3 n'appartenant pas à l'ensemble des terminologies utilisées, est appliqué à t_1 dès lors qu'il est établi entre CUI_1 auquel le terme t_1 est attaché et CUI_2 auquel le terme t_2 est attaché.

[3] Ce travail est en collaboration avec le Pr Christine Goldbreich de l'université de Versailles

FIG. 9.1 – Exemple d'alignement exact entre un terme MeSH et un terme SNOMED CT

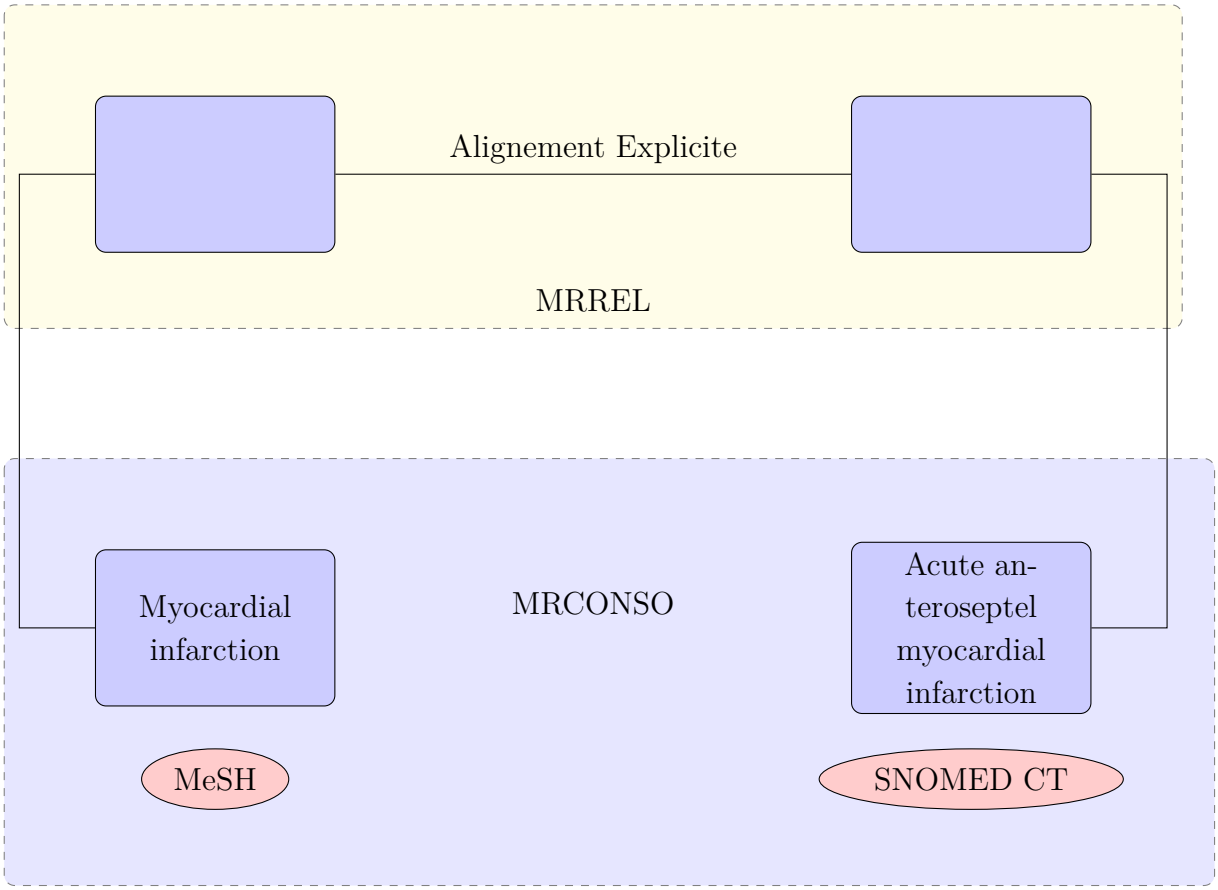

FIG. 9.2 – Exemple d'alignement partiel entre un terme MeSH et un terme SNOMED CT

Chaque terminologie en français a été alignée vers la SNOMED CT en utilisant notre méthode. Le tableau 9.1 donne le nombre et le pourcentage des termes préférés alignés vers au moins un terme préféré SNOMED CT. Au total, l'union des quatre terminologies est aligné vers 82% de la SNOMED CT.

Terminologie	**Termes préférés**	**Terme alignés**	**% des termes préférés**
CIM10	9 308	8 949	96
MedDRA	107 900	98 590	92
MeSH	17 867	9 359	52
SNOMED Int	24 767	14 024	57

Tab. 9.1 – Nombre et pourcentage des termes préférés alignés avec au moins un terme préféré SNOMED CT

D'un autre côté, l'alignement de la SNOMED CT vers l'union des quatre terminologies donne un nombre de 141 068 (45%) des termes préférés SNOMED CT alignés vers l'union de ces quatre terminologies. Cependant, un nombre de 170 245 termes préférés est non aligné (55%). Parmi ces termes, nous avons 146 603 concepts primitifs et 23 642 (8%) des concepts définis.

9.3 Le Projet PlaIR (Plateforme d'Indexation Régionale)

Lancé au sein du laboratoire LITIS, le projet PlaIR a pour objectif de mutualiser l'ensemble des travaux des laboratoires LITIS et LiDiFra (Linguistique Didactique, Francophone) portant sur l'indexation et la recherche d'information, que ce soit dans un univers de documents électroniques avec des vocabulaires contrôlés liés à des domaines métiers (médicale, droit ou les sciences de l'ingénieur) ou dans un univers de documents papiers numérisés en texte intégral sans domaine métier ciblé (comme dans le cas des documents d'archives et du patrimoine), de plus, l'objectif va permettre la réalisation d'une plateforme technologique d'indexation et de recherche d'information.

Chapitre 10

Conclusion

Ce travail apporte une approche automatique permettant de aligner les terminologies francophones dans le domaine de la santé. Les motivations étant, d'une part, mettre en place un modèle commun de représentation de toutes les terminologies , d'autre part, proposer des méthodes pour mettre en correspondance ces terminologies.
Concernant le premier point, nous avons décrit le SMTS (Serveur Multi-Terminologique de Santé) réalisé par MONDECA, LERTIM et CISMeF. L'objectif étant l'intégration de plusieurs terminologies médicales dans un même et unique serveur. Outre la gestion des terminologies de santé francophones, le SMTS va permettre aux professionnels de santé ainsi qu'aux applications un accès en temps réel à toutes les terminologies francophones.
Concernant la mise en relations des terminologies, nous avons proposé un certain nombre de méthodes lexicales et structurelles pour réaliser les alignements. Ces outils sont capables de trouver les termes les plus proches lexicalement entre différents termes de différentes terminologies. D'un autre côté, les alignements proposés ont été réalisés vers les terminologies francophones de l'UMLS (F_UMLS), considérée comme la plus grande base de données terminologique avec plus de 140 terminologies, l'utilisation de l'UMLS permet a avantages, d'une part, d'avoir une large couverture sur toutes les autres terminologies non francophones, d'autre part, l'alignement conceptuel de l'UMLS permettra de trouver plus d'alignements non repérés par les méthodes lexicales.
Dans la plupart de nos travaux, nous avons réalisé plusieurs évaluations manuelles suivant les contextes d'application de chaque terminologies. D'un autre côté, dans la plupart de nos travaux, des comparaisons avec l'outil MetaMap de la NLM ont été réalisées sur les versions anglaises des terminologies sources.
La projection des relations SNOMED CT entre les terminologies a permis d'enrichir ces terminologies par la création de plusieurs relations inter et intra terminologiques.

Tous les travaux que nous avons mentionnés sont intégrés actuellement au sein du PTS (Portail Terminologique de Santé) développé par CISMeF. Nos alignements inter terminologies seront aussi utilisés dans le cadre de la thèse de Sakji Saoussen pour la recherche d'informations Multi-Terminologique.
Néanmoins, un important travail reste à faire sur les autres type d'alignement (par combinaison, partiel) et sur l'amélioration de nos méthodes. Il est vraisemblable que je continue à travailler sur ces différentes problématiques d'interopérabilité entre terminologies.

Liste des publications

Communications internationales

1. Merabti, T; Massari, P; Joubert, M; Sadou, E; Lecroq, T;Abdoune, H; Rodrigues, JM & Darmoni, SJ. An automated approach to map a French terminology to UMLS. MedInfo2010, À paraitre.
2. Merabti, T; Letord, C; Abdoune, H; Lecroq, T; Joubert, M & Darmoni, SJ. Projection and inheritance of SNOMED CT Relations between MeSH Terms. Stud Health Technol Inform MIE2009, Volume 150, Pages 233-237, IOS Press, 2009.
3. Merabti, T; Pereira, S; Letord, C; Lecroq, T; Dahamna, B; Joubert, M & Darmoni, SJ. Searching Related Resources in a Quality Controlled Health Gateway: a Feasibility Study. eHealth Beyond the Horizon - Get IT There - Proceedings of MIE2008 - The XXIst International Congress of the European Federation for Medical Informatics, Göteborg, Sweden, May, Studies in Health Technology and Informatics, Volume 136, pages 235-240, 2008.
4. 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, pages 255-259, Springer, 2009. PSIP.
5. Joubert, M; Abdoune, H; Merabti, T; Darmoni, SJ & Fieschi, M. Assisting the Translation of SNOMED CT into French using UMLS and four Representative French-language Terminologies. AMIA symp.2009.

Publications nationales

1. Merabti, T; Joubert, M;Lecroq, T; Rath, A; Darmoni, SJ. Mapping biomedical terminologies using natural language processing tools and UMLS: mapping the Orphanet thesaurus to the MeSH. Ingénierie et Recherche Biomédicale / Biomedical Engineering and Research, 2010. À paraitre.
2. Pauchet, A; El Abed, M; Merabti, T; Prieur, E; Lecroq, T & Darmoni, SJ. Identification de répétitions dans les navigations au sein d'un catalogue de santé. RIA (Revue d'Intelligence Artificielle), Volume 23, Numéro 1, Pages 113-132, 2009.

Communications nationales

1. Merabti, T; Abdoune, H; Lecroq, T; Joubert, M & Darmoni, SJ. Projection des relations SNOMED CT entre les termes de deux terminologies (CIM10 et SNOMED 3.5). Risques, technologies de l'information pour les pratiques médicales : comptes rendus des treizièmes journées francophones d'informatique médicale (JFIM), Nice, France, Avril, Informatique et santé, Volume 17, pages 79-88, 2009.

Posters

1. Merabti, T; Pereira, S; Lecroq, T; Joubert, M & Darmoni, SJ. Inheritance of SNOMED CT Relations between Concepts by two Health Terminologies (SNOMED International and ICD-10). KR-MED 2008 - Representing and sharing knowledge using SNOMED International Conference, Phoenix, AZ, USA, June, 2008.
2. Joubert,M; Merabti, T; Vedenbusshe, PV; Abdoune, H; Dahamna, B; Fieschi, M & Darmoni,SJ. Modeling and Integrating terminologies into a French Multi-terminology server. MedInfo2010, À paraitre.
3. Bousquet, C; Sadou, E; Merabti, T; Trombert, B; Kumar, A; Darmoni, SJ & Rodrigues, JM. Multiaxial description of the French CCAM terminology for clinical procedures on the UMLS metathesaurus. MedInfo2010, À paraitre.
4. Merabti, T & Darmoni SJ. Web sémantique au sein de CISMeF. i-expo, salon de l'information numérique, Juin, 2009.

Bibliographie

(2000). ISO 1087-1:2000 terminology work - vocabulary - part 1:theory and application.

(2005). CEN TC 251 EN 12264:2005 informatique de santé - structure catégorielles des systèmes de concepts.

(2007). ISO 11715 health informatics - vocabulary for terminological systems.

Amardeilh, F. et Francart, T. (2004). A semantic web portal with hlt capabilities. *In Actes du colloque Veille Stratégique Scientifique et Technologique*.

Amardeilh, F., Laublet, P. et Minel, J. (2005). Annotation documentaire et peuplement d'ontologie à partir d'extractions linguistiques. *In Acte IC*.

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *In Proc. AMIA Symp. 2001*, pages 17–21.

Austin, C. (1968). Medlars. 1963-1967. Rapport technique, National Library of Medecine.

Avillach, P., Joubert, M. et Fieschi, M. (2007). A model for indexing medical documents combining statistical and symbolic knowledge. *Proc. AMIA Symp. 2007*, pages 31–35.

Bachrach, C. et Charen, T. (1978). Selection of MEDLINE contents, the development of its thesaurus, and the indexing process. *Med Inform (Lond)*, 3:237–254.

Bechhofer, S., van Harmelen, F., Hendler, J., Horrocks, I., McGuinness, D., Patel-Schneider, P. F. et Stei, L. A. (2004). OWL Web Ontology Language Reference. Rapport technique, w3c recommendation.

Beuscart, R., McNair, R., Darmoni, S., Koutkia, V., Maglaveras, N., Beuscart-Zephir, M. et Nohr, C. (2009). PSIP Project Consortium. *In Stud Health Technol Inform*, volume 148, pages 14–24.

Bodenreider, O. (2004). The Unified Medical Language System (umls): Integrating biomedical terminology. *Nucleic Acids Res*, 32:267–270.

Bodenreider, O., Nelson, S. J., Hole, W. T. et Chang, H. F. (1998). Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. *In Proc. AMIA Symp. 1998*, pages 815–819.

Bodenreider, O., Smith, B., Kumar, A. et Burgun, A. (2007). Investigating subsumption in SNOMED CT: an exploration into large description logic-based biomedical terminologies. *Artif Intell Med*, 39(3):183–95.

Bodenreider, O. et Zhang, S. (2006). Comparing the representation of anatomy in the FMA and SNOMED CT. *In AMIA Annu Symp Proc*, pages 46–50.

Booch, B., Rumbaugh, J. et Jacobson, L. (2000). *Le guide de l'utilisateur UML*. Eyrolles.

Bouaud, J., Séroussi, B., Dréau, H., falcoff, H., Riou, C., Joubert, M., Simon, C., Simon, G. et Venot, A. (2002). ASTI, un système d'aide à la prescription médicamenteuse basé sur les guides de bonnes pratiques. *Informatique et Santé*, 3:81–8.

Bourda, Y. et Hélier, M. (1999). Appliying IEEE Learning Object Metadata to Publishing Teaching Programs. Rapport technique, ED-MEDIA.

Bourigault, D., Aussenac-Gilles, N. et Charlet, J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, 18:87–100.

Bousquet, C., Sadou, E., Merabti, T., Trombert, B., Kumar, A., Darmoni, S. et Rodrigues, J. (2010). Multiaxial description of the French CCAM terminology for clinical procedures and mapping on the UMLS metathesaurus. *In Proc. MEDINFO. 2010*, Cap town, South Africa. À paraitre.

Brown, E., Wood, L. et Wood, S. (1999). The medical dictionary for regulatory activities (MedDRA). *Drug Saf*, 2:109–117.

Carter, J. S., Brown, S. H., Erlbaum, M. S., W, G., Elkin, P. L., T, S. et Tuttle, M. S. (2002). Initializing the VA medication reference terminology using UMLS metathesaurus co-occurrences. *In Proc. AMIA Symp. 2002*, pages 116–20.

Ceusters, W., Smith, B., Kumar, A. et Dhaen, C. (2004). Ontology-Based Error Detection in SNOMED-CT. *In Proc. MEDINFO. 2004*, pages 482–486.

Chevallier, J. (2006). *TOTHEM - Classification TOpographique et THÈmatique du domaine de la santé*. Èditions Glyphe, Paris.

Chute, C., Elkin, P., Sheretz, D. et Tuttle, M. (1999). Desiderata for a clinical terminology server. *In Proc. AMIA Symp. 1999*, pages 42–6.

Cimino, J. et Barnett, G. (1990). Automated translation between terminologies using semantic defintions. *MD Comput*, 7:104–109.

Cornet, R. (2008). Do SNOMED CT Relationships qualify ? *In Stud Health Technol Inform*, volume 136, pages 785–90.

Côté, R. (1972). From SNOP to SNOMED - A Challenge for the Medical record Librarian. *Bulletin of the Canadian Asssciation of Medical Record Librarians*, 5.

Côté, R. A., Rothwell, D. J., Patolay, J., Beckett, R. et Brochu, L. (1993). The Systematised Nomenclature of Human and Veterinary Medicine: SNOMED International.

Cutting, D., Hatcher, E. et Gospodnetic, O. (2004). *Lucene in Action*. Manning Publications.

Darmoni, S., Amsallem, E., Haugh, M. C., Lukacs, B., Chalhoub, C. et Leroy, J.-P. (2003a). Level of evidence as a future gold standard for the content quality of health resources on the internet. *Methods of Information in Medicine*, 42(3):200–225.

Darmoni, S., Joubert, M., Dahamna, B., Delahousse, J. et Fieschi, M. (2009a). Smts: a French Health Multi-Terminology Server. *In Proc. AMIA Symp. 2009*. InterSTIS.

Darmoni, S., Leroux, V., Thirion, B., Santamaria, P. et Gea, M. (1999). Netscoring : critères de qualité de l'information de santé sur internet. *Les enjeux des industries du savoir*, pages 29–44.

Darmoni, S., Sakji, S., Pereira, S., Merabti, T., Prieur, E., Joubert, M. et Thirion, B. (2009b). Multiple terminologies in a health portal: automatic indexing and information retrieval. *In Artificial Interlligence in Medecine*, Lecture Notes in Computer Science, pages 255–259, Verona, Italy. Springer.

Darmoni, S., Thirion, B., Ionut-Florea, F., Rogozan, A., Letord, C., Kerdelhué, G. et Dacher, J. (2007). Affiliation of a resource type to a MeSH term in a quality-controlled health gateway. *In Proc. Medinfo 2007*, pages 290–292.

Darmoni, S., Thirion, B., Leroy, J. et Douyère, M. (2001). The use of Dublin core metadata in a structured health resource guide on the internet. *Bull Med Libr Assoc*, 89(3):297–301.

Darmoni, S., Thirion, B., Platel, S., Douyère, M., Mourouaga, P. et Leroy, J. (2002). CISMeF-patient : a French counterpart to MEDLINE-plus. *J Med Libr Assoc*, 90:248–253.

Darmoni, S. J., Jarousse, E., Zweigenbaum, P., Le Beux, P., Namer, F., Baud, R., Joubert, M., Vallée, H., Cote, R. A., Buemi, A., Bourigault, D., Recourcé, G., Jenneau, S. et Rodrigues, J. (2003b). VUMeF: Extending the French involvement in the UMLS metathesaurus. *In Proc. AMIA Symp. 2003*, page 824.

de Keizer, N. F., Abu-Hanna, A. et Zwetsloot-Schonk, J. H. (2000). Understanding terminological systems. i: Terminology and typology. *Methods Inf Med*, 39(1):16–21.

Dekkers, M. et Weibel, S. (2003). State of the Dublin Core Metadata Initiative. *D-Lib Mag*, 9(40).

Deléger, L. (2009). *Exploitation de corpus parallèles et comparables pour la détection de correspondances lexicales : application au domaine médical*. Thèse de doctorat, Université Pierre et Marie Curie - Paris 6.

Dirieh Dibad, A., Sakji, S., Prieur, E., Joubert, M. et Darmoni, S. (2009). Recherche Mutli-terminologique en contexte : Étude préliminaire. *In Risques, technologies de l'information pour les pratiques médicales : comptes rendus des treizièmes journées francophones d'informatique médicale (JFIM)*, volume 17 de *Informatique et santé*, pages 101–112, Nice, France. Springer.

Doan, A., Noy, N. et Halvey, A. (2004). Introduction to the special issue on semantic integration. *SIGMOD Record*, 33:11–13.

Dougoulet, P., Fieschi, M. et Attali, C. (1997). Les enjeux de l'interopérabilité sémantique dans les systèmes d'information de santé. *In Informatique et Santé*, volume 2, page 203:212.

Douyère, M., Soualmia, L., Névéol, A., Rogozan, A., Dahamna, B., Leroy, J.-P., Thirion, B. et Darmoni, S. (2004). Enhancing the MeSH thesaurus to retrieve French online health resources in a quality-controlled gateway. *Health Info Libr J.*, 21(4):253–61.

El-Abed, M. (2008). Identification des répétitions dans les navigations au sein d'un catalogue de santé. Rapport de stage de Master 2 Recherche Informatique Théorique et Applications, 6 mois, Université de Rouen.

Euzenat, J. et Shvaiko, P. (2007). *Ontology Matching*. Hiedelberg: Springer-Varlag.

Fellbaum, C., éditeur (1998). *WordNet: an electronic lexical database*. MIT Press.

Ferru, P. et Kandel, O. (2003). Dictionnaire des résultats de consultation (révision 2003-04). *Doc Rech Med Gen*, 62:3–54.

Fieschi, M. (2005). Vers le dossier médical personnel. Les données du patient partagées : un atout à ne pas gâcher pour faire évoluer le système de santé. *Revue Droit Social*, pages 80–90.

Fung, K. et Bodenreider, O. (2005). Utilizing UMLS for semantic mapping between terminologies. *In Proc AMIA Symp*, pages 266–270.

Gruber, T. (1993). Toward principles for the design of ontologies used for knowledge sharing. *In Formal Ontology in Conceptual Analysis and Knowledge Representation*. Kluwar Academic Publishers.

Guarino, N. (1995). Formal ontology, conceptual analysis and knowledge representation. *International Journal of Human-Computer Studies*, 43(5/6):625–640.

Guarino, N. (1998). Some ontological principles for designing upper level lexical resources. *In International Conference on Language resources and evaluation*, pages 527–34.

Hamming, R. (1950). Error detecting and error correcting codes. Rapport technique, Bell System Technical Journal.

Imel, M. (2002). A closer look: the SNOMED clinical terms to ICD-9-CM mapping. *J AHIMA*, 73(6):66–9; quiz 71–2.

Jaccard, J. (1901). Distibution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la société vaudoise des sciences naturelles*, 37:241–272.

Jamoulle, M., Roland, M., Humbert, J. et Brûlet, J.-F. (2000). *Traitement de l'information médicale par la Classification internationale des soins primaires, deuxième version : CISP-2*. Care Edition, Bruxelles.

Jiang, G. et Chute, C. (2009). Auditioning the semantic completeness of SNOMED CT using formal concept analysis. *J Am Med Inform Assoc*, 78:86–94.

Johnson, H., Cohen, K., Baumgartner, W., Lu, Z., Bada, M., Kester, T., Kim, H. et Hunter, L. (2006). Evaluation of lexical methods for detecting relationships between concepts from multiple ontologies. *In Pac Symp Biocomput*, pages 28–39.

Joubert, M., Abdoune, H., Merabti, T., Darmoni, S. et Fieschi, M. (2009a). Assisting the translation of SNOMED CT into French using UMLS and four representative French-language terminologies. *In Proc. AMIA Symp. 2009*, pages 291–295. InterSTIS.

Joubert, M., Dahamna, B., Delahousse, J., Fieschi, M. et SJ, D. (2009b). SMTS: Un Serveur Multi-terminologies de santé. *In Risques, technologies de l'information pour les pratiques médicales : comptes rendus des treizièmes journées francophones d'informatique médicale (JFIM)*, volume 17 de *Informatique et santé*, pages 47–56, Nice, France. Springer. InterSTIS.

Joubert, M., Dufour, J., Aymard, S., Falco, L., Staccini, P. et Fieschi, M. (2003). Le projet CoMeDIAS : Accès à des bases de données hétérogènes au moyen de services internet. *Informatique et Santé*, 16.

Joubert, M., Fieschi, D. et Fieschi, M. (2002). ARIANE : un moteur de recherche de deuxième génération dans le domaine de la santé. *Informatique et Santé*, 13.

Joubert, M., Gaudinat, A., Boyer, C., Fieschi, M. et membres H.F.C (2007). WRAPIN: a tool for patient empowerment within EHR. *Stud Health Technol Inform*, 129:147–51.

Kim, W., Aronson, A. R. et Wilbur, W. J. (2001). Automatic MeSH term assignment and quality assessment. *In Proc. AMIA Symp. 2001*, pages 319–323.

Kondrak, G. (2005). N-gram similarity and distance. *In Proc of the 12th International Conference on String Processing and Information Retrieval*, pages 115–126, Buenos Aires, Argentina.

Lefevre, P. (2000). *La recherche d'informations : du texte intégral au thésaurus.* Editions Hermès.

Leroy, G. et Chen, H. (2001). Meeting medical terminology needs-the ontology-enhanced medical concept mapper. *IEEE Transactions on Information Technology in Biomedicine*, 5(4):261–270.

Lethord, C., Sakji, S., Pereira, S., Dahamna, B., Kergourlay, I. et Darmoni, S. (2008). Recherche d'information multi-terminologique : application à un portail d'information sur le médicament en Eeurope. *Ingénierie et Recherche Biomédicale*, Number 29:350–356.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. *Sov. Phys. Dokl.*, pages 707–710.

Lin, D. (1998). An information-theoretic definition of similarity. *In Proc. Int. Conf. on Machine Learning*, pages 296–304.

Lindberg, D., Humphreys, B. et McCray, A. (1993). The Unified Medical Language System. *Methods Inf Med*, 32(4):281–291.

Lord, P., Stevens, R., Brass, A. et Goble, C. (2003). Investigating semantic similarity measures across the gene ontology:the relationships between sequence and annotation. *Bioinformatics*, 19:1275–1283.

Mansour, I. (2008). Détction et désambiguïsation des abréviations. Rapport de stage de Master 2 Recherche Informatique Théorique et Applications, 6 mois, Université de Rouen.

Mayer, M. A., Darmoni, S., Fiene, M. et Al. (2003). MedCIRCLE - modeling a collaboration for internet rating, certification, labeling and evaluation of health information on the semantic world-wide-web. *In Medical Informatics Europe*, pages 667–672.

Maynard, D. et Ananiadou, S. (2001). Term extraction using a similarity-based approach. *In Didier Bourigault, Christian Jacquemin, and Marie-Claude Lhomme, editors, Recent advances in computational terminology*, pages 261–278.

Mazuel, L. et Charlet, J. (2009). Alignement entre ontologies de domaine et la SNOMED : trois études de cas. *In Ingénierie des Connaissances*.

McCray, A., Srinivasan, S. et Brown, A. (1994). Lexical methods for managing variation in biomedical terminologies. *In Annual Symposium on Computer Applications in Medical Care*, pages 235–239.

McCreight, E. (1976). A space-economical suffix tree construction algorithm. *Journal of the ACM*, 23:262–272.

McDonald, C., Huff, S. et Suico, J. (2003). Universal standard for identifying laboratory observations. *A 5 year update Clin Chem*, 49:624–633.

McKusick, V. (2004). *Mendelian Inheritance in Man; A Catalog of Human Genes and Genetic Disorders*. The Johns Hopkins University Press, Baltimore, MD.

Merabti, T. (2007). Distance inter-document. Rapport de stage de Master 2 Recherche Informatique Théorique et Applications, 6 mois, Université de Rouen.

Merabti, T., Abdoune, H., Lecroq, T., Joubert, M. et Darmoni, S. (2009a). Projection des relations SNOMED CT entre les termes de deux terminologies (CIM10 et SNOMED 3.5. *In Risques, technologies de l'information pour les pratiques médicales : comptes rendus des treizièmes journées francophones d'informatique médicale (JFIM)*, volume 17 de *Informatique et santé*, pages 79–88, Nice, France. Springer.

Merabti, T., Joubert, M., Lecroq, T., Rath, A. et Darmoni, S. (2010a). Mapping biomedical terminologies using natural language processing tools and UMLS: mapping the Orphanet thesaurus to the MeSH. *Ingénierie et Recherche Biomédicale*. À paraitre.

Merabti, T., Letord, C., Abdoune, H., Lecroq, T., Joubert, M. et Darmoni, S. (2009b). Projection and inheritance of SNOMED CT relations between MeSH terms. *In MIE2009*, volume 150, pages 233–7. IOS Press.

Merabti, T., Massari, P., Joubert, M., Sadou, E., Lecroq, T., Abdoune, H., Rodrigues, J. et Darmoni, S. (2010b). Automated approach to map a French terminology to UMLS. *In MedInfo2010*, Cap Town, South Africa. À paraitre.

Merabti, T., Pereira, S., Letord, C., Lecroq, T., Dahamna, B., Joubert, M. et Darmoni, S. (2008). Searching related resources in a quality controlled health gateway: a feasibility study. *In The XXIst International Congress of the European Federation for Medical Informatics (MIE'08)*, volume 136, pages 235–249.

Metzger, M., Gicquel, Q., Proux, D., Pereira, S., Kergorlay, I., Serrot, E., Segond, F. et Darmoni, S. (2009). Development of an automated detection tool for healthcare-associated infections based on screening. *In Proc. AMIA*, pages 1–2.

Miller, N., Lacroix, E. M. et Backus, J. E. (2000). MEDLINEplus: building and maintaining the national library of medecine's consumer health web service. *Bull Med Libr Assoc*, 88(1):11–7.

Mori, A., Consorti, F. et Galeazzi, E. (1998). Standards to support development of terminological systems for healthcare telematics. *Methods Inf Med*, 37:551–563.

Nelson, S. J., Brown, S. H., Erlbaum, M. S., Olson, N., Powell, T., Carlsen, B., Carter, J., Tuttle, M. S. et Hole, W. T. (2002). A semantic normal form for clinical drugs in the umls: early experiences with the vandf. *Proc AMIA Symp*, pages 557–561.

Névéol, A. (2005). *Automatisation des tâches documentaires dans un catalogue de santé en ligne*. Thèse de doctorat, INSA de Rouen.

Névéol, A., Mork, J. G., Aronson, A. R. et Darmoni, S. J. (2005). Evaluation of French and english MeSH indexing systems with a parallel corpus. *Proc. AMIA Symp. 2005*, pages 565–569.

Névéol, A., Zeng, K. et Bodenreider, O. (2006). Besides precision & recall: Exploring alternative approaches to evaluating an automatic indexing tool for MEDLINE. *In Proc. AMIA Symp. 2006*, pages 583–93.

OMG (2002). Meta object facility, version 1.4, april 2002. Rapport technique, OMG.

OMS, O. (1993). Classification statistique internationale des maladies et des problèmes de santé connexes. Dixième révision.

OMS, O. (2001). Classification internationale des handicaps : déficiences, incapacités et désavantages.

Paternostre, M., Francq, P., lamoral, J., Wartel, D. et saerens, M. (2002). Carry, un algorithme de désuffixation pour le français. Version électronique disponible sur http://www.galilei.ulb.ac.be.

Pauchet, A., El Abed, M., Merabti, T., Prieur, E., Lecroq, T. et Darmoni, S. (2009). Identification de répétitions dans les navigations au sein d'un catalogue de santé. *RIA (Revue d'Intelligence Artificielle)*, 23:113–132.

Pereira, S. (2007). *Indexation multi-terminologique de concepts en santé*. Thèse de doctorat, Université de Rouen.

Pereira, S., Massari, P., Buemi, A., Dahamna, B., Serrot, E., Joubert, M. et Darmoni, S. (2008). Evaluation of two French SNOMED indexing systems with a parallel corpus. *In KR-MED 2008 - Representing and sharing knowledge using SNOMED International Conference*, Phoenix, AZ, USA.

Pereira, S., Massari, P., Buemi, A., Dahamna, B., Serrot, E., Joubert, M. et Darmoni, S. (2009a). F-MTI : outil d'indexation multi-terminologique : application à l'indexation automatique de la SNOMED. *In Risques, technologies de l'information pour les pratiques médicales : comptes rendus des treizièmes journées francophones d'informatique médicale (JFIM)*, volume 17 de *Informatique de Santé*, pages 57–67, Nice, France. Springer.

Pereira, S., Sakji, S., Névéol, A., Kergourlay, I., Kerdelhué, G., Serrot, E., Joubert, M. et Darmoni, S. (2009b). Multi-Terminology indexing for the assignement of MeSH descriptors. *In Proc. AMIA Symp. 2009*.

Porter, M. F. (1980). An algorithm for suffix stripping. *Program*, 3(14):130–137.

Proux, D., Marchal, P., Segond, F., Kergourlay, I., Pereira, S., Gicquel, Q., Darmoni, S. et Metzger, M. (2010). Improving hospital document workflow with a risk patterns detection tool to detect potential hospital acquired infections. *In Biomedical Information Extraction, RANLP conference*. Accepted. ALADIN.

Prud'hommeaux, E. et Seaborne, A. (2008). SPARQL Query Language for RDF. Rapport technique, W3C Working Draft.

Rector, A., Bechhover, S. et Goble, C. (1997). The GRAIL concept modelling language for medical terminology. *Artif Intell Med*, 9(2):139–71.

Rector, A., Nowlan, W. et the GALEN consortium (1993). The GALEN Project:. *Comput Methods Programs Biomed*, 45:75–78.

Rector, A., Rogers, J., Zanstra, P. et Ven Der Haring, E. (2003). openGALEN : open source medical terminology and tools. *In Proc. AMIA Symp. 2003*.

Rekik, S. (2007). Modélisation de terminologies médicales. Rapport de stage de Master 2 Recherche Informatique Théorique et Applications, 6 mois, Université de Rouen.

Resnik, P. (1995). Using information content to evaluate semantic similarity in a taxonomy. *In the 14th International Joint Conference on Artificial Intelligence, Montreal*, pages 448–453.

Rocha, R., Rocha, B. et Huff, S. (1994). Automated translation between medical vocabularies using a frame-based interlingua. *In Proceedings of the 18th Annual Symposium on Computer Applications in Medical Care*, pages 690–694.

Roche, C. (2005). Terminologie et ontologie. *Revue Langages*, 157:48–62.

Rodrigues, J., Trombert Paviot, B., Marti, C. et P, V. (2005a). Integrating the modelling of EN 1828 and Galen CCAM Ontologies with protégé : toward a Knowledge acquisition tool for surgical procedures. *In Stud Health Technol Inform*, pages 69–77.

Rodrigues, J., Trombert-Paviot, B., Martin, C. et Vercherin, P. (2005b). Représentation du standard européen de terminologie en1828 et de galen ccam avec l'éditeur d'ontologie protégé : vers un système terminologique de troisième génération pour les interventions chirurgicales. *In JFIM2005*.

Rosse, C. et Mejino, J. (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *Journal of Biomedical Informatics*, 36:478–500.

Sakji, S. (2008). Recherche multi-terminologique de l'information de santé sur l'internet. *In CORIA (cinquième édition de la COnférence en Recherche d'information et Applicatoin*, pages 409–416, Tregastel, France.

Sakji, S., Dirieh Dibad, A., Kerogourlay, I., Joubert, M. et Darmoni, S. (2009a). Information retrieval in context using various health terminologies. *In RCIS, International Conference on Research Challenges in Information Science*, pages 453–458, Fez, Morocco. IEEE.

Sakji, S., Lethord, C., Pereira, S., Dahamna, B., Joubert, M. et Darmoni, S. J. (2009b). Drug information portal in Europe: Informatio retrieval with multiple health terminologies. *In Stud Health Technol Inform*, volume 150, pages 497–501.

Salton, G. et Buckley, C. (1988). Term weighting approaches in automatic text retrieval. *Information Processing & Management*, 33(4):495–512.

Salton, G. et McGill, M. J. (1983). *Introduction to Modern Information Retrieval*. McGraw-Hill, New York.

Sarker, I., Cantor, M., Gelman, R., Hartel, F. et Lussier, Y. (2003). Linking biomedical language information and knowledge ressources in the 21st Century: GO and UMLS. *In Pacific Symposium on Biocomputing*, volume 8, pages 439–450.

Skrbo, A., Bergovic, B. et Skrbo, S. (2004). Classification of drugs using the ATC system (anatomic, therapeutic, chemical classification) and the latest changes. *Med Arh*, 58(suppl 2):138–41.

Smith, B. (2003). *Blackwell Guide to the Philosophy of Computing and Information*, chapitre Ontology, pages 155–166. Oxford: Blackwell.

Smith, B. (2006). From concepts to clinical reality: an essay on the benchmarking of biomedical terminologies. *J Biomed Inform*, 39(3):288–298.

Smith, B., Ceusters, W. et Temmerman, R. (2005). Wüsteria. *Stud Health Technol Inform*, 116:647–652.

Smith, B. et Welty, C. (2001). Fois introduction: Ontology- towards a new synthesis. *In FOIS '01 : Proceedings of the international conference on Formal Ontology in Information Systems*.

Soualmia, L. (2004). *Étude et Évaluation d'Approches Multiples d'Expansion de Requêtes pour une Recherche d'Information Intelligente : Application au Domaine de la Santé sur Internet*. Thèse de doctorat, INSA de Rouen.

Soualmia, L., Barry, C. et Darmoni, S. (2009). Knowledge-based query expansion over a medical terminology oriented ontology. *In Artif Intell Med : 9th Conference on Artificial Intelligence in Medecine in Europe, AIME*, pages 209–213.

Sowa, J. (2000). *Knowledge representation: Logical, Philosophical and Computational Foundations*. Brooks/cole.

Spackman, K. (2000). SNOMED RT and SNOMED CT : promise of an international clinical terminology. *MD Computing*, 17(6):29.

Spackman, K. A., Campbell, K. E. et Côté, R. A. (1993). SNOMED RT: A Reference Terminology for Health Care. *In Proc AMIA Annu Fall Symp*, pages 640–4.

Stoilos, G., Stamou, G. et Kollias, S. (2005). A string metric for ontology alignement. *In International Semantic Web Conference*, volume 3729, pages 624–637.

Sun, J. et Sun, Y. (2006). A system for automated lexical mapping. *J Am Med Inform Assoc*, 13(3):334–43.

Thirion, B., Douyère, M., Soualmia, L., Dahamna, B., Leroy, J. et Darmoni, S. (2004). Metadata element sets in the CISMeF quality-controlled health gateway. *In International Confrence on Dublin Core and Metadata Applications*, page 12, Shanghai.

Uschold, M. et Grüninger, M. (1996). Ontologies : principles, methods, and applications. *Knowledge Engineering Review*, 11(2):93–155.

van Rijsbergen, C. (1979). *Information retrieval*. London, butterworth édition.

W3C (2004). Simple knowledge organization system. http://www.w3.org/2004/02/skos/. Rapport technique, World Wide Web Consortium.

Wang, Y., Patrick, J., Miller, G. et O'Hallaran, J. (2008). A computational linguistics motivated mapping of ICPC-2 PLUS to SNOMED CT. *BMC Med Inform Decis Mak*, 8 Suppl 1:5.

Weeber, M., Mork, J. et Aronson, A. (2001). Developing a test collection for biomedical word sens disambiguation. *In Proc. AMIA Symp. 2001*, pages 746–750.

Wegner, P. (1996). Interoperability. *ACM Computing Survey*, 28(1):285–7.

Weiner, P. (1973). Linear pattern matching algorithm. *In Proc. 14th IEEE Symposium on Switching and Automata Theory*.

WHO, W. (1992). WHO-ART : International monitoring of adverse reactions to drugs: adverse reaction terminology. Collaborating Center of International Drug Monitoring.

Wu, Z. et Palmer, M. (1994). Verb semantics and lexical selection. *In 32nd Annual Meetings of the Associations for Computational Linguistics*, pages 133–138.

Zeng, M. et Chan, L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. *Journal of the American Society for Information*, 55:377–395.

Zweigenbaum, P. (1999). Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *ISIS*, 3:27–47.

Zweigenbaum, P., Baud, R., Burgun, A., Namer, F., Jarousse, E., Grabar, N., Ruch, P., Le Duff, F., Thirion, B. et Darmoni, S. (2003). UMLF: construction d'un lexique médical francophone unifié. *In Actes des JFIM 2003*.

Annexe A

Étude de cas sur le Serveur Multi-terminologique de Santé

Nous présentons une étude de cas permettant la navigation dans le SMTS entre deux terminologies CIM10 et SNOMED 3.5. La première figure A.1 montre la page d'accueil du SMTS avec les terminologies médicales déjà incluses (dans la partie gauche de l'écran). Le choix d'une terminologie permet de développer le contenu de cette terminologie. Par exemple, le choix de la terminologie SNOMED 3.5 permet de voir toutes les catégories existantes dans cette terminologie (chapitres, axes, section...). Ceci permettra un accès suivant plusieurs choix : différents axes, différents chapitres, différentes sections ...

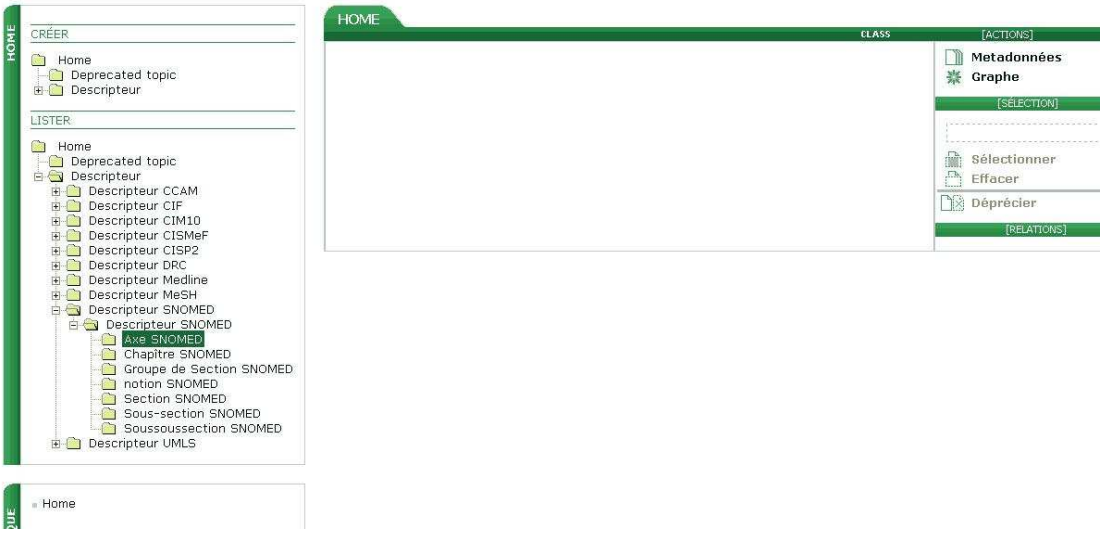

FIG. A.1 – Page d'accueil du SMTS

La figure A.2 montre un exemple d'accès sur l'axe D (des maladies) de la SNOMED 3.5. Comme montre la figure, les maladies sont classées en chapitres.

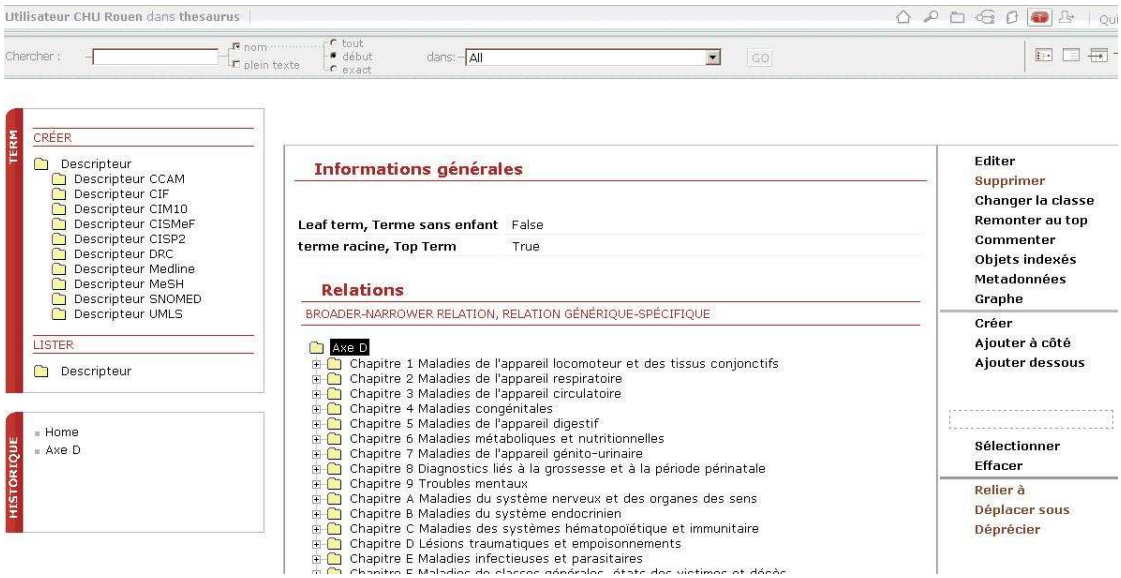

FIG. A.2 – Axe D des maladies classées par chapitre

Le développement de toutes les maladies de la section 3-1 (figure A.3) fait apparaitre toutes les maladies cardiaques recensées dans la SNOMED 3.5.

FIG. A.3 – Les maladies cardiaques dans la SNOMED 3.5

La sélection d'un terme (exemple de la figure A.3 : infarctus aigu du myocarde) fait apparaître tous les connaissances à propos de ce terme. La figure A.4 qui est une capture

d'écran d'une partie de haut de page correspondante à la maladie « infarctus aigu du myocarde » dans SNOMED 3.5, montre quelques informations liées aux termes :

– Le code (le Code CIM-9 lié au code SNOMED est « 410.9 »).

– Le fait que dans la hiérarchie il n'existe pas de terme plus précis.

FIG. A.4 – Haut de la page correspondant à « infarctus aigu du myocarde »

La deuxième partie de la page présentée dans la figure A.5, montre la partie description où on peut voir :

– La localisation de la maladie (« infarctus aigu du myocarde » est localisée au « myocarde » qui est un autre terme SNOMED) ;

– Le type de la maladie « infarctus aigu du myocarde » est de type « infarctus » qui est un autre terme SNOMED).

On peut surtout voir dans la rubrique SNOMED/CIM10 que « infarctus aigu du myocarde » est lié au code CIM10 « I21.9 ».

Ce lien inter-terminologique permet de naviguer entre les deux terminologies SNOMED 3.5 et CIM10 dans le SMTS. La figure A.6 montre le haut de la page correspondant à la description du code CIM10 « I21.9 ». On y trouve des informations spécifiques à la classification CIM10 comme le niveau du code par exemple.

La figure A.7 représente le bas de la page correspondant au code CIM10 « I21.9 », on y trouve tous les termes correspondant dans la CIM10 au code « I29.9 » comme par exemple : « rupture du cœur » ou « infarctus du myocarde ».

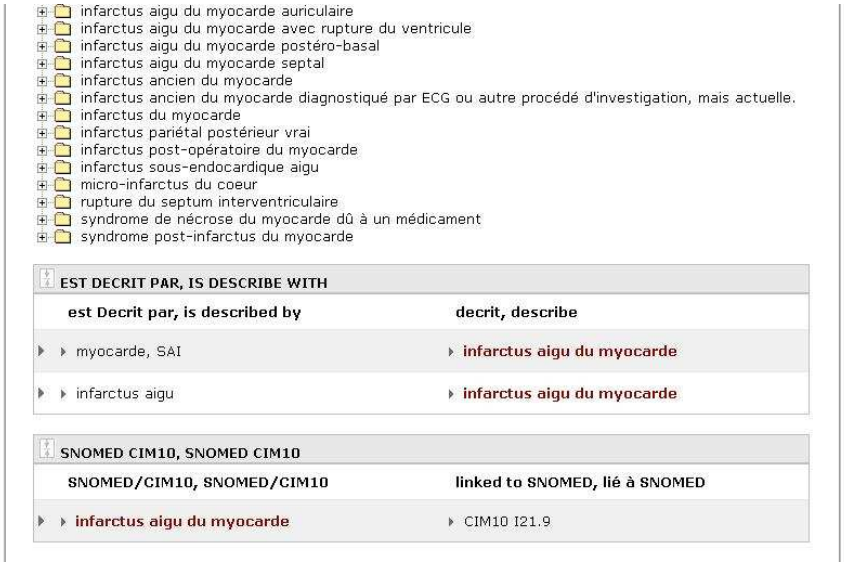

FIG. A.5 – Bas de la page correspondant à « infarctus aigu du myocarde »

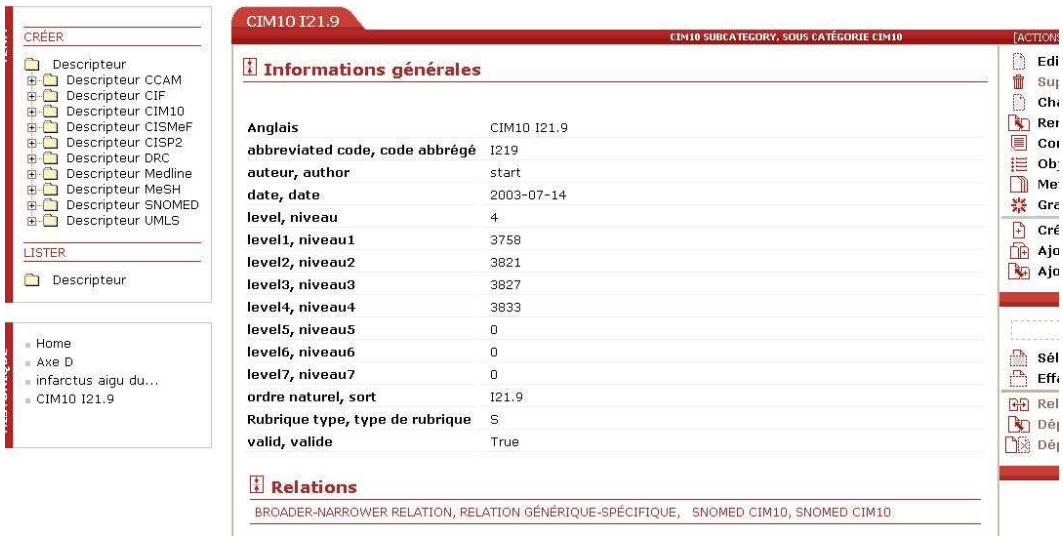

FIG. A.6 – Haut de la page correspondant au code CIM10 121.9

FIG. A.7 – Bas de la page correspondant au code CIM10 I29.9

Annexe B

Étude de cas sur le Portail Terminologique de Santé

Actuellement, 21 terminologies francophones sont intégrées dans le PTS en plus des terminologies déjà présentes dans le SMTS. La figure B.1 montre la page d'accueil avec toutes les terminologies affichées.

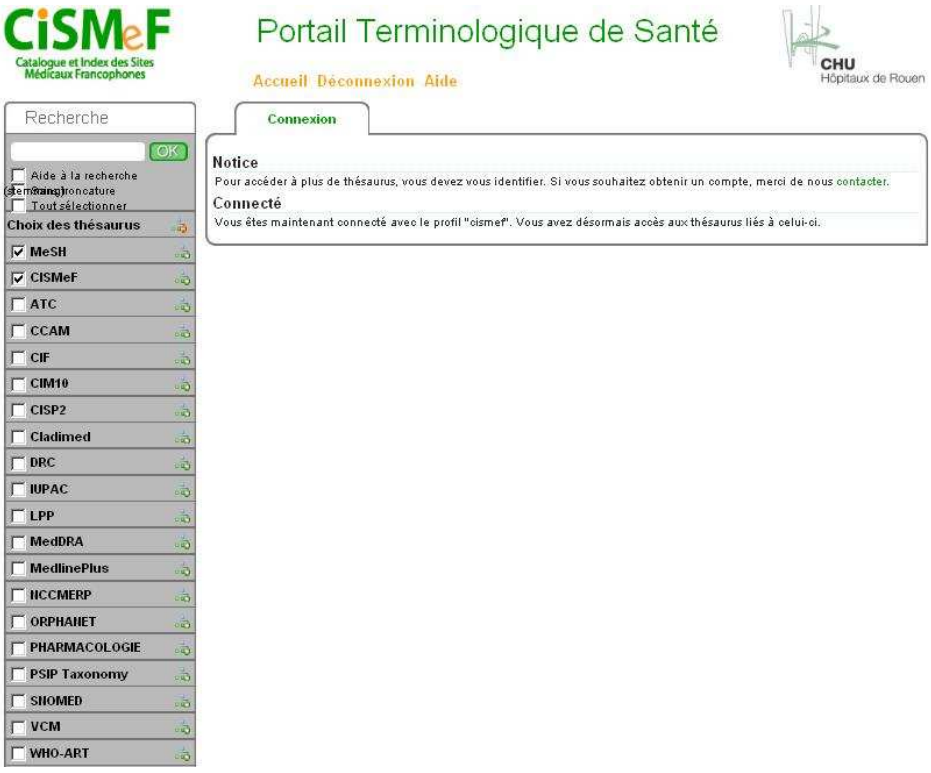

Fig. B.1 – Page d'accueil du PTS

La recherche (multi-)terminologique s'effectue dans le cadre gauche de l'écran. La troncature est activée par défaut, ce qui permet de ne saisir qu'une partie de mot « achondroplasie » par exemple (voir figure B.2). La recherche porte sur la liste des termes et de leurs synonymes de toutes les terminologies, en anglais et en français. En cochant « sans troncature », on recherche le mot exact, seul ou dans une expression. Une autre option est proposée permettant de restreindre la recherche sur une partie des terminologies intégrées dans le PTS. La liste des réponses est présentée à gauche avec pour chaque terminologie, le nombre de résultats trouvés.

FIG. B.2 – Recherche par troncature dans PTS

3 onglets d'informations sont disponibles pour chaque terme :

Description : il contient les généralités concernant le terme (définition(s), synonymes, relations avec d'autres termes...) ;

Hiérarchies : il permet de connaître la position hiérarchique du terme et de naviguer dans les arborescences ;

Ressources : il permet l'accès aux sites et documents de référence CISMeF et PubMed ou « CISMeF InfoRoute » un outil permettant un accès contextuel à plusieurs sites (figure B.3) de santé regroupés par leur contexte d'utilisation : « Outil de recherche », « Médicaments »...

Fig. B.3 – CISMeF InfoRoute

Actuellement dans le PTS, nous avons intégré deux résultats obtenus dans le cadre de cette thèse. Le premier concerne les projections des relations SNOMED CT entre les termes MeSH. La figure B.4 montre un exemple de deux relations intégrées : « localisation » et « association morphologique » pour le terme MeSH : « Angiocholite ».

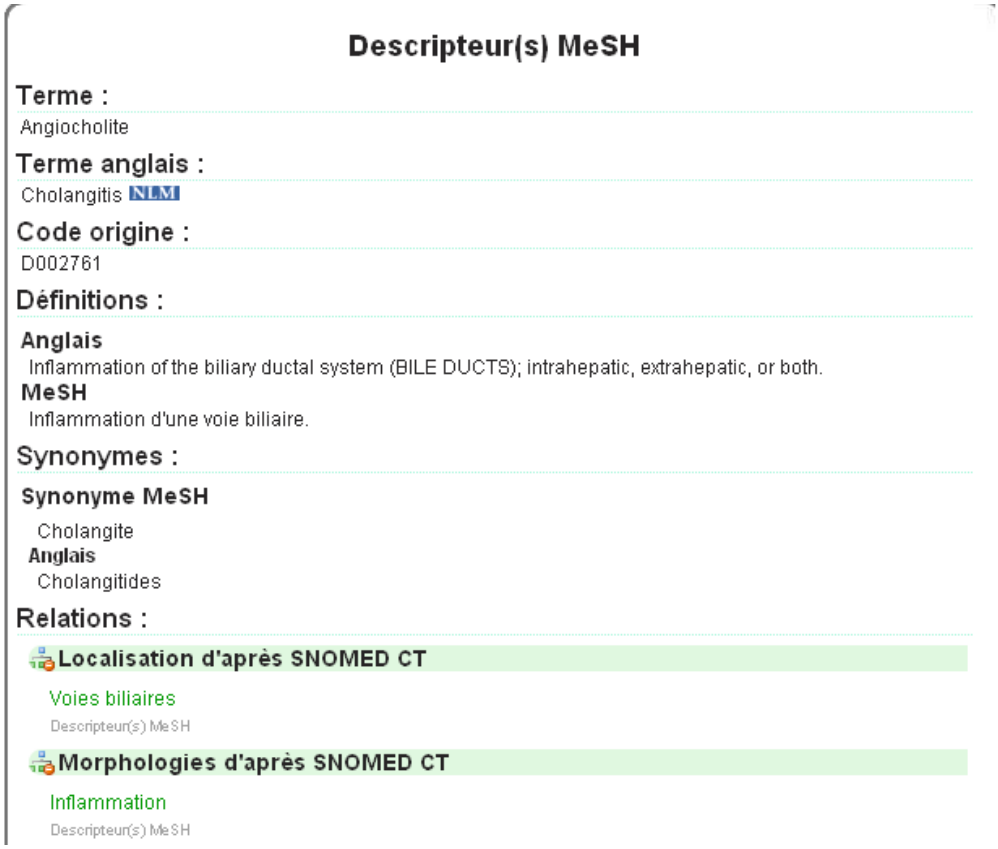

Fig. B.4 – Exemple de deux relations SNOMED CT intégrées dans le PTS

Le deuxième travail concerne tous nos résultats de matching, actuellement dans le PTS deux types de matchings sont intégrés :

1. matching provenant de l'UMLS : ce type de matching correspond au matching conceptuel de l'UMLS. Les termes sont matchés s'ils partagent le même concept UMLS ;
2. matching réalisé par nos méthodes en utilisant UMLS : deux matchings existent actuellement, ORPHANET vers F_UMLS et ATC vers F_UMLS ;
3. le dernier type de matching est réalisé pour toutes les terminologies du PTS (sans utilisation des matchings conceptuels de l'UMLS).

La relation entre les termes matchés utilisant les deux premiers types de matchings est nommée « Correspondance(s) UMLS (même concept) » pour préciser que les mat-

chings ont été obtenus en utilisant UMLS. La figure B.5 montre un exemple de cette relation pour le terme ORPHANET « syndrome de Marfan ».

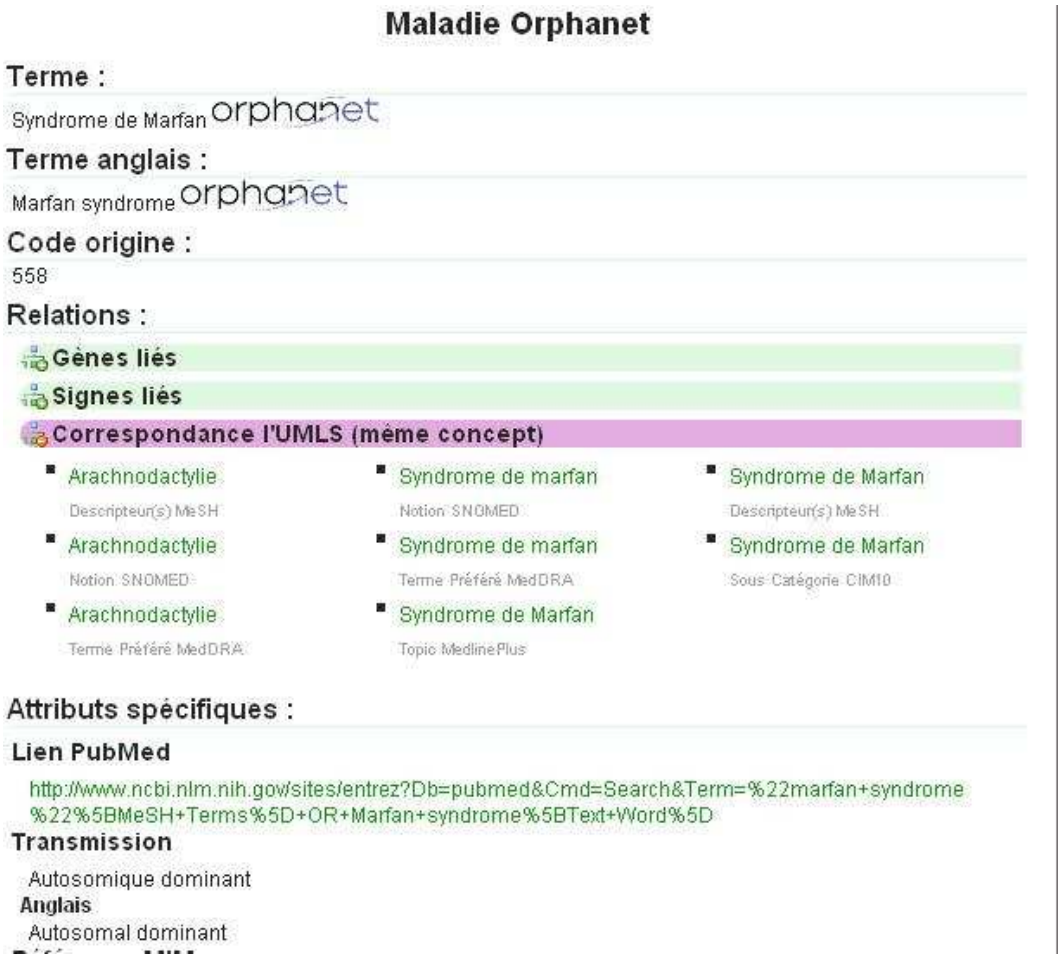

FIG. B.5 – Matching du terme ORPHANET « syndrome de Marfan » vers F_UMLS

Pour la relation entre les termes matchés utilisant le troisième type de matching, elle est nommé dans le PTS « Aligné automatiquement » pour préciser que les matchings ont été obtenus seulement par les méthodes automatiques développées dans le cadre de cette thèse. La figure B.6 montre un exemple de cette relation pour le terme MeSH « infarctus du myocarde ».

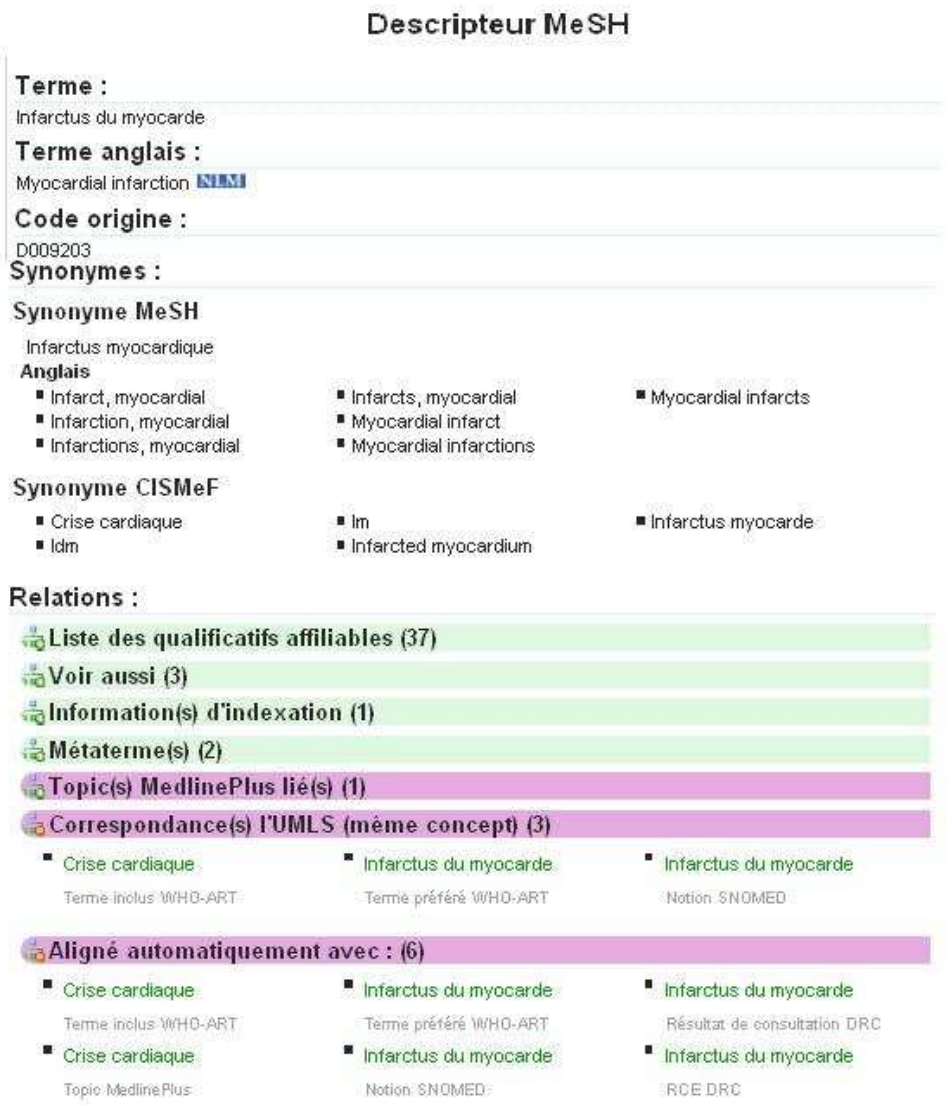

FIG. B.6 – Matching du terme MeSH « infarctus du myocarde »