Alignement entre des ontologies de domaine et la Snomed: trois études de cas

Laurent Mazuel, Jean Charlet

Alignement entre des ontologies de domaine et la Snomed: trois études de cas

Laurent Mazuel et Jean Charlet

INSERM UMR_S 872, Eq. 20
15, rue de l'École de Médecine, 75006 Paris
{Laurent.Mazuel,Jean.Charlet}@spim.jussieu.fr

Résumé : Les expériences sur les ontologies montrent de plus en plus clairement qu'elles ne représentent correctement et de façon consensuelle que des domaines réduits. Ainsi, les ontologies de domaines sont développées pour des applications particulières alors que des ontologies de référence tendent à être utilisées pour fédérer les résultats des applications spécifiques. Nous présentons dans cet article la construction, l'analyse et la discussion d'un alignement entre trois ontologies de domaine construites à l'INSERM UMR_S 872, Éq. 20 (*i.e.* OntoPneumo, OntoHTA et OntoReaChir) et la classification SNOMED v3.5.

Mots-clés : Alignement d'ontologies, ontologie médicale, SNOMED, terminologie de référence, terminologie d'interface.

1 Introduction

Depuis de nombreuses années, la médecine a produit de nombreuses terminologies pour des applications diverses, de tous types, classification, thésaurus et plus récemment des ontologies.

Les expériences sur les ontologies montrent de plus en plus clairement qu'elles ne représentent correctement et de façon consensuelle que des domaines réduits. En suivant ces travaux (Rosenbloom *et al.*, 2006), on parle de *terminologies de référence* et de *terminologies d'interface* : 1) les terminologies de référence ont des visées de représentations larges et de référentiel pour des futures études épidémiologiques. La plus connue est la SNOMED ; 2) les terminologies d'interface sont développées pour des applications spécifiques. On retrouve cette dichotomie au niveau des ontologies où des ontologies d'interfaces contenant des ontologies de domaines sont développées pour des applications particulières alors que des ontologies de référence – *i.e.*, toujours la SNOMED dans sa version ontologique, la SNOMED-CT– tendent à être utilisées pour fédérer les résultats, souvent de l'indexation au sens large, des applications spécifiques.

À l'INSERM UMR_S 872, Éq. 20, nous avons développé un certain nombre d'ontologies pour des applications, souvent d'aide au codage médical, mais aussi pour des motivations de modélisation liées à des études d'usage (Charlet *et al.*, 2008). Il existe par ailleurs une terminologie de référence, la SNOMED v3.5, dont la France a acquis les

droits d'usage pour tout le territoire et qui a vocation à être la terminologie de référence de la santé en France.

Dans ce contexte, les applications médicales diverses qui, pour la plupart encore une fois, indexent des données médicales liées à des patients, ne peuvent prétendre à l'avenir contribuer à l'indexation de données de santé pour des études épidémiologiques que si elles sont alignées avec la SNOMED v3.5[1]. Cet alignement est donc un passage obligé du développement de ces ontologies. C'est lui que nous allons étudier ici en analysant comment trois ontologies développées à l'INSERM UMR_S 872, Éq. 20 s'alignent avec la SNOMED. Ces trois ontologies sont une ontologie de la réanimation chirurgicale, OntoReaChir, une ontologie de la pneumologie, OntoPneumo, et une ontologie de l'hypertension artérielle, OntoHTA.

Dans la section 2, nous décrivons le contexte et l'objectif de ce travail ; dans la section 3, nous décrivons précisément les ontologies et terminologies impliquées dans ce travail ; dans la section 4 nous décrivons la méthode d'alignement mise en œuvre ; dans la section 5 nous présentons et discutons les résultats ; enfin, nous concluons en 6.

2 Contexte et objectif

Notre but va être ainsi de vérifier les possibilités d'alignement, ensuite à quels niveaux ils se font. Ainsi, une ontologie est maintenant classiquement découpée en 3 niveaux :

1. La *top*-ontologie, que l'on devrait plus précisément appeler « ontologie formelle » pour reprendre l'appellation des philosophes. C'est le niveau le plus abstrait structurant les connaissances de haut niveau avec des catégories dont l'organisation dépend de réflexions philosophiques.
2. La *core*-ontologie, fournissant les concepts structurant du domaine et décrivant les relations entre ces concepts – en médecine, on y trouve des concepts de *diagnostic*, *signe*, *structure anatomique* et des relations comme celles liées à la localisation d'une pathologie sur une structure anatomique.
3. L'ontologie du domaine, c'est-à-dire les concepts du domaine tels qu'ils sont manipulés par les professionnels – ici de santé. Dans notre équipe, le troisième et dernier niveau est celui que l'on construit avec les outils de TAL puisque l'on analyse les documents produits en activité avec ceux-ci.

Ce découpage nous servira de caractérisation des ontologies étudiées dont on peut déjà dire qu'elles sont, comme beaucoup d'autres, faites pour des usages spécifiques.

Par rapport à ce découpage et la structure connue des différentes ontologies, nous énonçons les hypothèses suivantes avant expérimentation :

– les feuilles des hiérarchies des ontologies de spécialité seront plus précises que les concepts de la SNOMED ;
– la SNOMED ne contenant pas de *top*-ontologie, la *top*-ontologie des spécialités, si elle existe, ne devrait pas s'apparier ;
– finalement, les appariements devraient se situer au niveau de la *core*-ontologie.

[1] Dans la suite de l'article, sauf précision particulière, nous utiliserons le vocable SNOMED pour la SNOMED v3.5.

Notre objectif est donc de confirmer ou d'infirmer ces hypothèses et, le cas échéant, de découvrir d'autres caractéristiques de ces ontologies.

3 Matériel

Nous décrivons dans cette section les modèles de connaissances que nous considérons dans nos alignements, à savoir la SNOMED d'un coté et les trois ontologies de spécialités de l'autre.

3.1 La classification SNOMED v3.5

La SNOMED v3.5 est une classification multi-axiale standardisée en français, contenant actuellement 116 000 concepts (pour environ 150 000 termes en comptant les synonymes). Elle est organisée en 11 axes (Terminologie, Diagnostic, etc.), chacun de ces axes étant organisé hiérarchiquement. Il existe de plus des liens non hiérarchiques entre ces axes. Par exemple, le concept « hémorragie ombilicale » de l'axe *Fonction* est associé à « ombilic » (axe *Terminologie*) et « hémorragie » (axe *Morphologie*).

La SNOMED-CT, évolution de la SNOMED v3.5[2], est actuellement la plus grande ontologie médicale, avec près de 344 000 concepts. Néanmoins, elle n'est actuellement définie qu'en anglais et n'est ainsi pas applicable dans le cadre de cet article, étant donné que nos ontologies de spécialité sont définies en français.

3.2 L'ontologie OntoPneumo

Le projet OntoPneumo visait à construire une ontologie de la pneumologie pour l'aide au codage médical (Baneyx, 2007). La construction de cette ontologie a fortement utilisé les ressources terminologiques du domaine à modéliser pour rendre compte, le plus précisément possible, non seulement des pratiques médicales actuelles en pneumologie mais également des vocabulaires utilisés par les médecins.

Cette construction est basée sur l'utilisation du logiciel SYNTEX-UPERY (Bourigault & Fabre, 2000) comme outils d'analyse de corpus de texte et de traitement (automatique) du langage pour obtenir un réseau de candidats termes, leurs proximités contextuelles et leurs liens avec le corpus source. L'éditeur DOE[3] a permis de construire l'ontologie selon la sémantique différentielle, les étapes de formalisation et d'opérationalisation étant réalisées à l'aide de l'éditeur d'ontologies PROTÉGÉ[4]. Par ailleurs, l'ontologie a été complétée par une analyse du thésaurus de spécialité de la pneumologie[5]. Ce thésaurus est conçu comme une sous-partie de la CIM-10 réduite aux pathologies de la pneumologie. Ainsi, par construction, les 337 termes de ce thésaurus sont bien inclus dans OntoPneumo. Pour finir, l'ontologie a été validée par un expert du domaine, médecin en milieu hospitalier dans un service de pneumologie.

[2]Les liens vers les concepts originaux SNOMED v3.5 sont d'ailleurs accessibles.

[3]The Differential Ontology Editor, `http://homepages.cwi.nl/~troncy/DOE/`

[4]`http://protege.stanford.edu/`

[5]Disponible sur le site de la Société de Pneumologie de Langue Française : `http://www.splf.org`

Cette ontologie du domaine compte actuellement 1 114 concepts, mais sans l'utilisation d'une *top*-ontologie (*i.e.* actuellement OntoPneumo est constitué de 25 arbres disjoints hiérarchiquement). En effet, l'intégration de la *top*-ontologie et de la *core*-ontologie du projet MENELAS[6] devrait mieux organiser la hiérarchie et ajouter environ 400 concepts. L'ontologie définit également une hiérarchie de 27 relations.

3.3 L'ontologie de l'hypertension artérielle, OntoHTA

OntoHTA est issue d'un projet de recherche sur les déterminants du raisonnement médical qui a abouti à la construction d'une première ontologie et a déjà eu pour effet de proposer une mise à jour des formulaires d'entrées de données cliniques dans le domaine de l'hypertension artérielle. Cette ontologie est en cours de construction par un médecin spécialiste (Steichen *et al.*, 2007) en tenant en partie compte de la SNOMED-CT, en particulier pour les termes associés aux concepts en anglais.

Comme dans le projet OntoPneumo, les outils de traitement automatique du langage SYNTEX et UPERY ont été choisis pour l'analyse des corpus (commentaires en texte libre et guides de bonne pratique). La modélisation ontologique a été réalisée, concept par concept, dans l'éditeur DOE.

Actuellement cette ontologie est une monohiérarchie strictement taxinomique, dans le respect des principes de la sémantique différentielle, et qui organise 506 concepts. Cette ontologie bénéficie d'une *top*-ontologie articulant l'ensemble des concepts.

3.4 L'ontologie de la réanimation chirurgicale, OntoReaChir

La réanimation chirurgicale est un domaine médical spécialisé dans la prise en charge des complications postopératoires et dans la traumatologie. Comme dans les deux ontologies précédentes, la base de l'ontologie a été construite à partir de corpus (800 comptes rendus hospitaliers) sur le logiciel SYNTEX-UPERY (Le Moigno *et al.*, 2002). Par ailleurs, l'élément de référence utilisé pour l'évaluation de l'ontologie est la version du thésaurus de spécialité — correspondant à peu près au thésaurus de la CIM-10 – émise en 1999.[7]

Ceci a abouti à une ontologie constituée d'une hiérarchie taxinomique de 2 039 concepts, ainsi que d'une hiérarchie de 200 relations. Cette ontologie possède une *top*-ontologie très détaillée ne correspondant pas à une *top*-ontologie spécifiée préalablement mais proche de celle de MENELAS (Charlet *et al.*, 1996). La partie basse est ainsi celle qui correspond le plus au thésaurus initial du domaine. OntoReaChir a été récemment reprise pour être décrite formellement en OWL.

4 Méthode

Cette section décrit la méthode que nous avons utilisée pour produire les alignements entre la SNOMED et nos trois ontologies. Dans une première section, nous justifierons

[6] `http://estime.spim.jussieu.fr/Menelas/`

[7] Version disponible sur le site de la société française d'anesthésie et de réanimation : `www.sfar.org`.

nos choix pour la production de ces alignements. Dans un deuxième temps, nous présenterons en détail notre méthode semi-automatique.

4.1 Méthodes d'alignements applicables pour nos ontologies

Il existe différentes méthodes d'alignement, plus ou moins utilisables en fonction des situations et des formalismes considérés pour les ontologies (Euzenat & Shvaiko, 2007). Dans notre situation, l'absence d'instances aussi bien dans la SNOMED que dans nos trois ontologies supprime les possibilités liées à ce type de méthode (Ichise *et al.*, 2003). De même, les approches demandant une troisième ontologie (avec un rôle de *médiateur* (Aleksovski *et al.*, 2006)) ne sont pas applicables, car la SNOMED est elle-même une ontologie médiatrice (de part son aspect générique et sa grande couverture générale du domaine). Les méthodes d'alignement structurel (Breitman *et al.*, 2005) reposent sur l'idée que si deux noeuds sont alignés, alors leurs ancêtres et leurs enfants doivent s'aligner mutuellement entre eux (cette hypothèse permettant à la fois de tester la cohérence des propositions d'appariements ainsi que d'en proposer de nouveaux). La structure particulière de nos ontologies (*e.g.* l'absence de *top*-ontologie pour OntoPneumo), ainsi que la structure de la SNOMED (qui bien qu'organisée hiérarchiquement n'est pas parfaitement analogue à une hiérarchie de subsomption comme dans les ontologies) font qu'il nous était difficile d'appliquer cette méthode de manière automatique. Néanmoins, nous avons tenu compte manuellement, lorsque cela était possible, de la cohérence structurelle des alignements proposés.

Finalement, nous ne pouvons utiliser que les méthodes morpho-syntaxiques pour produire automatiquement notre alignement. L'alignement morpho-syntaxique consiste à chercher un alignement entre concepts en ne se basant que sur les labels des termes. Cette étape utilise différents outils de TAL pour proposer des alignements basés sur la correspondance entre les chaînes de caractères représentant les concepts. Toute méthode d'alignement commence toujours par une première étape de ce type, afin de fournir un alignement initial de travail. Ainsi, notre calcul initial réside aussi sur ce type d'algorithmes. D'autre part, la complétude et la vérification de l'alignement sont effectuées manuellement. La section suivante décrit en détail ces processus.

4.2 Méthode d'alignement utilisé

Cette section décrit la méthode que nous avons utilisée pour calculer notre alignement. Dans un premier temps, nous présenterons les méthodes automatiques mises en place pour produire une proposition d'alignement initial, puis les méthodes de vérification et complétion manuelles.

4.2.1 Description de la fonction d'appariement morpho-syntaxique

Notre approche morpho-syntaxique se découpe en deux parties. Tout d'abord, nous normalisons et simplifions les chaînes de caractères représentant les concepts pour les ramener à des formats équivalents. Ensuite, nous utilisons la distance de Levenshtein et

la distance de Stoilos *et al.* (2005)[8] pour augmenter la portée des propositions d'alignements.

Dans la partie sur la normalisation syntaxique, nous effectuons plusieurs opérations successives :

- retirer les diacritiques (*i.e.* les accents, cédilles, etc.). Ceci peut poser quelques problèmes d'ambiguïtés que nous résolverons dans la partie manuelle (*e.g.* les termes « côte - coté » ou « aine - aîné » ne sont plus distinguables sans diacritiques) ;
- réduire l'ensemble de la casse aux lettres minuscules ;
- normaliser le nombre d'espaces et les remplacer par le caractère « _ » ;
- supprimer les mots de liaisons pour ne garder que les noms (*e.g.* suppression de « l' », « le », « la », etc.) ;
- pour la SNOMED, retirer certains suffixes inutiles pour l'alignement comme par exemple la chaîne « SAI » (*i.e.* Sans Autre Indication).

Nous utilisons ensuite la distance de Levenshtein normalisée[9] avec un seuil à $0,97$, ce qui est (empiriquement) suffisant pour rattraper une partie des fautes d'orthographe sans impliquer un trop grand nombre d'erreurs d'alignement. Nous complétons ce premier alignement par un calcul suivant la distance de Stoilos, avec un seuil à $0,9$. Les deux étapes sont complémentaires : la distance de Levenshtein est robuste face aux fautes d'orthographes alors que la distance de Stoilos est plus efficace pour l'analyse des sous-chaînes de caractères.

4.2.2 Validation et complétion manuelles des alignements

Malgré les méthodes précédentes, un alignement à la main est nécessaire pour compléter et valider l'alignement obtenu, par exemple :

- pour les abréviations : « ALAT » $\equiv$ « aspartate aminotransférase », « AVC » $\equiv$ « accident vasculaire cérébrale » ;
- pour les différences de notation : « Dosage de la ... » $\equiv$ « Mesure de la ... », « Syndrome de ... » $\equiv$ « Maladie de ... » ;
- pour les erreurs induites par les mesures : Par exemple, l'alignement entre « Dosage du facteur X » de OntoPneumo et « Dosage du facteur V » de la SNOMED.

Pour simplifier notre travail, nous avons développé un outil permettant facilement d'aligner une ontologie avec la SNOMED (figure 1). Ce logiciel affiche sur la partie gauche de l'écran la hiérarchie de l'ontologie considérée avec un code de couleur en fonction de la saisie de l'utilisateur dans la partie droite :

- le vert lorsqu'un équivalent existe. Il est alors noté dans la colonne de droite ;
- le rouge lorsqu'il n'existe pas d'équivalent du concept dans la SNOMED ;
- le jaune lorsque le concept n'a pas actuellement de statut.

Cette interface simple permet d'éditer les alignements, de récupérer les statistiques de chaque nœud (*i.e.* obtenir pour le sous-arbre de chaque nœud le nombre d'appariements effectués, le nombre de concepts notés sans appariements possibles et calculer les proportions respectives) et offre un outil de visualisation permettant de mieux appréhender la répartition des alignements en fonction des branches de l'ontologie considérée.

[8] Nous remercions le relecteur qui nous a suggéré cette distance pour améliorer cette étape.

[9] Dans l'intervalle $[0, 1]$, 1 pour des chaînes identiques et 0 pour des chaînes totalement différentes.

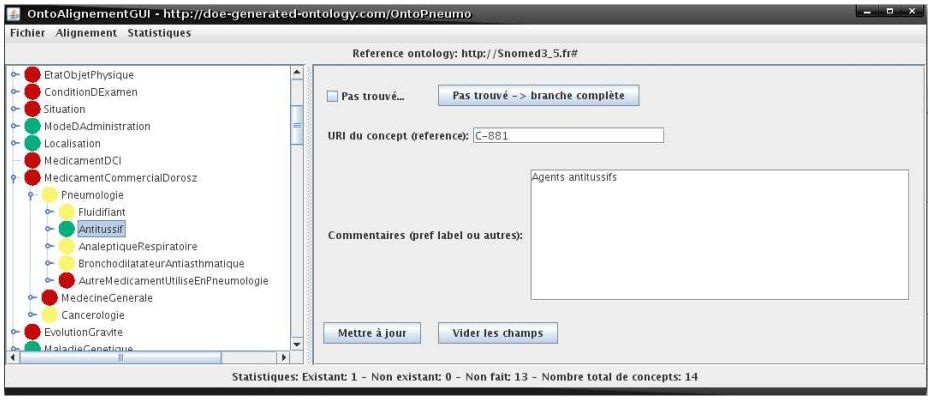

FIG. 1 – Logiciel d'alignement entre la SNOMED et une ontologie, ici OntoPneumo.

Ontologie	Nb. concepts	Nb. appariements automatiques		Nb. final appariements
		Directs	Validés	
OntoPneumo	1114	669	613	787
OntoHTA	506	159	144	228
OntoReaChir	2039	1046	987	1187

TAB. 1 – Résultats des alignements.

Cette partie est importante, l'alignement manuel effectué en complément permet d'augmenter sensiblement le nombre d'appariements (pratiquement le double pour HTA).

5 Résultats et discussion

Cette section présente les alignements obtenus entre nos trois ontologies et la SNOMED (tableau 1 et figure 2). Nous commencerons par discuter des alignements ontologie par ontologie, puis nous conclurons avec une discussion générale de ces résultats.

5.1 Résultats par ontologies de spécialité

5.1.1 OntoPneumo

OntoPneumo est l'ontologie de spécialité parmi les trois que nous étudions qui s'aligne le mieux avec la SNOMED (avec 75% de recouvrement). Ceci peut s'expliquer pour deux raisons :

– OntoPneumo ne contient pas de *top*-ontologie, ses concepts « hauts » s'apparient

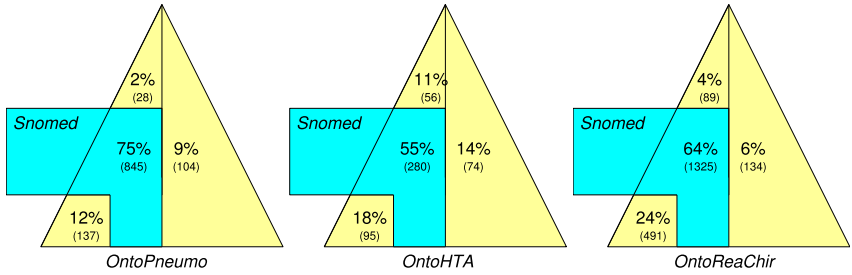

FIG. 2 – Schéma de la répartition moyenne des alignements pour les trois ontologies de spécialités. La partie droite représente la proportion de sous-arbres complets (*i.e.* de la racine jusqu'aux feuilles) sans aucun alignement. En plus des concepts qui sont directement alignés, la partie « SNOMED » inclut aussi tous les concepts qui ont à la fois un ancêtre *ET* un descendant alignés, et sont donc conceptuellement inclus dans la SNOMED (ce qui explique pourquoi ce nombre est plus grand que le nombre d'alignements noté au tableau 1).

donc bien avec la SNOMED, car ils ne sont pas très génériques (seul 2% des concepts hauts de OntoPneumo ne s'alignent pas avec la SNOMED, figure 2).

– OntoPneumo inclut le thésaurus de spécialité de la pneumologie (*cf.* section 3.2). Or, ce thésaurus est construit en utilisant la CIM-10 qui est elle-même incluse dans dans la SNOMED.

Ainsi, les axes de diagnostics sont pratiquement complètement représentés dans la SNOMED. Par contre, dans le cas de diagnostics composés, la SNOMED est assez faible. On entend par « diagnostic composé » les diagnostics liant une pathologie à une partie du corps, tel que « mésenthéliome pleural », « ablation du pouls carotidien » ou encore « granulome hyalin pulmonaire ». En effet, la SNOMED définit de manière générale une pathologie mais ne propose pas d'instances en fonction de l'anatomie sous-jacente. Ainsi, les pathologies pulmonaires de OntoPneumo existent dans la SNOMED, mais sous leur forme générale. Ceci implique que pour ce type de sous-arbres, les feuilles de l'ontologie de spécialité ne s'apparient pas dans la SNOMED (12% pour OntoPneumo), mais que les concepts plus hauts (et plus génériques), oui. A l'inverse, dans le cas de pathologies annexes (cas des maladies cardio-vasculaire ou psychiatrique par exemple), la SNOMED est alors totalement suffisante.

La SNOMED n'ayant pas vocation à coder les concepts non médicaux, les sous-arbres traitant de « rôle hospitalier » comme « médecin » ou « infirmière », ou traitant d' « instrument » comme « bistouri » ou « sonde », ainsi que les sous-arbres couvrant les médicaments ne sont absolument pas alignables avec la SNOMED[10]. Ceci est représenté par les 9% de concepts inclus dans des sous-arbres totalement non alignables.

[10]Le cas des médicaments est néanmoins particulier, car la SNOMED définit les principes actifs (« acide acétylsalicylique » ou « paracétamol »), mais pas les dénominations commerciales (« Aspegic », « Doliprane »). Un alignement est donc envisageable, mais sans équivalence stricte.

5.1.2 OntoHTA

L'ontologie de l'hypertension artérielle est celle qui s'aligne le moins bien avec la SNOMED. La raison principale réside dans la différence de la sémantique de spécialisation de la hiérarchie de subsomption. Autrement dit, les liens hiérarchiques de OntoHTA sont souvent assez peu compatibles avec les liens hiérarchiques de la SNOMED. Par exemple, les procédures de OntoHTA sont triés en « procédure par appareil », « procédure par pathologie », etc. alors que les procédures de la SNOMED sont classées par spécialités médicales, « procédure dentaire », « procédure psychiatrique », etc. ce qui implique que beaucoup de concept sont proches mais pas équivalents.

Comme dans OntoPneumo, les parties liées à la spécialité étudiée sont plus détaillées que dans la SNOMED. Par exemple, le concept « abolition du pouls » (présent dans les deux modèles de connaissances), n'est pas spécialisé dans la SNOMED, alors que OntoHTA la spécialise 18 fois en fonction des pouls possibles (*e.g.* « abolition du pouls carotidien », « abolition du pouls pédieux », etc.). Un autre point intéressant concerne la révision possible de OntoHTA grâce à l'alignement obtenu ou le repérage d'imprécisions dans la SNOMED. Par exemple, les concepts « polykystose rénale » et « maladie kystique congénitale du rein » sont considérés comme synonymes dans la SNOMED, alors qu'ils sont pères et « fils unique »[11] dans OntoHTA. Cette particularité peut dénoter une incomplétude d'OntoHTA ou une imprécision de la SNOMED, assumée ou pas par les constructeurs.

5.1.3 OntoReaChir

OntoReaChir possède une très grande *top*-ontologie, très détaillée. Il faut ainsi en moyenne parcourir 7 à 8 nœuds de profondeur pour arriver sur un nœud proche de la conceptualisation de la SNOMED.[12] La *top*-ontologie s'aligne ainsi assez peu. De plus, 9% des arbres sont totalement non alignables avec la SNOMED (de la racine aux feuilles), décrivant des concepts abstraits, utiles pour la définition de concepts définis, non applicables à la SNOMED.

Malgré tout, cette ontologie a été construite dans le but de couvrir les concepts du thésaurus de la spécialité du domaine. Or, à l'instar de OntoPneumo, ce thésaurus utilise les concepts de la CIM-10, couverts par la SNOMED. Ceci implique qu'un grand nombre de concepts de OntoReaChir s'apparie avec la SNOMED (64% de recouvrement, soit 1325 concepts). Ainsi, au final, la répartition des alignements est sensiblement équivalente à la répartition de OntoPneumo, malgré le nombre extrêmement différent de concepts.

On constate aussi des classements hiérarchiques très différents d'avec la SNOMED, impliquant une non-cohérence de la structure de l'alignement. Par exemple, la SNOMED classe les os par nom, puis les sous-classes représentant les parties de ces os (*e.g.* le concept « humérus » a comme fils « épiphyse de l'humérus »). Dans OntoReaChir, c'est l'inverse, les os sont classés par type de partie, puis instanciés au vrai os (*e.g.* le concept « épiphyse » a comme fils « épiphyse de l'humérus »).

[11] Nous entendons par « fils unique » un concept sans frères et étant ainsi l'unique fils du nœud initial.

[12] Exemple de descente dans la hiérarchie à partir de la racine : « Concept », « ObjetAbstrait », « ObjetDescription », « ObjetFonction », « ObjetFonctionPhysique », « ObjetFonctionPhysiquePathologie », ...

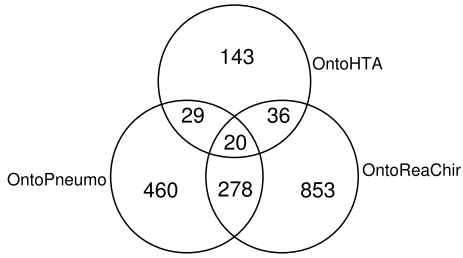

FIG. 3 – Recoupement et répartition des alignements obtenus entre les trois ontologies spécifiques étudiées.

5.1.4 Recoupement entre les alignements des ontologies de spécialité

Comme nous l'évoquions dans la section 4.1, ces alignements à la SNOMED peuvent permettre d'aligner aussi les ontologies spécifiques entre elles. Nous avons donc étudié la répartition des alignements communs à plusieurs ontologies spécifiques (figure 3).

Il y a 20 alignements communs aux trois ontologies. Il est intéressant de constater que ces concepts communs à toutes les ontologies tournent autour de la description de l'état de santé du patient (*e.g.* « tension artérielle »), de la description d'examen générique (*e.g.* « echocardiographie », « échographie Doppler ») et de la description des symptômes (*e.g.* « AVC »).

Il existe 56 alignements communs entre OntoHTA et OntoReaChir (36+20). Les alignements communs supplémentaires tournent autour des diagnostics sur le rein (*e.g.* « néphropathie ») et les maladies cardiaques (*e.g.* « souffle systolique »). Pour les 49 alignements communs entre OntoHTA et OntoPneumo, les alignements communs sont majoritairement issus des examens sanguins standards (*e.g.* « dosage du cholestérol LDL », « dosage du potassium ») ainsi que les maladies emboliques (*e.g.* « thromboembolisme aortique »).

Enfin, il existe un nombre étonnement grand d'alignements communs entre OntoPneumo et OntoReaChir (au total 298 alignements). La majeure partie (131) de ces alignements communs concerne la description morphologique du corps humain (*e.g.* « jambe » ou « prostate »). Le reste est divisé de manière a peu près équitable entre la définition d'organismes néfastes (*e.g.* « virus de l'hépatite C »), d'examen sanguin (*e.g.* « ASAT ») et de pathologique pneumologique (*e.g.* « détresse respiratoire aiguë »).

L'étude de ces alignements communs montre que même dans une ontologie de spécialité donnée il reste des informations issues de la médecine généraliste. Cette conclusion tend à montrer l'importance de l'appariement entre des classifications généralistes comme la SNOMED et les ontologies spécifiques.

5.2 Discussion

La SNOMED est une classification générique. En ce sens, elle manque de précisions pour énoncer les spécialité d'un domaine (*e.g.* les pathologies pulmonaires dans OntoPneumo). D'autre part, elle manque parfois de granularité, et deux concepts père/fils de la

SNOMED peuvent facilement être séparés par plusieurs descendants dans les ontologies de spécialité. Par exemple, dans l'ontologie de la réanimation chirurgicale, le concept « thorax » possède le fils « hémi-thorax » qui lui-même possède deux fils « hémi-thorax droit » et « hémi-thorax gauche ». Dans la SNOMED, « hémi-thorax droit » et « hémi-thorax gauche » sont directement des fils de « thorax ». Autre exemple tiré de OntoPneumo, le concept « lobectomie » est défini et possède plusieurs fils, alors que dans la SNOMED ce concept n'existe pas, mais il existe directement les opérations précises comme « lobectomie thyroïdienne unilatérale » (22 définitions de « lobectomie » au total !). Ceci tend à confirmer l'importance de la définition d'ontologies de domaines, et non d'une utilisation directe d'un modèle dit générique tel que la SNOMED.

Un autre exemple pour étayer cette conclusion réside non plus dans la structure, mais dans le vocabulaire. Celui des ontologies de spécialité, hérité de comptes rendus hospitaliers, est plus proche du véritable vocabulaire médical que celui de la SNOMED. Par exemple, la SNOMED ne définit aucune abréviations, là ou les abréviations tels que « AVC », « ASAT », « ALAT », sont couramment utilisées en clinique et apparaissent dans les ontologies de spécialité.

L'une des autres remarques concerne l'utilisation des thésaurus de spécialité spécifiés à partir de la CIM-10. Cette situation, alliée au fait que la SNOMED contient totalement la CIM-10, simplifie et optimise grandement les alignements. Ainsi, dans un objectif d'utilisation pratique de la SNOMED, son rapport à la CIM-10 permet d'envisager des applications d'aide au codage médical beaucoup plus précises que précédemment.

6 Conclusion

Nous avons présenté dans cet article les résultats d'une étude de cas sur l'alignement de trois ontologies de domaine (OntoPneumo, OntoHTA et OntoReaChir) avec la classification SNOMED v3.5. La répartition de ces alignements tend à montrer l'utilité de ces ontologies de domaine par rapport à l'utilisation directe d'un modèle dit générique. En effet, la granularité interne est plus adaptée dans les ontologies de spécialité, ainsi que le niveau de détail des concepts les plus spécifiques. D'autre part, le formalisme d'ontologie est plus complet et pensé pour la définition de concepts définis issus d'une étape de post-coordination, ce qui implique que des blocs entiers ontologies de spécialités ne soient absolument pas représentés dans la SNOMED.

Une évaluation naturelle de ce travail serait d'étudier les alignements avec la SNOMED-CT. La SNOMED-CT possédant les liens vers la SNOMED v3.5, nos alignements sont a priori directement récupérables comme base d'étude. D'autre part, l'ontologie OntoHTA a été construite sur la base de la SNOMED-CT (à l'inverse des deux autres ontologies qui ont plutôt utilisé la CIM-10). Ainsi, il est fort probable que cette étape fournisse des résultats d'étude très intéressants.

Deuxièmement, les ontologies de domaine de cet article ne possèdent pas de concepts définis issus de la post-coordination ou s'il y en a – c'est le cas pour OntoPneumo–, ils n'ont pas été utilisés. Or, il pourrait être intéressant d'étudier comment ces concepts particuliers peuvent s'aligner avec la SNOMED, et quelles en seraient les conséquences.

Enfin, la difficulté des alignements additionnés de la nécessité de développements d'ontologies de domaines spécialisés pousse à envisager l'alignement avec la SNOMED

durant la construction de la ressource elle-même.

Remerciements

Nous remercions particulièrement Audrey Baneyx et Olivier Steichen de leur participation à la lecture et à l'interpretation des résultats correspondant à leurs ontologies.

Références

ALEKSOVSKI Z., TEN KATE W. & VAN HARMELEN F. (2006). Exploiting the structure of background knowledge used in ontology matching. In *Proc. Workshop on Ontology Matching in ISWC2006* : CEUR Workshop Proceedings.

BANEYX A. (2007). *Construire une ontologie de la pneumologie : aspects théoriques, modèles et expérimentations.* PhD thesis, Université Pierre et Marie Curie (Paris VI). Disponible à `http://tel.archives-ouvertes.fr/tel-00136937/fr/`.

BOURIGAULT D. & FABRE C. (2000). Approche linguistique pour l'analyse syntaxique de corpus. *Cahiers de Grammaires*, (25), 131–51. numéro spécial « sémantique et corpus ».

BREITMAN K., FELICÍSSIMO C. & CASANOVA M. (2005). CATO–A Lightweight Ontology Alignment Tool. *Proc. 17th Conf. on Advanced Information Systems Engineering (CAISE'05)*.

CHARLET J., BACHIMONT B., BOUAUD J. & ZWEIGENBAUM P. (1996). Ontologie et réutilisabilité : expérience et discussion. In N. AUSSENAC-GILLES, P. LAUBLET & C. REYNAUD, Eds., *Acquisition et ingénierie des connaissances : tendances actuelles*, chapter 4, p. 69–87. Cepaduès-éditions.

CHARLET J., BANEYX A., STEICHEN O., ALECU I., DANIEL C., BOUSQUET C. & JAULENT M.-C. (2008). Utiliser et construire des ontologies en médecine : Le primat de la terminologie. *Techniques et Sciences Informatiques. À paraître*.

EUZENAT J. & SHVAIKO P. (2007). *Ontology matching*. Heidelberg (DE) : Springer-Verlag.

ICHISE R., TAKEDA H. & HONIDEN S. (2003). Integrating multiple internet directories by instance-based learning. *Proceedings of the eighteenth International Joint Conference on Artificial Intelligence (IJCAI03)*.

LE MOIGNO S., CHARLET J., BOURIGAULT D. & JAULENT M.-C. (2002). Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale. In B. BACHIMONT, Ed., *Actes des 6 ^{es} Journées Ingénierie des Connaissances*, p. 229–38, Rouen, France.

ROSENBLOOM S. T., MILLER R. A. & JOHNSON K. B. (2006). Interface terminologies : facilitating direct entry of clinical data into electronic health record systems. *J Am Med Inform Assoc*, **13**(3), 277–88.

STEICHEN O., DANIEL-LE BOZEC C., JAULENT M.-C. & CHARLET J. (2007). Construction d'une ontologie pour la prise en charge de l'hypertension artérielle. In F. TRICHET, Ed., *Actes des 18 ^{es} Journées Ingénierie des Connaissances*, p. 241–52, Grenoble, France : Cépaduès. ISBN 978.2.85428.790.5.

STOILOS G., STAMOU G. & KOLLIAS S. (2005). A string metric for ontology alignment. *Lecture notes in computer science*, **3729**, 624.