Construire une ontologie de la Pneumologie Aspects théoriques, modèles et expérimentations

Audrey Baneyx

UNIVERSITÉ PIERRE ET MARIE CURIE - PARIS 6

THÈSE DE DOCTORAT DE L'UNIVERSITÉ PARIS 6

SPÉCIALITÉ : INFORMATIQUE MÉDICALE

Pour l'obtention du grade de

DOCTEUR DE L'UNIVERSITÉ PARIS 6

Thèse de doctorat présentée et soutenue publiquement le 06 février 2007 par

Audrey Baneyx

CONSTRUIRE UNE ONTOLOGIE DE LA PNEUMOLOGIE

ASPECTS THÉORIQUES, MODÈLES ET EXPÉRIMENTATIONS

Composition du jury :

Madame	Nathalie	AUSSENAC-GILLES	Rapporteur
Monsieur	Stefan J.	DARMONI	Rapporteur
Madame	Anne	DOUCET	Examinateur
Madame	Marie-Christine	JAULENT	Examinateur
Monsieur	Bruno	BACHIMONT	Examinateur
Monsieur	Jean	CHARLET	Directeur de thèse

LABORATOIRE INSERM UMR_S 872 – SANTÉ PUBLIQUE ET INFORMATIQUE MÉDICALE

INTITULÉ ET ADRESSE DE L'UNITÉ OÙ LA THÈSE A ÉTÉ PRÉPARÉE :

INSERM UMR_S 872
Équipe 20
Santé Publique et Informatique Médicale
Centre de recherche des Cordeliers
15, rue de l'École de Médecine
75006 Paris – France

À Julien

Remerciements

Au-delà de la formalité d'usage, c'est avec un grand plaisir que je remercie les membres de mon jury :

Monsieur Jean Charlet, mon directeur de thèse, pour m'avoir proposé cette thèse et m'avoir permis de la faire dans les meilleurs conditions qui soient, pour nos nombreuses discussions et pour m'avoir présentée à ses amis de la communauté IC. Je lui adresse également un grand merci pour m'avoir fait profiter de ses talents de typographe et ses coups de main en LaTeX. Il est évident que sans lui cette thèse serait moins bien présentée. Il m'a beaucoup appris. J'espère avoir toujours autant de volonté et d'enthousiasme que lui pour mener mes recherches futures, qu'il trouve dans ces quelques mots l'expression de ma gratitude.

Madame Marie-Christine Jaulent pour son bouillonnement perpétuel d'idées et son enthousiasme communicatif, pour avoir toujours pris le temps de relire mes articles, les bons comme les moins bons, et pour savoir créer une ambiance de travail conviviale et motivante. Ce n'est pas donné à tout le monde.

Madame Nathalie Aussenac-Gilles, a accepté d'être le premier rapporteur de mon travail de thèse. Elle a relu avec attention ce mémoire et ses remarques, pertinentes et constructives, m'ont permis de regarder mes contributions aux domaines de l'Ingénierie des connaissances et de l'Informatique médicale d'un œil neuf.

*Le docteur Stefan Darmoni, a bien voulu être le second rapporteur de ce travail. Je le remercie de m'avoir donné son point de vue de médecin sur certains des aspects de ce travail. Ses encouragements quant aux résultats de Med*CKARe *et ses suggestions de collaborations pour mes futurs travaux de recherche sont une grande motivation.*

En 2002-2003, Madame Anne Doucet m'a accueillie dans le DESS d'Intelligence artificielle qu'elle dirigeait à l'université Paris 6. Cette année d'études a sans aucun doute été la plus intéressante et la plus enrichissante pour moi. Je suis particulièrement heureuse qu'elle ait accepté d'examiner mon travail.

Les travaux de monsieur Bruno Bachimont ont servi de point de départ à la partie méthodologique de mes recherches. Ses recherches en Ingénierie ontologique et les liens qu'il fait avec la philosophie sont, pour moi, une source d'inspiration. Je le remercie d'avoir bien voulu juger mon travail.

Je tiens également à remercier chaleureusement :

Monsieur Didier Bourigault qui m'a permis d'utiliser son outil SYNTEX-UPERY *et d'obtenir les résultats de qualité sur lesquels j'ai fondé mes travaux de recherche.*

Le Dr F-X Blanc, le Pr B. Housset et le Pr T. Similowski pour leur participation au projet PERTOM*ed et le Pr B. Maitre, le Dr N. Roche, le Pr C. Chouaid, le Pr J. Cadranel, le Pr M. Humbert, et le Dr A. Duguet, pour avoir rassemblé les ressources nécessaires à mon travail.*

Ce travail doit également beaucoup aux nombreux échanges avec mes collègues (ex-) doctorants rencontrés deci delà. J'ai une pensée particulière pour Véronique Malaisé, Natalia Grabar, Nadia Nadah et Sandra Bringuay que je remercie de leurs encouragements et de leur bonne humeur.

Un clin d'œil à tous mes copains thésards et post-doctorants de Paris, de Marseille et d'ailleurs . . . et aux autres aussi, parce qu'il parait qu'il n'y a pas que les thèses dans la vie !

Un grand merci à ma famille, et en particulier à mes parents, pour leur soutien tout au long de ma scolarité, leurs enseignements et leur confiance.

J'adresse une mention toute particulière à Simon qui m'a tenu compagnie tout l'été et motivé chaque jour pour écrire ce manuscrit dans les temps. Je souhaite que ce qu'il fera plus tard l'intéresse autant que moi.

Les mots de la fin sont pour exprimer toute ma reconnaissance à Julien pour m'avoir toujours soutenue dans mes choix, pour nos discussions sans fin, parce qu'il regarde dans la même direction que moi et bien plus encore . . .

Table des matières

CHAPITRE 1

Introduction générale

« Où que tu sois, creuse profond. »
Friedrich Nietzsche
Le gai savoir (1882)

Ce chapitre est une introduction à notre mémoire de thèse. Nous y abordons successivement les conditions administratives et le contexte scientifique dans lequel nous avons mené nos recherches, puis nous apportons des précisions sur les domaines concernés par nos travaux, enfin, nous détaillons le plan de ce mémoire.

1 Contexte

1.1 Contexte administratif

La recherche présentée dans ce mémoire s'est déroulée au sein du laboratoire d'informatique médicale[1], UMR_S 872, équipe 20, de l'Institut National de la Santé et de la Recherche Médicale[2] de décembre 2003 à novembre 2006. Les trois années de ce travail de thèse ont été financées par une allocation doctorale attribuée par l'école doctorale[3] 393 – Santé Publique : épidémiologie et sciences de l'information biomédicale – de l'Université Pierre et Marie Curie[4], Paris 6.

[1] http://www.spim.jussieu.fr/
[2] http://www.inserm.fr/
[3] http://www.u707.jussieu.fr/spsib/index.html
[4] http://www.upmc.fr/

1.2 Contexte scientifique

Depuis une vingtaine d'années, l'accès aux connaissances médicales est un enjeu majeur pour les professions de santé comme pour le grand public. Face à la multiplication des sources d'informations potentiellement accessibles et face à l'augmentation vertigineuse de la production textuelle, les limites actuelles des outils de traitement de l'information ne proviennent pas de leurs performances pour stocker et traiter rapidement des gros volumes, mais de leur incapacité à prendre en compte les spécificités des vocabulaires métier des utilisateurs. Si on met à part le cas des moteurs de recherche dits « généralistes », la plupart des outils de traitement de l'information proposés sur le marché (traduction, recherche d'information spécialisée, veille…) laissent la possibilité à l'utilisateur d'adapter le système à son domaine en y introduisant ses propres ressources lexicales (appelées selon les systèmes « vocabulaire métier », « ontologie », « lexique spécialisé »…). Cependant, les promoteurs de ces outils laissent l'utilisateur relativement démuni quant aux outils et méthodes pour construire des ressources terminologiques et ontologiques adaptées au domaine et à l'application qui les exploite (Bourigault *et al.*, 2004).

La communauté francophone d'Ingénierie des connaissances travaille depuis plus d'une dizaine d'années sur le problème de la construction de ressources terminologiques et ontologiques à partir de corpus. Elle a produit des résultats, tant théoriques que méthodologiques et logiciels, qui ont été éprouvés dans plusieurs projets applicatifs. Depuis 2000, elle est reconnue comme étant en avance sur cette problématique au niveau international (Aussenac-Gilles *et al.*, 2000). Le développement de ces ressources terminologiques et ontologiques pour faciliter l'usage des terminologies nationales et internationales, disponibles notamment dans le domaine de la médecine, revêt une importance particulière pour le recueil d'information (aide au codage des diagnostics et des actes, réalisation d'études épidémiologiques…) et pour l'accès aux connaissances médicales (base de données bibliographiques MEDLINE, base de connaissances VIDAL sur les médicaments, documents disponibles sur l'Internet…). Les utilisateurs potentiels de ce type de ressources sont le grand public, les étudiants et les médecins, ainsi que des organismes et institutions publiques ou semi-publiques (services dans les hôpitaux, sociétés savantes de médecine…). Il faut également souligner la pertinence de telles recherches dans la mouvance des Sciences et Technologies de l'Information et de la Communication, dans le cadre de la société de l'information et dans le contexte du Web sémantique. En effet, l'interopérabilité attendue des systèmes d'information de santé au sein du Web sémantique passe par la constitution de ressources terminologiques et ontologiques, en particulier les ontologies (Charlet *et al.*, 2004).

Dans cette thèse, notre réflexion a porté sur la collecte, l'organisation, la représentation et la formalisation des connaissances en médecine, tout particulièrement, dans le domaine de la pneumologie. En médecine, la pneumologie est la branche qui s'occupe de maladies des poumons et du tractus respiratoire. Elle est, en général, considérée comme une branche de la médecine interne, bien qu'elle soit très proche des soins intensifs lorsqu'il s'agit de patients qui nécessitent une ventilation mécanique.

Nous avons été amenés à considérer le problème dans son ensemble, afin de comprendre les mécanismes qui sous-tendent la constitution de ressources terminologiques et ontologiques à partir de textes. Nous avons également considéré chaque tâche séparément, afin de proposer, pour chaque étape, si ce n'est une solution, au moins un savoir-faire personnel susceptible d'ap-

porter des éléments de réponse. L'objectif principal de cette thèse consiste à mettre au point une ontologie dans le domaine de la pneumologie pour faciliter, d'une part, l'aide au codage médico-économique des pathologies et, d'autre part, la représentation des connaissances relatives au patient, dans ce domaine de spécialité. La méthode de travail adoptée est une démarche expérimentale courante dans le domaine de l'Ingénierie des connaissances. Il s'agit d'une démarche ascendante qui consiste à partir des problématiques concrètes rencontrées pour aller vers la résolution des questions scientifiques sous-jacentes. Nous essayons de mettre en avant cette approche dans l'annonce du plan de ce mémoire (*cf.* section 3) en soulignant les problématiques clés abordées à chaque chapitre. Selon cette démarche, nous avons tout d'abord cerné les besoins des pneumologues en termes de représentation des connaissances dans leur domaine de spécialité. Ensuite, nous avons mis au point une méthodologie, destinée à l'ingénieur des connaissances, fondée sur la méthode ARCHONTE définie par B. Bachimont (2002). Enfin, nous avons développé un outil d'aide au codage médico-économique semi-automatique à l'usage des pneumologues. Nous sommes revenus, lorsque cela était nécessaire, sur les différentes étapes du processus pour ajuster la méthode de modélisation initiale et améliorer les résultats obtenus. Ce sujet de thèse s'inscrit dans le cadre du projet PERTOMed – production et évaluation de ressources terminologiques et ontologiques dans le domaine médical – dont les objectifs scientifiques seront détaillés plus loin (*cf.* chapitre 2, sec. 1). Plusieurs acteurs ont contribué à la réussite de ce travail parmi lesquels des médecins pneumologues de la Société de Pneumologie de Langue Française[5] et tout particulièrement le docteur F.-X. Blanc pneumologue au CHU du Kremlin-Bicêtre en tant qu'expert du domaine d'application.

2 Domaines concernés

Notre sujet de thèse se situe à la croisée de plusieurs domaines de recherche tels que l'Ingénierie des connaissances, la modélisation informatique, le traitement automatique du langage, l'informatique médicale, la recherche en pneumologie et la gestion et l'optimisation des coûts en médecine. Nous nous sommes spécifiquement intéressés à l'Ingénierie des connaissances et à l'une de ses spécialisations, l'Ingénierie ontologique, en considérant ces domaines par rapport aux problématiques de l'informatique médicale. Nous avons examiné ces trois domaines suivant un point de vue personnel. Nous avons tenté d'en cerner la nature, les enjeux et envisagé d'en utiliser les méthodes et outils pour faire avancer notre propre travail. Nous espérons que le travail réalisé sera profitable, bien qu'à des degrés divers, à chacune de ces disciplines.

Avant de détailler notre problématique dans le chapitre suivant (*cf.* chapitre 2), nous présentons brièvement ces trois principaux champs de recherche auxquels ce travail apporte sa contribution la plus significative.

2.1 Informatique médicale

Le nombre des connaissances disponibles dans les domaines de la médecine, de la biologie et de la santé publique s'accroît. En médecine notamment, plusieurs raisons y concourrent : le

[5] `http://www.splf.org/`

développement de la connaissance clinique et physio-pathologique, la technicité croissante des examens complémentaires, la multiplication et la diversification des structures de prise en charge des malades, l'allongement de la durée de vie des individus. L'augmentation du nombre de paramètres nécessaires à la prise en charge des patients pose le problème de la maîtrise de l'information. En santé publique, l'information pertinente concerne des groupes d'individus et comprend des données démographiques, épidémiologiques, sanitaires, sociales, économiques voire politiques. Face à l'augmentation des connaissances médicales et des paramètres de soins à prendre en compte, il apparaît nécessaire de recourir aux méthodes de traitement de l'information et à l'informatique. L'informatique est la science du traitement automatique de l'information et bénéficie de l'apport des mathématiques, de la statistique, de la linguistique, des sciences cognitives et de la philosophie. L'informatique est aussi une technique, portée par le développement technique et industriel de l'électronique.

Au même titre que la médecine, l'informatique médicale est multidisciplinaire et se situe à l'intersection des technologies de l'information et des différentes disciplines de la médecine et de la santé. Dès 1984, M.S. Blois dans *Information and Medicine - The Nature of Medical Descriptions* résume l'hétérogénéité de la science médicale de manière tout à fait éloquente et justifie la nécessité de l'informatique médicale :

> *« It is sometimes asserted that medical science is no different than any other science. I would strongly disagree with this view ; medical science (human biology) in its describing, reasoning, explaining, and predicting, necessarily draws upon a number of lower-level sciences, while physics, for example, does not. This obvious state of affairs (that medicine rests upon a hierarchy of natural sciences) has profound consequences. Because medicine derives its experimental content from a set of sciences (including both « hard » and « soft » sciences), the processing of the observational data of medicine faces a number of problems. This is one of the reasons why there is a medical information science, and why there is not a « physics » information science*[6] *»* (Blois, 1984).

Cela dit, nous pensons, contrairement à ce que dit M. S. Blois dans la citation ci-dessus, que la médecine vue dans son intégralité n'est pas seulement une science mais également une pratique et que c'est à ce titre que le traitement des données d'observation se révèle compliqué et mobilise les compétences de plusieurs domaines scientifiques. C'est en tant que pratique que nous considérerons la médecine dans la suite de nos travaux de recherche. Sachant que c'est le traitement des connaissances qui est compliqué, nous pensons que l'argument avancé par M.S. Blois s'applique à toutes les spécialités. En médecine et en santé publique, l'informatique est d'abord une méthode imposant la formalisation de l'information et permettant le partage et la

[6] « On affirme parfois que la science médicale ne diffère pas des autres sciences. Je suis fortement en désaccord avec cette idée. La science médicale (la biologie humaine) dans ses dimensions descriptives, résonantes, explicatives et prédictives, utilise nécessairement un certain nombre de sciences de plus bas niveau là où la physique, par exemple, ne le fait pas. Cet état de fait évident (que la médecine domine une hiérarchie de sciences naturelles) a des conséquences profondes. Puisque la médecine tire son contenu expérimental d'un ensemble de sciences (les sciences « dures » comme les sciences « molles »), le traitement des données d'observation en médecine fait face à un certain nombre de problèmes. C'est une des raisons pour lesquelles il y a une science de l'information en médecine et pour lesquelles il n'y a pas une science de l'information en physique. »

diffusion des connaissances. Plusieurs bénéfices peuvent être espérés de l'application de ces principes : augmenter la fiabilité des données (saisie, enregistrement, transmission), comprendre les mécanismes d'interprétation et de raisonnement médical, sélectionner les données les plus pertinentes parmi la masse des informations disponibles, rationaliser les choix au niveau individuel ou collectif par l'application de protocoles, partager l'information et faciliter l'accès à la connaissance. Ainsi, l'informatique médicale s'attache à développer et à évaluer des méthodes et des systèmes pour l'acquisition, le traitement et l'interprétation des données « patient » avec l'aide des connaissances issues de la recherche scientifique. Pour réaliser ces objectifs, le domaine utilise des méthodes scientifiques qui lui sont propres mais qui héritent de l'informatique, des mathématiques et des sciences de gestion. L'informatique médicale a pour vocation de traiter les domaines entiers de la médecine et de la santé, de l'informatisation des dossiers médicaux (Degoulet & Fieschi, 1991; Clemmer, 1995; Renard *et al.*, 2000) au traitement d'images par la robotique (Voros *et al.*, 2006) en passant par la recherche d'information (Soualmia & Darmoni, 2005) et par l'informatisation des guides de bonnes pratiques (Shiffman *et al.*, 2004). Certains secteurs du domaine ont un caractère nettement appliqué mais ils génèrent des recherches fondamentales.

La gestion du langage médical par l'outil informatique pose un problème majeur. En effet, il faut rester proche de la structure naturelle de l'information pour la dénaturer le moins possible (question de la standardisation du langage médical) et adopter la représentation informatique la plus efficace (question de la gestion et de la structure des données). Cela soulève deux problèmes au centre des réflexions de l'informatique médicale : (1) comment faut-il organiser les informations de façon à obtenir le système le plus efficace et le plus informatif ? (2) comment peut-on représenter les informations présentes dans le langage en leur conservant le maximum de richesse (et donc d'ambiguïté) sans renoncer à appliquer les algorithmes de traitement automatique de l'information ? Le principal défi de l'informatique médicale est l'adaptation et le transfert des méthodes et des systèmes développés et rendus opérationnels pour une spécialité médicale à une autre spécialité (par exemple de la pneumologie à la cardiologie – Van Bemmel & Musen, 1997).

2.2 Ingénierie des connaissances

L'Ingénierie des connaissances est un domaine de recherche de l'Intelligence artificielle qui historiquement s'intéressait à la conception et à la réalisation de systèmes experts et de systèmes à base de connaissances. Il s'agit d'une science récente et pluridisciplinaire qui étudie les concepts, méthodes et techniques qui permettent de modéliser ou d'acquérir des connaissances (Charlet, 2005). Elle s'intéresse tout particulièrement à l'identification, la compréhension et l'exploitation du sens pour le partage et la diffusion d'informations. Pour cela, elle réalise des systèmes informatiques « intelligents » portant sur la résolution d'un problème donné, ce qui nécessite trois étapes essentielles : la modélisation de ce problème, la mise au point de la méthode de résolution et l'implémentation informatique du modèle obtenu (Charlet, 2002). Ce domaine de recherche se trouve au carrefour de plusieurs réflexions parmi lesquelles la linguistique pour étudier la formulation des connaissances, la logique pour l'élaboration de modèles formels, la psychologie, la sémiotique, l'ergonomie, sans oublier l'informatique pour mettre en œuvre les modèles. Ses do-

maines d'applications sont nombreux et variés : l'acquisition de connaissances à partir de corpus textuels, la recherche d'information dans des bases de données ou sur le web, l'indexation, la réalisation de mémoire d'entreprise, l'annotation documentaire, la maintenance de thésaurus... Notre travail constitue un apport à l'Ingénierie des connaissances dans la mesure où il vise à développer une méthodologie et à mettre en œuvre des techniques pour la construction de modèles de connaissances conceptuels à partir de textes, pour répondre à une tâche définie. Faire de la recherche dans ce domaine implique de faire de nouvelles propositions d'ingénierie des connaissances, générales ou pour des domaines et applications particuliers. Ces propositions peuvent être des modèles (notre ontologie OntoPneumo), des outils (MedCKARe notre outil d'aide au codage), des combinaisons d'outils ou d'approches (expérimentation de combinaison entre des patrons lexico-syntaxiques et l'analyse distributionnelle[7]), des méthodologies (reprise de la méthode ARCHONTE et apports de précisions).

2.3 Ingénierie ontologique

Née des besoins de représentation des connaissances, l'Ingénierie ontologique est, à l'heure actuelle, au cœur des travaux menés dans le domaine de l'Ingénierie des connaissances (*cf.* figure 1.1). Elle cherche à mettre au point des représentations et des modèles pour permettre aux systèmes informatiques de manipuler la partie sémantique des informations. Au même titre que l'Ingénierie des connaissances, l'Ingénierie ontologique est un domaine pluridisciplinaire puisque la construction d'ontologies – l'objet de travail de l'Ingénierie ontologique – demande à la fois une analyse sémantique, et donc linguistique des informations, la mise en place d'outils de raisonnement et de calcul, la définition de langages de représentation et la réalisation de systèmes informatiques pour les utiliser. C'est également un domaine en pleine expansion qui a des répercussions sur les systèmes d'aide à la décision, les systèmes d'enseignement assisté par ordinateur, les systèmes de gestion de connaissances... Une véritable ingénierie s'est constituée autour du développement d'ontologies dans des domaines aussi variés que la médecine (Le Moigno *et al.*, 2002b), le droit (Lame, 2002), la biochimie (Gene Ontology Consortium, 2001), l'indexation de séquences audiovisuelles (Troncy, 2004), l'électronique (Bellatreche *et al.*, 2006)... Cette ingénierie a pour objectifs la construction d'ontologies, leur exploitation, leur maintenance et, de manière générale, leur gestion tout au long de leur cycle de vie (Gandon, 2006). Dans ce sens, les ontologies apparaissent aujourd'hui comme des composants logiciels avancés qui s'insèrent au centre des systèmes informatiques pour leur apporter une dimension sémantique qui, jusqu'à présent, leur faisait défaut. Un des plus importants projets d'Ingénierie ontologique à ce jour consiste à ajouter au web une sur-couche de connaissances qui permettrait de faire de la recherche d'informations au niveau sémantique et non plus seulement au niveau syntaxique (Charlet *et al.*, 2005). Le « Web sémantique » se veut un web dont le contenu peut être exploité et surtout appréhendé par des machines. À l'heure actuelle, les machines savent bien gérer la quantité (*cf.* les moteurs de recherche classiques) mais avec des performances de qualité moyenne, que le Web sémantique a la prétention d'améliorer. L'un des objectifs est alors de transformer la masse ingérable humainement des pages web en un gigantesque index hiérarchisé. Le Web sé-

[7] Travaux réalisés en commun avec Véronique Malaisé.

mantique a donc pour but de donner aux informations un sens que même les ordinateurs pourront « comprendre » (Berners-Lee *et al.*, 2001). Il peut être vu comme une infrastructure qui complète le contenu informel du web actuel avec de la connaissance formalisée.

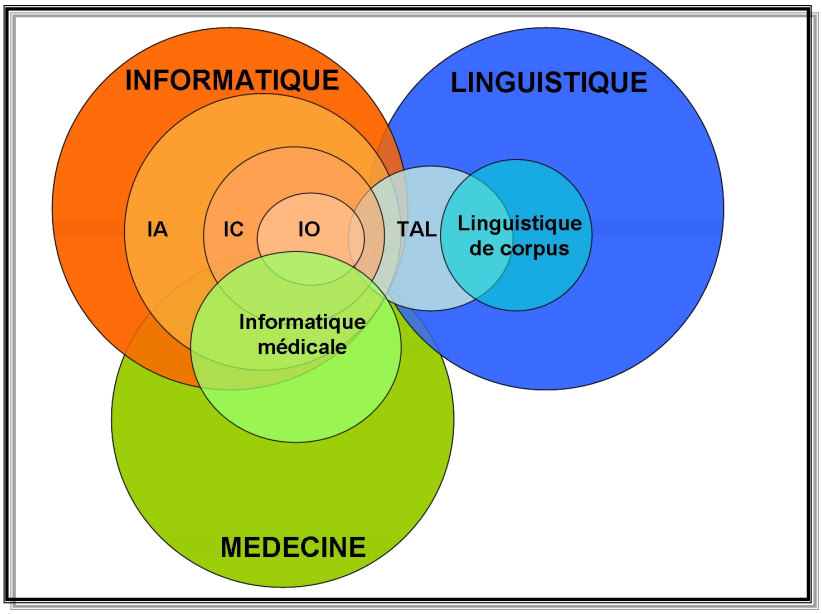


FIG. 1.1 – Interdisciplinarité de l'Ingénierie des connaissances, de l'Ingénierie ontologique et de l'Informatique médicale.

Dans le domaine de l'Ingénierie ontologique, notre travail de thèse contribue à faire connaître l'intérêt des ontologies dans le milieu médical, à préciser une méthodologie de construction générique qui s'adresse à un ingénieur des connaissances, à utiliser et améliorer des outils existants et à construire une application qui repose sur une ontologie et qui réponde aux besoins particuliers d'une communauté d'utilisateurs. Ainsi, nous proposons une étude de cas complète proposant des contributions méthodologiques (collecte des ressources nécessaires à la construction de l'ontologie, mise au point d'une méthodologie, développement de la hiérarchie ontologique, validation et évaluation du modèle) et techniques (développement d'un outil d'aide au codage médico-économique et à la représentation des connaissances dans le domaine de la pneumologie utilisant l'ontologie, adaptation des ressources linguistiques – ontologie, thésaurus, lexiques – aux besoins de l'outil, utilisation de logiciels d'Ingénierie des connaissances et de techniques du Traitement automatique de la langue et des Statistiques).

3 Organisation du mémoire

Nous venons, dans ce chapitre, de présenter le contexte général dans lequel s'inscrit notre sujet de thèse. Dans la section 1, nous avons précisé quel était le contexte administratif dans lequel nous avons travaillé ces trois dernières années. Nous avons ensuite apporté des précisions

sur le contexte scientifique qui nous a servi de cadre. Dans la section 2, nous avons présenté les trois domaines de recherche qui concernent notre travail : l'Ingénierie des connaissances, plus spécifiquement l'Ingénierie ontologique, ayant comme cadre applicatif l'informatique médicale.

Le chapitre 2 est entièrement consacré à la problématique et aux enjeux de notre travail. Nous présentons, section 1, les objectifs du projet PERTOMed puis nous précisons notre problématique en section 2. La section 3 concerne les hypothèses sur lesquelles nous avons fondé notre travail par rapport à certaines des difficultées que nous avons rencontrées. Enfin, la section 4 propose une synthèse de notre problématique et souligne l'originalité de notre approche.

Dans la suite du plan nous mettons l'accent sur les problématiques sous-jacentes auxquelles nous avons été confrontés et que nous avons tenté de résoudre :

Le chapitre 3 est consacré à la représentation des connaissances, principalement aux solutions terminologiques et ontologiques.

Comment est-on passé de la notion de classification à celle d'ontologie ? Quels sont les rapports entre terminologie, thésaurus et ontologie ? Nous définirons ces notions en section 1. Quelles sont les ressources structurées existantes en médecine ? Dans quelle mesure paraissent-elles adéquates et utilisables ? La section 2 présente un bilan des ressources terminologiques et ontologiques existantes dans le domaine de la médecine. Comment construit-on une ontologie ? À partir de quelles décisions théoriques ? Quels sont les principes à suivre ? Quels langages de représentation des connaissances peut-on adopter et pour quels usages ? Quels sont les différents outils qui peuvent nous aider à développer une ontologie ? Nous répondrons à ces questions en faisant un état de l'art des formalismes de représentation en section 3 et des méthodes de construction d'ontologies en section 4 qui s'intéressera également, en section 5, aux langages existants pour exprimer les ontologies et, en sections 6 et 7, aux différents types d'outils d'Ingénierie ontologique et de Traitement automatique du langage aidant à la construction d'ontologies.

Le chapitre 4 est consacré à la modélisation ontologique des connaissances en médecine et détaille nos expérimentations dans le domaine de la pneumologie. La question qui nous intéresse dans ce chapitre peut être formulée ainsi : quelles propositions pour construire une ontologie en médecine lorsque le concepteur n'est pas un expert du domaine qui doit être modélisé ?

Pourquoi et comment utiliser la méthode ARCHONTE ? La section 1 rappelle l'intérêt de mettre le texte au centre du processus méthodologique de construction d'une ontologie et explique quels sont les principes et les différentes étapes de la méthode ARCHONTE. Qu'est-ce qu'un corpus de référence ? Comment identifier les ressources pertinentes ? Quelle taille doit faire un tel corpus pour être « exhaustif » ? C'est à ces questions que la section 2 s'intéresse. Les sections suivantes présentent les différentes étapes méthodologiques qui mènent à l'élaboration de l'ontologie de la pneumologie. Elles précisent les libertés prises par rapport à la méthode de départ et justifient ces adaptations. Il s'agit ici de rendre compte d'une expérience particulière et d'un savoir-faire. Qu'est-ce qu'une ontologie de haut niveau ? Dans quelle mesure peut-on parler d'une ontologie de la médecine ? Dans la section 6, nous donnons notre avis sur ces questions en nous appuyant sur l'expérience de réutilisation, dans nos travaux, du haut de l'ontologie MENELAS consacrée à la cardiologie. Enfin, la section 9 détaille les résultats obtenus et discute les faiblesses et les points forts de notre approche méthodologique.

Le chapitre 5 s'intéresse à l'utilisation des ontologies dans les systèmes informatiques. Une fois l'ontologie construite, comment l'adapter et comment l'exploiter dans une application in-

formatique ? Est-ce un plus ? Nous détaillons notre expérimentation avec l'outil de codage semi-automatique des comptes rendus d'hospitalisation nommé MedCKARe. Nous souhaitons montrer dans ce chapitre ce qu'implique, en terme d'ingénierie, l'utilisation d'une ontologie dans une application. Du point de vue de l'Informatique médicale, l'utilisation d'un tel modèle couplée avec des techniques et outils du Traitement automatique du langage, ouvrent de nouvelles voies : gain de temps, amélioration de l'aide apportée à l'utilisateur, ... Le développement de MedCKARe a été guidé par notre volonté de répondre, le mieux possible, aux besoins des médecins.

Quels sont les outils existants pour le codage médico-économique ? Existe-t'il des outils permettant de visualiser les connaissances que le médecin a sur son patient ? La section 1 analyse les besoins des pneumologues en matière de codage pertinent et détaille les objectifs que nous nous sommes fixés avec MedCKARe. La section 2 dresse un bref état de l'art des outils existants pour coder de l'information en médecine. La section 3 présente brièvement les ressources dont nous disposons pour faire fonctionner MedCKARe. La section 4 présente l'extracteur terminologique que nous avons utilisé. La section 5 explique le fonctionnement de notre application et détaille chaque étape du processus de reconnaissance des informations à coder : quels choix de représentation pour réaliser un dictionnaire ? quels sont les patrons lexico-syntaxiques les plus pertinents pour identifier les connaissances qui nous intéressent ? Comment peut-on gérer les relations dans le cadre du codage médical ? Comment faire le lien entre ontologie et thésaurus pour reconnaître un maximum d'informations dans les textes ? La section 6 détaille les résultats obtenus d'un point de vue qualitatif et quantitatif pour les deux types de codage que notre outil propose : le codage médico-économique PMSI et le codage médical qui rend compte des connaissances que le médecin a sur le patient. La section 6.4 discute les faiblesses et les points forts de notre outil et propose des pistes d'améliorations. Enfin, la section 7 rappelle les enjeux de ce travail, les résultats et les améliorations qui devront être apportées.

Le chapitre 6 s'intéresse à la question de l'évaluation et de l'évolution d'une ressource terminologique et ontologique : Quels sont les critères judicieux pour évaluer une ressource terminologique et ontologique ? Comment et quand faut-il faire évoluer une ressource terminologique et ontologique ? Existe-t-il des méthodes ou des principes génériques ? Dans le cas des ressources terminologiques et ontologiques construites à partir de textes, dans quelle mesure doit-on faire évoluer le corpus de référence ? Comment envisager le processus de maintenance ? Autant de questions que nous avons essayé de traiter au mieux dans ce chapitre. La section 1 positionne notre propos, puis la section 2, dresse un état de l'art des différents critères d'évaluation d'une ressource termino-ontologique. La section 3 revient sur notre expérience de construction de l'ontologie de la pneumologie et de modélisation du thésaurus de spécialité du point de vue de l'évolution et de la maintenance. Enfin, la section 4 remet en avant les axes de réflexions qui nous ont semblés importants.

Le chapitre 7 présente les perspectives et conclusion à donner à ces recherches. La section 1 revient sur la question de la réutilisabilité de la top-ontologie du projet MENELAS. La section 2 est consacrée aux serveurs de terminologies et aux services associés qu'ils proposent. Cette section s'intéresse tout particulièrement au serveur PERTOMed et à sa mise en place. Nous décrivons, en section 3, le projet RNTL 2006 DaFOE4App dans lequel l'application MedOC, présentée en section 4, constitue la suite logique de notre travail de thèse. La section 5 clôt ce mémoire en rappelant les différentes étapes de notre travail de doctorat, des objectifs initiaux aux

différentes réalisations.

CHAPITRE 2

Problématique scientifique et enjeux

« Seul celui qui agit apprend. »
Friedrich Nietzsche
Ainsi parlait Zarathoustra (1883)

Dans ce chapitre, nous situons la problématique scientifique de notre travail de thèse, et ses enjeux, en expliquant quels étaient nos objectifs au sein du projet PERTOM*ed (*cf. *section 1). L'énoncé de ces objectifs nous permet, en section 2, de formuler notre problématique et de distinguer deux axes d'investigation distincts. Nous formulons, en section 3, plusieurs hypothèses de recherche sur lesquelles fonder notre travail de départ pour répondre à certaines des difficultés que nous avons rencontrées. Enfin, en section 4, nous faisons la synthèse et soulignons l'originalité de nos travaux ainsi que nos contributions aux domaines de l'Ingénierie des connaissances et de l'Informatique médicale.*

1 Projet PertoMed

Le projet de recherche PERTOMed – production et évaluation de ressources terminologiques et ontologiques dans le domaine de la médecine[1] – a été financé par le Centre National de la Recherche Scientifique[2] dans le cadre du programme « Traitement des Connaissances, Apprentissage et Nouvelles Technologies de l'Information et de la Communication[3] ». D'une manière générale, ce projet visait à développer une infrastructure proposant un ensemble de méthodes et

[1] http://pertomed.spim.jussieu.fr/
[2] http://www.cnrs.fr/
[3] http://www.dr4.cnrs.fr/tcan/index.html

d'outils opérationnels pour la production et l'utilisation de ressources terminologiques et ontologiques en médecine. Plusieurs réflexions menées autour de la nature de cette infrastructure ont, par ailleurs, suscité notre intérêt pour les serveurs de terminologies (*cf.* chapitre 7, section 2). PERTOMed avait trois objectifs principaux :

Production de ressources terminologiques
Le premier objectif était de produire des ressources terminologiques et ontologiques dans plusieurs secteurs de la médecine : réanimation chirurgicale, pneumologie et pharmacovigilance. Ces ressources ont été construites en étroite collaboration avec des groupes d'utilisateurs avec lesquels sont, en particulier, mises en place des procédures d'évaluation de ces ressources dans leur contexte d'usage.

Partage et développement d'expertise
Le deuxième objectif du projet était de créer une forte synergie entre les différentes actions de productions de ressources terminologiques et ontologiques, en mettant en place une réflexion pour partager et confronter les retours d'expérience et profiter des avancées des uns et des autres. Ces réflexions se sont développées selon trois axes transversaux fondamentaux : méthodologie, normes et évaluation.

Développement de méthodes et d'outils logiciels pour l'appariement de terminologies
Le problème de la construction de ressources terminologiques et ontologiques sectorielles à partir de corpus spécialisés étant considéré comme acquis à la suite de la réalisation du premier objectif, le point fondamental est alors celui de la mise en correspondance, pour un secteur donné, de terminologies élaborées dans des contextes différents ou dans des langues différentes. L'enjeu qui consiste à développer des outils pour l'appariement est très important pour la maintenance des terminologies, pour la recherche d'information monolingue ou multilingue, pour l'accès du grand public aux informations médicales disséminées sur la toile et pour une meilleure prise en compte des vocabulaires réellement utilisés par les médecins pour le codage des actes médicaux.

Au sein de ce projet, notre travail répond à deux objectifs :

- Proposer aux médecins pneumologues un environnement d'aide au codage et à la représentation des connaissances médicales reposant sur le modèle conceptuel d'une ontologie du domaine. Pour cela, nous avons travaillé en étroite collaboration avec le service de Pneumologie du Kremlin Bicêtre et avec la Société de Pneumologie de Langue Française.
- Développer une méthodologie de construction d'ontologies à partir de corpus textuels, fondée sur la sémantique différentielle vue selon B. Bachimont (2002). L'intérêt est de mettre au point un processus méthodologique très précis destiné à un ingénieur des connaissances, non spécialiste du domaine à modéliser, de manière à ne faire appel à l'expert médical que pour des moments précis de validation.

Dans la section suivante, nous allons préciser les problématiques que soulève la réalisation de ces deux objectifs.

2 Problématique

2.1 Limites du codage médico-économique des pathologies

Depuis la mise en place du Programme de Médicalisation des Systèmes d'Information (PMSI), le codage des diagnostics et des actes médicaux est devenu une obligation légale des structures de soins. Ce programme est un outil de description et de mesure médico-économique de l'activité hospitalière. Introduit en France dans le milieu des années 80 par Jean de Kervasdoué, alors responsable de la Direction des Hôpitaux, il a d'abord été présenté comme un outil épidémiologique avant de devenir un outil d'allocation budgétaire. Il a également pour mission de favoriser un meilleur échange entre les partenaires hospitaliers : médecins, soignants, administratifs. Il a été généralisé dans le secteur hospitalier public en 1994 et dans le secteur hospitalier privé en 1996. D'abord utilisé en court séjour MCO (Médecine, chirurgie et obstétrique), le PMSI est maintenant utilisé pour les soins de suite et de réadaptation (PMSI SSR) et la psychiatrie (PMSI PSY). Le PMSI est obligatoire depuis la loi du 31 juillet 1991 qui oblige les établissements de santé à procéder à l'évaluation et à l'analyse de leur activité. Cette obligation cherche à réduire les inégalités de ressources entre les établissements de santé et touche toutes les spécialités médicales. Loin d'être un processus anodin, elle demande une vraie disponibilité des médecins chargés de ce codage ainsi que l'acquisition de nouvelles compétences. Habituellement, les informations nécessaires au codage sont recueillies à la fin de chaque séjour d'un patient à partir du compte rendu d'hospitalisation par des médecins qui peuvent n'avoir jamais vu le patient. La codification du diagnostic principal et des diagnostics secondaires est effectuée, dans ces comptes rendus, à partir de la Classification Internationale des Maladies[4] (CIM-10 – *cf.* section 2.1).

2.1.1 Standardisation du langage médical

Les projets actuels de standardisation du contenu des dossiers médicaux et de codage des pathologies tentent d'unifier le langage médical, en niant la complexité de la démarche de soins. La validité et l'intérêt de ce codage sont pour l'instant réduits par la difficulté d'obtenir un fort consensus parmi les médecins-codeurs sur la signification des codes et leur adéquation avec la situation qu'ils sont censés caractériser, notamment en ce qui concerne les pathologies associées ou comorbidités. Ainsi, les études statistiques réalisées à partir du codage généralisé et exhaustif des pathologies incluent de nombreux biais, conduisant à des risques d'analyses erronées et des prises de décision inadaptées (Suesser, 2000). Qui plus est, les médecins utilisent le codage PMSI a des fins médico-économiques comme cela est prévu, mais également à des fins d'études épidémiologiques. Cela ajoute à la confusion ambiante.

Il faut également insister sur la difficulté méthodologique d'obtenir, parmi une population de médecins, un recueil de données homogène pour un même état pathologique. Sachant que ces médecins ne sont ni des enquêteurs, ni des chercheurs, mais des praticiens exerçant une activité quotidienne de soins déterminée essentiellement par sa dimension clinique et thérapeutique.

S'y ajoute une difficulté supplémentaire lorsqu'il s'agit de hiérarchiser le codage en une affection principale et une affection secondaire. En effet, ce choix peut se révéler ardu car, pour

[4] `http://www.med.univ-rennes1.fr/noment/cim10/`

un grand nombre de situations, il ne repose pas sur des critères objectifs et laisse une place importante à la variabilité entre utilisateurs ou, au cours du temps, pour un même utilisateur. Les choix effectués sont d'autant plus importants qu'ils influent sur la valorisation financière de l'activité.

Si, pour les raisons précédemment évoquées, les praticiens ne trouvent pas dans le codage des actes et des pathologies un instrument utile d'évaluation de leur pratique, ils n'y verront qu'un outil de surveillance contraignante de leur activité, ce qui est souvent le cas (Charbonnel *et al.*, 1996). Tel qu'il est envisagé à ce jour, le codage des pathologies semble poser plus de problèmes qu'il n'en résout. Dans ces conditions, une des problématiques relatives à l'aide au codage médico-économique, consiste à se demander comment rendre compte des richesses d'un langage de spécialité, tel que le langage médical, dans une nomenclature standardisée.

2.1.2 Limites des thésaurus de spécialité

La procédure de codification est la plus souvent réalisée manuellement par les praticiens qui s'aident d'un thésaurus de spécialité (Bensadoun, 2001). Ces thésaurus, proposés par les sociétés savantes, sont construits pour permettre aux médecins de coder à partir de leur terminologie usuelle mais il est aujourd'hui manifeste que les outils d'aide au codage fondés sur ces thésaurus sont inadaptés aux besoins du praticien (Friedman *et al.*, 2004). En effet, les libellés de ces thésaurus se révèlent ambigus (par exemple, à un même code sont associées plusieurs pathologies) et non exhaustifs. Le mode de classification choisi est difficile à appréhender, et le maintien de la consistance ainsi que de la cohérence du thésaurus est impossible à assurer manuellement. De plus, il est évident que le sens des libellés de ces nomenclatures médicales (SNOMED, CIM-10, ...) repose sur les facultés d'interprétation du lecteur humain : ces nomenclatures ne sont donc pas non plus adaptées à une exploitation par l'ordinateur. C'est pourquoi il semble indispensable de décrire la sémantique et l'organisation des objets du domaine médical de manière formelle, afin de se doter de modélisations conceptuelles non contextuelles et non ambiguës. Une telle modélisation est appelée ontologie (Charlet, 2002; Bachimont, 2000; Staab & Studer, 2004). Nous détaillons la problématique résultante de la construction d'ontologie dans la section 2.2 et précisons notre définition d'ontologie dans le chapitre 3, section 1.

La volonté de décrire les informations propres aux patients dans le cadre d'un modèle conceptuel formel, à l'aide d'une ou plusieurs ontologies est une direction de recherche particulièrement explorée dans le domaine de l'informatique médicale (Spackman & Campbell, 1998). La plupart des chercheurs qui travaillent dans ce domaine, espèrent, à la suite de la construction de ce modèle conceptuel du patient, en avoir une représentation canonique favorisant l'interopérabilité des dossiers médicaux. Nous nous demandons alors : (1) quelles sont les connaissances médicales déterminantes pour « parler » du diagnostic médical et, de manière plus générale, rendre compte de l'activité pneumologique ; (2) quelles sont les connaissances pertinentes en pneumologie pour le codage PMSI.

2.1.3 Proposition

Par rapport aux systèmes classiques d'aide au codage (Friedman *et al.*, 2004), nous recherchons d'abord une représentation conceptuelle, là où les travaux habituels recherchent directement une représentation dans un thésaurus, tel que le MeSH, le métathesaurus d'UMLS ou la SNOMED (Dolin *et al.*, 2001). Pour cela, nous allons construire une ontologie de la pneumologie. Nous sommes ainsi beaucoup plus proches des travaux de A. Rector (1998) sur les classifications formelles. Cette démarche nous démarque également des récents travaux sur la SNOMED-CT que ses concepteurs visent à « ontologiser », en lui faisant jouer simultanément le rôle de l'ontologie formelle pour la véracité des inférences faites dessus et le rôle de thésaurus pour l'interaction avec le praticien (Spackman & Campbell, 1998). À une époque où des groupes de pression tentent d'imposer la SNOMED-CT comme l'unique moyen de représentation de la connaissance médicale, il nous paraît indispensable de proposer et de faire connaître des solutions alternatives.

La solution que nous proposons passe par le développement et la mise en œuvre d'un environnement de codage de l'information médicale qui dépasse le seul cadre médico-économique imposé par le PMSI. Ce travail est ancré dans une réflexion sur les moyens de coder l'information médicale afin de représenter les patients et leurs pathologies. Le problème du codage se situe au niveau du passage au formalisme : l'expression des connaissances, qui se présente le plus souvent et le plus naturellement sous forme textuelle, doit être transformée en un codage qui, lui, est toujours réducteur en termes de représentation du sens. Nous proposons un outil qui permet le codage en même temps qu'il permet aux médecins d'assumer le caractère réducteur de ce processus (*cf.* chapitre 5).

Contrairement aux outils commerciaux existants, nous ne souhaitons pas automatiser complètement la procédure de codage. L'idée est de permettre aux médecins de se réapproprier le processus de codage. C'est pourquoi nous proposons un système de représentation des connaissances avec lequel le médecin puisse interagir pour construire la représentation du patient qu'il désire, en tenant compte de ses propres capacités de choix et d'interprétation mais en l'aidant dans sa tâche. Nous souhaitons donc proposer : (1) un codage médico-économique prenant en compte d'autres critères que les seuls médicaux et, (2) une représentation des connaissances médicales, que nous appelons « codage médical » à des fins d'indexation. Ainsi, par exemple, les pneumologues pourront rechercher tous les cas de patients atteints de « sténose serrée de la trachée à la fois par compression extrinsèque et par envahissement de la muqueuse » diagnostiquée par « une endoscopie bronchique ». Pour l'instant aucun outil ne permet de faire de telles recherches sur des comptes rendus d'hospitalisation. Ce codage médical est évidemment une représentation réductrice de ce que contient le dossier médical du patient mais c'est parce qu'elle est réductrice que cette représentation permet au médecin un rappel rapide, une représentation résumée de ce qu'il sait sur le patient.

2.2 Construction d'ontologies en médecine à partir de textes

La construction d'ontologies à partir de textes constitue un enjeu important aussi bien pour la communauté des chercheurs en Traitement automatique des langues que pour celle de l'Ingénierie des connaissances. Les systèmes de traitement de l'information qui doivent fonctionner

dans des domaines de connaissances spécialisés comme la médecine ne peuvent être efficaces que s'ils s'appuient sur des ressources terminologiques et ontologiques, construites pour le domaine concerné et en vue d'une application particulière (Bourigault & Aussenac-Gilles, 2003). L'enjeu, dès lors, est d'élaborer des méthodes d'acquisition des connaissances à partir de textes qui spécifient (1) comment utiliser les outils de Traitement automatique des langues, nécessaires à l'analyse de corpus, et (2) les environnements de modélisation des connaissances, nécessaires à la construction d'ontologies. Nous verrons dans le chapitre 3, section 4, qu'aucune méthodologie générale de construction d'ontologies n'a, pour l'instant, réussi à s'imposer. Ceci dit, quelle que soit la méthodologie adoptée, le processus de construction doit faire l'objet d'une collaboration qui rassemble des experts du domaine à modéliser, des ingénieurs des connaissances et les futurs utilisateurs (Farquhar *et al.*, 2000). Cette collaboration ne peut être fructueuse que si les objectifs du processus de construction et les besoins qui en découlent sont clairement définis. Selon M. Uschold (1995), l'ingénieur des connaissances doit s'interroger sur trois aspects du processus de construction : l'objectif opérationnel, le domaine de connaissance et les utilisateurs. Il paraît évident que l'ingénieur des connaissances doit rencontrer les futurs utilisateurs pour établir un cahier des charges pertinent. Nous allons examiner les deux autres aspects ci-dessous. Notre cas particulier étant le développement d'une ontologie à partir de textes, nous ajouterons quelques réflexions sur cette spécificité.

2.2.1 Objectif opérationnel

Notre représentation conceptuelle des connaissances en médecine passe par le développement d'ontologies. Brièvement, une ontologie est un système formel dont l'objectif est de représenter les connaissances d'un domaine spécifique au moyen d'éléments de base, les concepts, définis et organisés les uns par rapport aux autres (Rector, 1998). Il est cependant difficile de repérer et de classifier les objets d'un domaine car les critères de classification dépendent des buts poursuivis et n'ont rien d'immuables (Charlet, 2002). Le contenu, la forme, la couverture, le degré de formalisation, ... sont choisis en fonction du rôle que doit jouer l'ontologie dans l'application cible. Une fois construite et acceptée par une communauté particulière, cette ontologie traduit un consensus explicite et un certain niveau de partage, deux aspects essentiels pour permettre son exploitation par différentes applications ou agents logiciels. Ce point de vue soulève un paradoxe important : d'une part, dans un souci de réutilisabilité, l'ontologie gagne à cultiver une certaine indépendance vis-à-vis des différentes applications dans lesquelles elle peut être utilisée et, d'autre part, sa construction elle-même doit être guidée par l'usage dans l'application cible. Se pose alors le problème de l'utilisation opérationnelle des ontologies, c'est-à-dire de leur mise en œuvre pratique (Fürst, 2004b).

Il est indispensable de préciser soigneusement l'objectif opérationnel de l'ontologie, en particulier au travers de scénarios d'usage. C'est ce que nous avons fait dans le cahier des charges de notre outil (Baneyx *et al.*, 2005b). Nous avons clairement énoncé quel était notre contexte d'usage et ses problématiques dans la section 2.1 de ce chapitre. Nous développons une ontologie pour qu'elle serve de pivot dans un outil de codage médical et médico-économique, dans le domaine de la pneumologie. Pour construire cette ontologie, nous avons travaillé en collaboration avec des pneumologues de la SPLF qui nous ont fourni les ressources textuelles de base.

L'étape, indispensable, de validation de l'ontologie a été assurée par le docteur F.-X. Blanc du CHU du Kremlin-Bicêtre. Les entretiens que nous avons eus avec ce médecin et le secrétariat médical du service de pneumologie nous ont permis de comprendre leur manière de fonctionner et leurs besoins. Comment les médecins codent-ils ? Quelle est la part de l'activité quotidienne dévolue au codage ? Pourquoi n'utilisent-ils pas les outils industriels disponibles ? Quels seraient pour eux les spécificités d'un outil adéquat ? Quelles sont les informations relatives aux patients que les médecins souhaitent voir apparaître rapidement à l'écran ? Quelles sont les informations qu'ils peuvent avoir besoin de retrouver et sous quelle forme ? À partir de là, nous avons pu envisager différentes fonctionnalités qui sont détaillées au chapitre 5 de ce mémoire.

2.2.2 Domaine de connaissances

Le domaine de connaissances doit être délimité aussi précisément que possible, et découpé, si besoin est, en plusieurs aspects : les connaissances du domaine, les connaissances de raisonnement et les connaissances de haut niveau qui, par leur degré d'abstraction supérieur, peuvent être communes à plusieurs domaines.

Notre premier souci concernant la construction de l'ontologie est d'identifier quel est le domaine de connaissance que nous cherchons à modéliser. Cette question peut paraître triviale, pourtant c'est loin d'être le cas en médecine. À priori, notre domaine de connaissance est la pneumologie. Pourtant, l'ensemble des connaissances que nous devons conceptualiser n'appartient pas seulement à cette spécialité. Pour preuve, l'analyse d'un de nos corpus par un outil de Traitement automatique du langage révèle que le syntagme nominal qui a la plus haute fréquence[5] est « cure de chimiothérapie ». Nous aurions volontier classé cette connaissance en dehors du champ de la pneumologie et certainement comme appartenant au domaine de la cancérologie. Quelle que soit sa spécialité médicale, un médecin se sert constamment de connaissances appartenant à d'autres spécialités pour soigner et cela se retrouve dans ses écrits. Dans ce cas, comment définir les limites des connaissances à modéliser et comment identifier les connaissances pertinentes ?

2.2.3 Degré de formalisme et granularité

La formalisation, autre facette des ontologies, est nécessaire pour permettre un raisonnement automatique afin de décharger les utilisateurs d'une partie de leur tâche d'exploitation (Charlet *et al.*, 2006). Ces utilisateurs doivent, au même titre que l'objectif opérationnel, être identifiés car le choix du degré de formalisation et de la granularité de l'ontologie dépendent de leurs besoins et de leurs compétences. Les principales problématiques concernant les questions de formalisme et de granularité auxquelles nous nous sommes intéressés relèvent de l'interprétation du sens des connaissances médicales en logique de description :

1. Comment passer de l'expression linguistique des connaissances à une représentation formelle et calculable, propre à l'exploitation informatique ?
2. Comment ne pas réduire de manière excessive l'expressivité du langage médical en le formalisant ?

[5] Il s'agit du nombre total d'occurrences du terme qu'il soit isolé ou au sein d'un syntagme.

3. Comment définir quelles sont les primitives de représentation et leur signification dans le processus de modélisation ?

2.2.4 Modélisation à partir de textes

Selon l'article de N. Aussenac-Gilles et D. Sörgel (2005), les ontologies construites à partir de textes peuvent représenter et saisir les objectifs d'un domaine de connaissances. Notre ontologie est construite dans un domaine *a priori* peu formel où les connaissances s'expriment principalement en langue. Nous avons constitué des corpus de textes du domaine et les avons soumis à une analyse complète dans le but d'en retirer des traces de conceptualisation sous-jacente (Bourigault *et al.*, 2004). La communauté francophone d'Ingénierie des connaissances travaille depuis une dizaine d'années sur le problème de la construction de ressources terminologiques et ontologiques à partir de corpus (Aussenac-Gilles *et al.*, 2000). Elle a produit des résultats, tant théoriques que méthodologiques et logiciels, qui ont été éprouvés dans un certain nombre de projets applicatifs, tels que MENELAS [6] (Zweigenbaum *et al.*, 1998), TERMINAE (Aussenac-Gilles *et al.*, 2002), DOE (Troncy & Isaac, 2002) ...

La modélisation d'ontologies à partir de textes pose la question du statut de la ressource textuelle et place le corpus comme source privilégiée de connaissances. Le choix et le traitement de ces corpus occupent une place d'importance dans le processus de construction de l'ontologie. Ainsi, il faut veiller à collecter des textes de genre et de taille divers, adéquats avec ceux qui seront traités dans l'application cible et ayant un vocabulaire diversifié. La préparation du corpus, c'est-à-dire l'opération qui consiste à traiter les textes de base pour en faire un corpus, est une opération délicate et coûteuse en temps. Il faut bien souvent convertir ces textes dans un format exploitable et les baliser pour les outils de Traitement automatique des langues. Cette phase du travail nous a amenés à nous interroger sur le nombre de textes et sur le nombre de mots nécessaires pour obtenir un corpus de bonne qualité. Pour répondre à cette question, nous avons utilisé la loi de G. K. Zipf (1949). Nous expliquons brièvement la teneur de cette loi et ses résultats sur nos corpus au chapitre 6, section 2.1.

2.2.5 Proposition

La solution que nous proposons consiste à :

1. Fournir une méthodologie d'Ingénierie ontologique, destinée à être mise en œuvre par un ingénieur des connaissances, qui ne nécessite les compétences d'un expert du domaine que dans les différentes phases de validation du travail.

   Cette méthodologie s'appuie, d'une part, sur l'analyse de corpus textuels par des outils de Traitement automatique des langues et, d'autre part, sur l'utilisation de la méthode ARCHONTE mise au point par B. Bachimont (2000) avec la participation du groupe de réflexion Terminologie et Intelligence Artificielle[7]. Cette méthode, présentée au chapitre 4, section 1, place les questions de définition du sens au centre du processus de construction. Nous l'avons légèrement modifiée pour l'adapter à nos besoins.

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

[7] `https://stid-bdd.iut.univ-metz.fr/~termwatch/TIA/` ou `http://tia.loria.fr`

2. Placer l'ontologie réalisée au sein d'une application informatique.
   Elle servira de pivot à l'outil d'aide au codage et permettra, d'une part, de faire le lien entre les différentes ressources linguistiques requises pour identifier les pathologies à coder pour des raisons économiques (codage PMSI) dans les CRH et, d'autre part, de représenter les données du patient sous la forme de graphes (codage médical).

3 Hypothèses de travail

Plusieurs chercheurs ont posé l'hypothèse de la pertinence du modèle ontologique pour résoudre certains problèmes du domaine. La lecture de ces problèmes dans la littérature nous a amenés à dresser une liste d'hypothèses qui concernent nos propres problématiques. Nous en présentons certaines ci-dessous :

Identifier les connaissances pertinentes d'un domaine
Dès le début de notre travail, nous faisons l'hypothèse que, dans le domaine de la médecine, le CRH véhicule un vocabulaire métier fiable et consensuel.

Délimiter le domaine de connaissance
L'objectif opérationnel, c'est-à-dire les nécessités de l'application cible, va tracer les limites du domaine. Ainsi, si les pneumologues ont besoin de coder des informations, telles que « cure de chimiothérapie » ou « diabète », c'est qu'elles seront présentes dans les CRH et accessibles à la modélisation.

Faire face au volume des données fournies par les outils de Traitement automatique du langage

La méthodologie que nous proposons pour repérer les données pertinentes dans la masse de celles qui nous sont fournies lors de l'analyse des CRH comprend plusieurs étapes (*cf.* chapitre 4). Premièrement, nous utilisons les informations issues de l'analyse syntaxique et de l'analyse distributionnelle pour faire un premier classement des notions intéressantes en fonction de critères tels que la productivité, la fréquence, etc. Ensuite, nous regroupons les notions, précédemment sélectionnées, en les reliant aux axes conceptuels majeurs du domaine. Cette deuxième étape est proche des démarches dites de « clustering ».

Organiser les objets d'un domaine
La sémantique différentielle de F. Rastier (1994), reprise et adaptée dans la méthode ARCHONTE nous permettra de situer chaque objet du domaine dans l'arbre ontologique par rapport à ses pères et frères ontologiques. La construction du sens se fait dès lors de manière compositionnelle par le parcours de l'arbre. Cet arbre fixe le nouveau contexte interprétatif de chaque objet.

Assumer la réduction du sens par le passage au formalisme

Nous adoptons l'hypothèse selon laquelle la représentation des connaissances dans un système formel est un ensemble d'engagements ontologiques (Davis *et al.*, 1993). Si toutes les

représentations sont des approximations imparfaites de la réalité, chaque approximation se concentre sur une certaine vue du monde. En choisissant une représentation, nous prenons inévitablement un ensemble de décisions (comment voir le monde, que voir dans le monde) qui réduisent le sens des objets du monde parce qu'elles en réduisent la complexité. Il ne s'agit pas d'un effet accidentel dû au choix de la représentation des connaissances mais d'un effet essentiel dû à la décontextualisation de la représentation. C'est à la fois la force et la faiblesse de la représentation formelle. Il nous faut donc adopter une sélection judicieuse d'engagements ontologiques qui vont nous fournir l'opportunité de cibler notre attention sur des aspects du monde (ici la pneumologie à l'intérieur du monde médical) que nous pensons pertinents. En Ingénierie des connaissances, l'ontologie détermine des catégories de concepts qui existent pour répondre aux besoins d'un domaine d'application. Ces catégories représentent les engagements ontologiques du concepteur et c'est ce qui lui permet d'assumer la réduction du sens générée par le passage du linguistique au formel.

4 Synthèse et originalité des travaux

Contributions méthodologiques

L'originalité de notre recherche réside dans la mise au point d'une méthodologie d'Ingénierie ontologique unifiée (tenant compte des principes et des méthodes de l'Ingénierie ontologique, de l'Ingénierie des connaissances, du Traitement automatique des langues, de la logique et de la sémantique différentielle) pour la construction d'ontologies, à partir de textes. À cet effet, nous avons complété et précisé la méthodologie ARCHONTE mise au point par B. Bachimont, en montrant comment s'enchaînent les différentes étapes : 1) analyse terminologique fondée sur l'analyse des textes, 2) identification, sélection et extraction des termes pertinents, 3) normalisation, 4) formalisation et 5) opérationalisation.

Nous expérimentons, en collaboration avec V. Malaisé, la complémentarité de deux modes d'analyse de la langue en usage dans les textes, l'analyse distributionnelle et l'approche par patrons lexico-syntaxiques et montrons que l'utilisation conjointe de ces méthodes facilite la structration hiérarchique des concepts de l'ontologie.

Enfin, nous expliquons comment la réutilisation d'une ontologie de haut niveau de la médecine vient guider la réorganisation d'une première structuration des concepts.

Contributions techniques

Bien que nous utilisions des outils et des approches éprouvés, nous avons tenté d'en articuler l'utilisation du mieux possible. Pour cela nous avons mis au point un certain nombre de programmes informatiques permettant d'aider le passage des uns aux autres, notamment en ce qui concerne la conversion des formats. À cette occasion, nous avons pu cerner quels étaient leurs atouts et leurs limites. Cela nous a fait réfléchir aux moyens d'enchaîner ces outils, à leur complémentarité et aux meilleurs moments pour les utiliser.

Nous avons suivi le cycle de vie complet d'une ontologie, de sa création à son évaluation dans une application.

Enfin, nous avons démontré l'utilité et la valeur ajoutée de l'ontologie au sein d'un sys-

tème d'aide au codage en médecine. Il s'agit d'un travail exploratoire qui s'appuie sur la collaboration de personnes ayant des domaines de compétences variés.

Contributions pratiques

L'ontologie OntoPneumo est la première de nos contributions pratiques. Il s'agit d'un résultat en soi. Elle peut être réutilisée pour des tâches approchantes et nécessiterait alors d'être complétée et en partie remaniée.

Nous avons généralisé notre contribution méthodologique, autant que faire ce peut, dans un guide qui se trouve en annexe A de ce mémoire.

Notre application d'aide au codage MedCKARe utilise OntoPneumo et fait ainsi la démonstration de l'utilité d'une telle modélisation dans un système informatique. La réalisation de cette application nous a également permis de montrer et de mesurer l'impact de l'application cible sur le contenu du modèle et sur la manière de le construire. Le travail d'adaptation et celui de prise en compte des besoins requis pour intégrer l'utilisation de l'ontologie dans l'application cible est loin d'être négligeable.

Enfin, MedCKARe offre également une piste intéressante en matière d'aide au codage et répond à un certain nombre de besoins formulés par les médecins pneumologues.

CHAPITRE 3

Représenter des connaissances : terminologies et ontologies

« Celui qui dit que deux et deux font quatre, a-t-il une connaissance de plus que celui qui se contenterait de dire que deux et deux font deux et deux ? »
Jean le Rond d'Alembert
Discours préliminaire à l'Encyclopédie (1751)

*Ce chapitre propose, de façon non exhaustive, un état de l'art en matière de représentation des connaissances. Nous essayons de synthétiser les acquis des différents sujets abordés et de mettre l'accent sur les nombreux problèmes qu'il reste à traiter. La section 1 s'intéresse aux notions clés du domaine en adoptant un point de vue épistémologique et précise la place de chacune de ces notions au sein du processus de représentation. La section 2 dresse une revue des principales ressources terminologiques et ontologiques disponibles en médecine, certaines ayant été utilisées dans notre travail. Les sections suivantes s'intéressent à la construction d'ontologies et présentent un aperçu de l'existant en matière de formalismes de représentation (*cf. *section 3), de méthodes et de méthodologies de construction (*cf. *section 4), de langages (*cf. *section 5) et d'outils (*cf. *sections 6 et 7).*

1 De la notion de terminologie à celle d'ontologie : épistémologies et définitions

Lors de l'expression orale ou écrite, le mot constitue un maillon essentiel de la chaîne de compréhension. Un mot mal compris, ou auquel on accorde un sens différent de celui qui lui est assigné par l'auteur de l'énonciation, pose problème dans le cheminement logique du raison-

nement et dans l'échange d'informations. L'idée n'est pas ici d'imposer le sens d'un mot, mais d'indiquer clairement le sens assigné à certaines notions dans le contexte de cette thèse.

1.1 Terminologie

Une terminologie présente l'ensemble des termes particuliers à une science, un domaine ou un art (Larousse, 1988), à un groupe de personnes ou à un individu (Office de la langue française, 2000[1]). P. Lefèvre, dans son livre sur la recherche d'informations (2000), propose une définition plus précise : *« Les terminologies sont des listes de termes d'un domaine ou sujet donné représentant les concepts ou notions les plus fréquemment utilisés ou les plus caractéristiques, cette liste étant ou non structurée. »*. La terminologie, considérée comme science, s'intéresse au recensement des concepts d'un domaine et des termes qui le désignent pour faciliter l'échange de connaissances dans une langue et d'une langue à l'autre. Pour cela, il faut, d'une part, normaliser et figer l'expression des concepts du domaine en fixant les termes qui les désignent et, d'autre part, rendre compte de l'agencement relatif des concepts recencés (Wüster, 1981; Zweigenbaum, 1999). Ainsi se profilent les trois sommets du triangle aristotélicien (*cf.* figure 3.1) : la chose, c'est-à-dire l'objet du monde (par exemple le chat Félix), le signe, c'est-à-dire le terme, la chaîne de caractère, ou la photo qui désigne cet objet particulier dans le monde (par exemple Félix, le terme qui désigne ce chat parmi tous les chats) et le concept, c'est-à-dire l'idée du chat Félix.

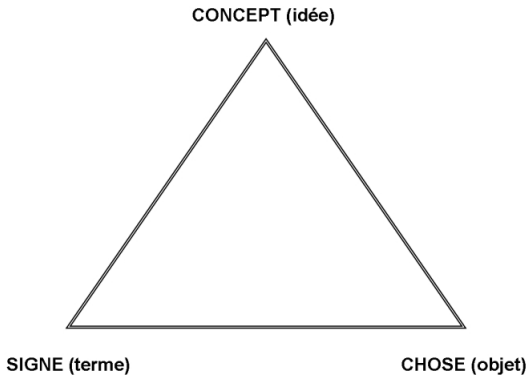


FIG. 3.1 – Triangle aristotélicien.

Les concepts présents dans la terminologie peuvent être reliés par des relations qui témoignent, par exemple, d'un rapport de spécialisation-généralisation (relation *« est-un »*) comme le montre la figure 3.2. Dans cette figure, *« scintigraphie pulmonaire de ventilation »* est plus spécifique que *« scintigraphie de l'appareil respiratoire »*.

Une terminologie modélise ainsi un système de concepts sous la forme d'un système de termes normalisés. Le principal intérêt des terminologies est de réduire, voire de supprimer, l'ambiguïté. En effet, puisque par définition, une terminologie de référence spécifie une norme pour un domaine donné, alors le sens de chaque terme est figé et il n'existe qu'une interprétation

[1] D'après le Grand Dictionnaire Terminologique - `http://w3.granddictionnaire.com`

06.01.06	Scintigraphie de l'appareil respiratoire *À l'exclusion de: recherche d'une thrombose artérielle pulmonaire, par injection de traceur radio-isotopique spécifique (DFQL001)*				
GFQL004 (F, G, P, S, U)	**Scintigraphie pulmonaire de ventilation**	1	0	**189,35**	2
GFQL007 (F, G, P, S, U)	**Scintigraphie pulmonaire de perfusion**	1	0	**193,19**	2
GFQL006 (F, G, P, S, U)	**Scintigraphie pulmonaire de ventilation et de perfusion** *(ZZQL017)*	1	0	**382,54**	2
GFQL001 (F, G, P, S, U)	**Tomoscintigraphie pulmonaire de ventilation**	1	0	**284,03**	2

FIG. 3.2 – Extrait de la classification commune des actes médicaux.

possible pour l'utilisateur. La possibilité de hiérarchiser les concepts permet de relier explicitement un terme générique imprécis, par exemple *« cancer »*, aux termes plus spécifiques qui peuvent le préciser *« cancer de la trachée », « cancer trachéal in situ »*.

Cependant, M. Slodzian (2000) développe longuement des arguments épistémologiques et linguistiques démontrant que le triangle aristotélicien tel qu'il est présenté ci-dessus n'est pas, pour l'Ingénierie des connaissances, une structure figée. Elle discute, d'une part, le principe du mot isolé, c'est-à-dire pris hors contexte, comme point d'entrée privilégié de la terminologie et, d'autre part, interroge le credo selon lequel le sens préexiste.

> *« As a matter of fact, the necessity of reducing drastically the complexity and multiplicity of linguistic facts leads to select one type of semantic paradigm, the one which restate the antique credo that the sign represents the concept and the concept represents the object or referent. And the cognitive semantics paradigm doesn't change the equation : the sign is given as linguistic and the meaning is given as conceptual, but we have the same one-to-one relationship concept vs word. […] As Rastier notes, the paradigmatic option of cognitive semantics leads to a strict semasiological approach, resting upon the prelinguistic prejudice, born from the philosophy of language, that to one word corresponds one signified ; and as this is obviously not the case, one must find for it a preferential signified, or more precisely a basic conceptualization*[2] *»* (Slodzian, 2000).

Nous trouvons que les observations faites et les arguments donnés par M. Slodzian et F. Rastier sont tout à fait pertinents. Ceci dit, l'objet de notre travail est de construire des ontologies pour représenter une région particulière du réel, à un moment donné. À ce moment-là (moment est utilisé dans son acceptation d'origine), la signification des concepts qui sont décontextualisés

[2]En effet, la nécessité de réduire rigoureusement la complexité et la multiplicité de faits linguistiques amène à choisir un type de paradigme sémantique, celui qui réitère le credo habituel selon lequel le signe représente le concept et le concept représente l'objet ou le référent. Le paradigme cognitif et sémantique ne change pas l'équation : le signe est donné comme linguistique et le sens est donné comme conceptuel, mais nous avons la même relation linéaire concept versus mot. […] Comme le note Rastier, l'option paradigmatique de la sémantique cognitive mène à une approche sémiologique stricte reposant sur le préjudice prélinguistique, né de la philosophie du langage, selon lequel à un mot correspond un signifié. Ce n'est bien évidemment pas le cas, il faut trouver pour ce mot un signifié préférentiel ou plus précisément une conceptualisation de base.

est figée. Dans ce contexte, en respectant ces contraintes, alors le triangle aristotélicien présenté ci-dessus est bien figé.

Enfin, nous terminerons cette section en ajoutant qu'il existe des terminologies de natures diverses adaptées aux différents objectifs de traitement de l'information : classification pour le recueil de données, nomenclature pour la description d'observations cliniques et thésaurus pour la recherche d'information. Cela dit, P. Lefèvre (2000) distingue les terminologies des thésaurus et avance que les thésaurus possèdent une structure de réseaux sémantiques et qu'en cela ils ne peuvent pas être inclus dans les terminologies. Nous précisons les notions de classification, nomenclature et thésaurus dans les sections suivantes.

1.2 Classification

Selon J. Charlet (2002), une classification est l'action de distribuer par classes et par catégories. D. Bourigault (2004) offre une définition allant dans le même sens mais plus complète : *« une classification est la répartition systématique en classes, en catégorie d'êtres, de choses ou de notions ayant des caractères communs notamment afin d'en faciliter l'étude. C'est aussi le résultat de cette opération. »*. Une classification consiste donc à partitionner l'ensemble des objets pour les distribuer en classes et sous-classes constituées d'éléments de plus en plus semblables, ici les termes de signification proche. La structure de la classification et la granularité des classes dépend des objectifs poursuivis par son concepteur. La définition de classes plus spécifiques à l'intérieur de classes plus générales, hiérarchise la classification. La classification internationale des maladies (CIM - *cf.* section 2.1) et la classification commune des actes médicaux (CCAM - *cf.* section 2.2) sont de bons exemples de classifications hiérarchiques dans le domaine médical bien qu'elles n'aient pas le même niveau de profondeur.

Les classifications portant sur un domaine particulier de la connaissance sont généralement bien admises par les spécialistes du domaine. Les classifications à vocation universelle ne peuvent faire abstraction d'un point de vue et sont, de ce fait, l'objet de nombreuses critiques. Elles apportent cependant toujours un éclairage sur la nature de la connaissance. Classer les connaissances, c'est dire comment elles se situent les unes par rapport aux autres.

En situant la notion de classification par rapport à celle de terminologie telle que définie ci-dessus, section 1.1, on peut dire que les concepts d'une classification sont ses classes (Zweigenbaum, 1999). Les termes d'une classification appartiennent souvent à un métalangage : *« Asthme SAI*[3] *»*, *« Autres … pathologies »*, *« Cancer primitif bronchique de stade … I, II, III »*, *« À l'exclusion de*[4] *… »*.

Les principales caractéristiques des classifications permettant d'évaluer leur capacité d'expression sont : la nature du principe de classement, la prise en compte d'axes multiples et les types des relations exprimées. Une classification correspond à une catégorisation récursive du domaine selon un critère qui s'applique à l'ensemble des éléments d'une classe, les critères se succédant de classe en sous-classe par ordre d'importance décroissante. Le lien sémantique qui

[3] SAI = sans autre indication.

[4] *« Échographie de l'appareil respiratoire, à l'exclusion de : échographie et/ou échographie-doppler de contrôle ou surveillance de pathologie ».*

préside à la catégorisation peut être un lien de spécification-généralisation (type *« est-un »*, par exemple : *« une bronchite aiguë est une pathologie infectieuse »*) ou de partition (type *« fait-partie-de »*, par exemple : *« le poumon fait partie de l'appareil respiratoire »*). Une classification monoaxiale répartit en plusieurs classes disjointes l'ensemble des objets et revient à construire une hiérarchie de classes à partir d'une racine unique et commune. Les classes d'un niveau doivent couvrir l'ensemble du domaine de ce niveau (exhaustivité) sans se recouvrir (exclusivité) afin qu'un objet trouve une place et une seule. Nous verrons qu'en pratique, il s'avère très difficile de répartir les objets du domaine selon un seul critère. Cette difficulté reconnue a entraîné le développement de répartitions multiaxiales comme le montre l'exemple du decripteur du nez dans le thesaurus *Medical Subject Headings* (MeSH – *cf.* figure 3.8, page 42) que nous présentons en section 2.3.

1.3 Nomenclature

Le mot *nomenclature* vient du latin *nomenclatura* qui désigne l'action d'appeler par le nom. Dans notre domaine, une nomenclature désigne un ensemble de termes techniques, présentés selon un classement méthodique. Il n'y a aucun agencement particulier des termes ni de définition explicite, l'objectif recherché étant l'exhaustivité. Il s'agit d'un recueil ouvert de données dont l'intérêt est de recenser tous les concepts d'un domaine, sans se restreindre à un objectif spécifique. Il s'agit d'un certain type de terminologie. La principale différence entre les notions de classification et de nomenclature tient à la précision de l'objectif poursuivi. En effet, la classification est clairement orientée vers un objectif précis tandis que la nomenclature a pour seul but l'exhaustivité. Ainsi, selon P. Zweigenbaum (1999), lorsque le but est de décrire des informations cliniques le plus précisément et fidèlement possible, les classifications telles que définies ci-dessus (section 1.2) trop orientées vers la résolution d'un objectif particulier, se révèlent peu adaptées. Une nomenclature importante dans le domaine médical est la Nomenclature Systématique des Médecines Humaine et Vétérinaire (SNOMED - Côté *et al.*, 1993). Nous la présentons en section 2.5 de ce chapitre.

1.4 Thésaurus

Un thésaurus est un ensemble structuré de termes d'un vocabulaire, par exemple les termes techniques utilisés en médecine, représentés de façon normalisée par des descripteurs ou des mots clés (Foskett, 1997). Les termes sont organisés de manière conceptuelle et reliés entre eux par des relations sémantiques. Un thésaurus est donc un ensemble organisé de termes, choisis pour leur capacité à faciliter la description d'un domaine et à harmoniser la communication et le traitement de l'information. Chaque terme, appelé descripteur, est aussi peu ambigu que possible et est préféré à des termes voisins ou synonymes, les non-descripteurs, pour tous les échanges significatifs. En pratique, le thésaurus forme un répertoire alphabétique pour l'analyse du contenu, le classement et donc l'indexation de documents, sachant que dans de nombreux cas, les thésaurus proposent également une définition des termes utilisés. En mode consultation et exploitation des données, le thésaurus devient un instrument de recherche : disposant des vocabulaires et règles de l'indexation, l'utilisateur peut optimiser ses requêtes.

Un thésaurus s'élabore comme un sous-ensemble du vocabulaire usuel et d'au moins un vocabulaire spécialisé. C'est un vocabulaire contrôlé puisqu'il résulte d'un long processus de tri des mots, appellations et expressions utilisées de manière informelle dans un domaine particulier. Il s'agit d'une démarche pragmatique de rationalisation des termes descriptifs. Des outils d'analyse automatique de textes permettent l'extraction des termes les plus fréquents d'un corpus et, dans une certaine mesure, facilitent l'émergence de leurs relations sémantiques. Pour construire le thésaurus, les termes ainsi identifiés sont inventoriés, comparés, mis en relation et finalement hiérarchisés pour rendre compte des traits essentiels du domaine. Cette hiérarchie s'appuie sur une typologie : chaque terme appartient à une catégorie qui le situe par rapport à tous les autres termes retenus et qui fixe de cette manière sa priorité d'emploi. La hiérarchie des termes peut tout-à-fait être différente d'un thésaurus à un autre et même, sous réserve d'incohérence, dans un usage ou un autre du même thésaurus. Il demeure toujours une dimension arbitraire dans l'étape de hiérarchisation, soit dans le choix des termes, soit dans leur position hiérarchique bien qu'il existe des normes pour guider l'élaboration des thésaurus[5]. Finalement, en partant du niveau le plus haut correspondant au domaine du thésaurus, nous trouvons en premier les subdivisions majeures représentant les composantes du domaine - subdivisions souvent nommées microthésaurus - puis pour chaque subdivision, la hiérarchie propre aux descripteurs. Un thésaurus peut également concerner plusieurs domaines et plusieurs langues.

Concernant ce que la communauté d'Ingénierie des connaissances nomme « thésaurus sémantique », nous ne dirons ici que quelques mots pour situer cette notion par rapport à celles que nous avons précédemment abordées. C. Roussey *et al.* (2002) séparent explicitement une terminologie du domaine de sa conceptualisation et définissent le thésaurus sémantique comme *« une normalisation des notions du domaine auxquelles sont associées des terminologies »*. Selon cette définition, on peut rapprocher (voir assimiler) les thésaurus sémantiques des serveurs de terminologie médicaux dans lesquels le couple thésaurus-ontologie joue le même rôle que celui assumé par le thésaurus sémantique (*cf.* projet GALEN- section 2.9).

Pour conclure cette section, les thésaurus et les classifications permettent de traduire un message dans un vocabulaire normalisé. Lorsqu'il y a transmission d'information, l'émetteur code le message en fonction d'un langage et du contexte d'énonciation, l'interprétation correcte par le récepteur suppose l'emploi du même langage et la connaissance du contexte. Or le contexte conditionne le codage, ainsi, dans le cas d'un malade hospitalisé pour chimiothérapie d'un cancer ayant développé une aplasie, le dossier serait codé selon l'étiologie cancéreuse par un épidémiologiste mais sous la rubrique *aplasie* si on s'intéresse à la charge en soins.

1.5 Taxinomie

Le mot *taxinomie* vient du grec *taxis*, rangement, et de *nomos*, loi. Il s'agit de la partie de la biologie visant à établir une classification systématique des êtres vivants[6]. Le Petit Robert définit

[5]Norme ISO 2788-1986 : Principes directeurs pour l'établissement et le développement des thésaurus monolingues.
Norme ISO 5964-1985 : Principes directeurs pour l'établissement et le développement des thésaurus multilingues.

[6]Exemple d'une classification visant à couvrir tout le vivant : `http://tolweb.org/tree/`.

la taxinomie comme étant (1) l'étude théorique des bases, lois, règles, principes d'une classification ; (2) une classification d'éléments. Le terme taxonomie fut inventé, sous cette orthographe, par Augustin Pyrame de Candolle pour définir la théorie des classifications. L'orthographe fut corrigée en taxinomie par Émile Littré mais l'autre forme reste pourtant très répandue.

Toutes les classifications se présentent sous la forme d'un arbre (classement arborescent, *cf.* figure 3.3), depuis une racine incluant tous les êtres vivants existants ou ayant existé, jusqu'aux individus. Chaque nœud de l'arbre définit un taxon, qui groupe tous les sous-taxons qu'engendre le nœud.

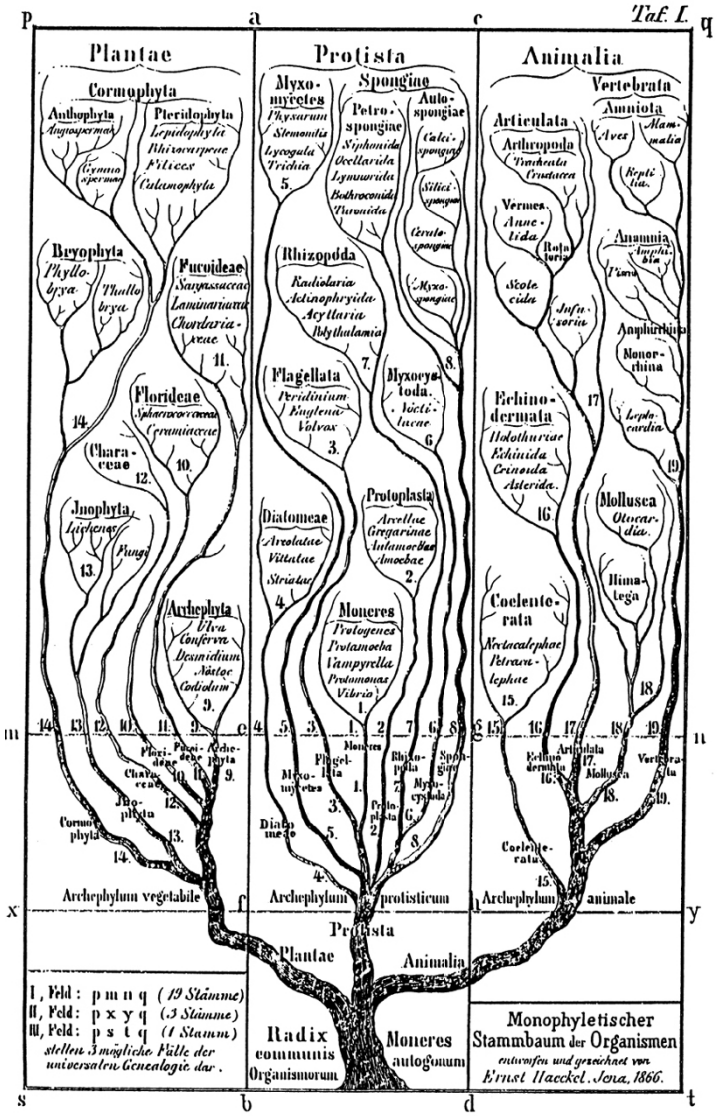


FIG. 3.3 – Arbre phylogénétique des êtres vivants selon Haeckel (1866).

Il n'en a pas toujours été ainsi. Le scientifique suédois Carl von Linné (1707 - 1778) posa

les fondations de la systématique, et fut l'auteur d'une classification dont les grands principes sont la base de la systématique scientifique jusqu'au milieu du XX^{e} siècle. Cette classification traditionnelle, fortement anthropocentrique, fait encore, en ce début du XXI^{e} siècle, partie du bagage culturel commun. Pourtant, elle reflète des causes de la diversité des êtres vivants telles qu'on les pensait voici 250 ans, mais qui n'ont plus rien à voir avec ce que nous en pensons aujourd'hui. En effet, l'anthropocentrisme est battu en brèche avec les théories de Charles Darwin qui recommande en 1859 une classification purement généalogique : s'il y a eu évolution, les espèces doivent être classées selon leur degré d'apparentement évolutif. Il faudra attendre près d'un siècle pour accepter la généalogie comme inaccessible (qui descend de qui ?) et pour se concentrer sur la phylogénie (qui est plus proche de qui ?). C'est dans la deuxième moitié du XX^{e} siècle qu'apparaît l'approche phylogénétique pour laquelle le critère fondamental du choix de la classification est qu'elle doit refléter strictement la phylogénie, c'est-à-dire les degrés d'apparentement entre espèces (Darlu & Tassy, 1993). La notion même d'une telle phylogénie est une conséquence de la théorie de l'évolution, et le succès prédictif des arbres phylogénétiques une des preuves de cette théorie (Lecointre & Le Guyader, 2001).

Dans la culture occidentale, la spécialisation est la méthode la plus intuitive d'organisation des connaissances. Le raisonnement s'appuie sur la structure de la hiérarchie taxinomique qui doit donc être suffisamment rigoureuse et explicite. Le principe d'exclusivité implique que les nœuds fils directs d'un concept père soient disjoints deux à deux. Pour reprendre l'exemple d'O. Dameron dans sa thèse de doctorat (2003), une hiérarchie taxinomique qui décompose *Etre vivant* en *Végétal* et *Animal* puis ce dernier en *Carnivore*, *Herbivore* et *Omnivore* pose le problème du respect de ce principe, le problème du statut de l'héritage multiple et la question de ce qui est contingent et de ce qui ne l'est pas. En effet, si on définit un carnivore comme étant un animal qui ne se nourrit *que* de viande alors le principe d'exclusivité est respecté. Par contre, si on considère qu'un carnivore est un animal qui se nourrit de viande alors le principe n'est pas respecté entre *Carnivore* et *Omnivore*. Ce problème peut être résolu en décomposant *Animal* en *Carnivore* et *Herbivore* tout en autorisant l'héritage multiple pour *Omnivore*. N. Guarino (1999) souligne la nécessité de tenir compte des implications des choix taxinomiques en cas d'héritage multiple afin d'éviter les réductions et les collisions de sens ainsi que les généralisations abusives. *Animal* pourrait également être défini par d'autres attributs que celui de la nature de sa nourriture, comme par exemple en *Vertébré* et *Invertébré*[7]. Cet exemple, bien que très simplifié, montre que décider du définitoire et du contingent est loin d'être évident et qu'il est donc difficile de repérer et classifier les objets d'un domaine.

1.6 Ontologie

Depuis le début des années 90, l'Ingénierie des connaissances a grandement contribué à diffuser le terme « ontologie ». T.R. Gruber reprend ce terme – précédemment utilisé par J.F. Sowa, C.S. Pierce et R.J. Brachman – et l'introduit en informatique. Il ne fait alors pas référence à Aristote ou au domaine de la philosophie mais bien au domaine de la sémantique au sens où

[7] Il faut noter que cette dernière proposition de classification va à l'encontre de la phylogénie moderne puisqu'elle se base sur ce que les animaux n'ont pas en commun plutôt que sur ce qu'ils ont en commun.

l'entendent les théoriciens de la représentation des connaissances. L'ontologie, en tant que pratique, a été définie par Aristote comme étant la science de l'Être. Le terme lui-même date du XVII^{e}. Bien que des débats préexistent, la naissance de l'Ingénierie des connaissances puis l'essort du Web sémantique placent la définition, la construction et l'utilisation d'ontologie au centre de nombreux travaux. On parle plus volontiers d'ontologies (au pluriel) afin de refléter les multiples facettes que recouvre cette appellation. Plusieurs auteurs, par exemple N. Guarino (1996) ou encore O. Dameron (2003), passent en revue les différentes définitions de la littérature afin d'examiner le type de représentation des connaissances dénoté par le terme ontologie. En 1993, T.R. Gruber propose une première définition et introduit la notion de conceptualisation : *« une ontologie est une spécification partagée d'une conceptualisation »*, (Gruber, 1993). Cette définition permet de nombreuses interprétations et demande à être clarifiée et précisée. Les travaux menés par N. Guarino et C. Welty permettent de définir cette notion de conceptualisation (Guarino & Welty, 2000a). Il s'agit, d'une part, de préciser la notion de domaine d'application ou de discours en distinguant parmi les entités peuplant un domaine, les individus et les propriétés, et, d'autre part, de préciser quelles sont les propriétés essentielles, unicité, identité …. Contrairement aux définitions précédentes, la définition proposée par B. Bachimont s'appuie directement sur la représentation des connaissances en tant que domaine d'étude et sur la formalisation en tant que moyen de mise en œuvre. Dans ce contexte, l'ontologie réconcilie la dimension sémantique et la dimension syntaxique nécessaire aux langages formels informatisés (Bachimont, 2000) :

> *« Définir une ontologie pour la représentation des connaissances, c'est définir, pour un domaine et un problème donnés, la signature fonctionnelle et relationnelle d'un langage formel de représentation et la sémantique associée. »*

L'ontologie permet donc de fixer une sémantique aux objets primitifs de la représentation d'un domaine. Pour cela, il convient d'identifier quelles sont les notions élémentaires qui, assemblées et combinées, vont constituer l'ensemble des connaissances représentées pour le domaine en question. Nous verrons en section 1 l'approche que B. Bachimont propose pour traiter cette difficulté. L'ontologie est également un artefact informatique qui doit être opérationnel. Elle devient, dans ce but, un ensemble de concepts et de relations spécifiques du domaine et atteint un niveau d'abstraction supplémentaire. Reprenant les travaux de T.R. Gruber (1993) et ceux de M. Uschold et M. Grüninger (1996), J. Charlet propose une définition rigoureuse et affinée de ce qu'est une ontologie (Charlet, 2002). C'est cette définition que nous retiendrons dans la suite de ce manuscrit.

> *« Une ontologie implique ou comprend une certaine vue du monde par rapport à un domaine donné. Cette vue est souvent conçue comme un ensemble de concepts* – e.g. *entités, attributs, processus –, leurs définitions et leurs interrelations. On appelle cela une conceptualisation. »*
>
> *[…]*
>
> *« Une ontologie peut prendre différentes formes mais elle inclura nécessairement un vocabulaire de termes et une spécification de leur signification. »*
>
> *[…]*
>
> *« Une ontologie est une spécification rendant partiellement compte d'une conceptualisation. »*

Cette définition précise les précédentes et introduit ce que sont les constituants de l'ontologie.

1.6.1 Constituants d'une ontologie

Comme nous l'avons dit précédemment, les ontologies rassemblent les connaissances propres à un domaine donné. En représentation des connaissances, ces ontologies existent sous la forme de concepts et de relations, et permettent d'en fixer la sémantique selon un degré de formalisme variable. Nous plaçons la formalisation ontologique des connaissances dans un monde tel que décrit dans la théorie des modèles. L'hypothèse de départ est donc qu'il existe des objets individuels qui peuvent être énumérés. Les concepts et les relations de l'ontologie sont organisés sous une forme hiérarchique qui admet une relation de subsomption. Nous détaillons ci-dessous les différents composants de l'ontologie considérée en tant qu'objet informatique.

1.6.1.1 Concepts

Les connaissances portent sur des objets auxquels on fait référence à travers des concepts (également appelés classes dans certains travaux). Un concept peut représenter un objet matériel (par exemple, un comprimé de médicament), une notion (par exemple, la quantité) ou bien une idée (Uschold & King, 1995). Un concept peut être divisé en trois parties, un terme (que nous désignerons sous le nom de label), une notion et un ensemble d'objets. Le label d'un concept est l'expression linguisitique utilisée couramment pour y faire référence. La notion désigne ce qui est appelé, au sens de la représentation des connaissances, l'intension du concept. Elle contient sa sémantique qui est définie à l'aide de propriétés, d'attributs, de règles et de contraintes. L'ensemble d'objets forme ce qui est appelé l'extension du concept. Il s'agit des objets auxquels le concept fait référence, autrement dit, de ses instances. Par exemple, le label de concept « hôpital » renvoie aussi bien à la notion d'hôpital en tant que lieu où l'on soigne qu'à l'ensemble des objets de ce type : hôpital du Kremlin-Bicêtre, hôpital de la Pitié-Salpêtrière ...

L'intension et l'extension d'un concept sont deux aspects bien distincts : deux extensions peuvent ne pas être des ensembles disjoints tandis que deux intensions ont comme propriété de s'exclure mutuellement. B. Bachimont (2000), distingue ainsi le « concept formel », qui désigne l'extension d'un concept et admet une sémantique référentielle[8], du « concept sémantique » qui désigne l'intension d'un concept et admet une sémantique différentielle. R. Troncy illustre ces aspects avec un exemple extrait d'une ontologie du cyclisme :

> *« La comparaison des extensions permet de définir une relation d'héritage extensionnelle entre les concepts : un concept sera subsumé par un autre si et seulement si son extension est incluse dans celle de son parent. [...] Les transcriptions des notions différentielles en concepts formels peuvent admettre des extensions qui ont un sous-ensemble commun. Ainsi,* `Rouleur` *et* `Grimpeur` *sont bien des notions (différentielles) qui s'excluent, mais les concepts formels correspondants ont des extensions qui peuvent avoir en commun plusieurs individus : par exemple l'individu* `LanceArmstrong` *est habituelement répertorié dans ces deux catégories. On peut donc définir dans l'ontologie référentielle un concept formel*

[8] Les sémantiques différentielle et référentielle sont expliquées en détails au chapitre 4, sections 1.1 et 1.2

> `Grimpeur-Rouleur` *dont la référence sera l'intersection des extensions des concepts* `Rouleur` *et* `Grimpeur`*, et dont* `LanceArmstrong` *sera un élément » (Troncy, 2004).*

Attention cependant, des concepts peuvent partager une même extension mais pas une même intension et être désignés par le même terme. Cette difficulté est l'écho des différents points de vue que l'on peut avoir sur un même objet. C'est le cas pour « table » qui est à la fois un meuble constitué d'un plateau et de quatre pieds et un meuble sur lequel on peut poser des objets et autour duquel on peut s'asseoir : les extensions sont les mêmes mais les intensions sont différentes. Il s'agit d'un cas typique d'homonymie de termes. Dans le même ordre d'idées, il paraît indispensable de gérer les synonymies. Ces deux problèmes compliquent la tâche de l'ingénieur des connaissances lorsqu'il doit choisir comment désigner, dans la langue, un concept. Certains auteurs, parmi lesquels (Gómez-Pérez *et al.*, 1996), soulignent la nécessité de désigner un concept par plusieurs termes. Nous proposons plutôt de désigner un concept par un seul label et de relier ce label à un ensemble de termes préférés. Cela a l'avantage de marquer une différence nette entre le statut de label et celui de terme utilisé dans la langue. Par exemple, le label « Cephalee » a comme ensemble de termes préférés « céphalée, migraine, céphalgie ».

1.6.1.2 Propriétés

Les propriétés sont des caractéristiques valuées attachées aux concepts. Une des grandes difficultés de l'ingénieur des connaissances qui développe une ontologie est de choisir si une connaissance doit être modélisée sous la forme d'une propriété ou bien à l'aide d'une relation liant un autre concept. Une solution courante est de choisir la propriété lorsque la valeur de la connaissance est un entier ou une chaîne de caractères. Les valeurs que peut admettre une relation sont d'un genre plus complexe puisqu'il s'agit d'un autre concept présent dans l'ontologie.

1.6.1.3 Relations

Les relations représentent un type d'interaction entre les concepts d'un domaine. Elles lient les concepts primitifs (ou simples) entre eux pour construire des représentations conceptuelles complexes que nous appelons concepts définis. Elles sont caractérisées par un terme, voire plusieurs, et une signature qui précise le nombre d'instances de concepts que la relation lie, leurs types et l'ordre de ces concepts, c'est-à-dire la façon dont la relation doit être lue. Par exemple, la relation « diagnostique » lie une instance du concept « personnelMedical » et une instance du concept « pathologie », dans cet ordre. Des exemples de relations binaires sont : « observe-par », « associe-a », « qualifie-de », ou encore « connecte-a ». G. Kassel (2002) et N. Guarino *et al.* (1995) formalisent les principales relations jugées utiles à la modélisation d'une ontologie : « instance-de », « sorte-de », « appartenance-a », « dependance » ... Le point de vue que nous venons de présenter sur les relations est influencé par les logiques de description. Il est propre à notre modèle d'ontologie (et aux graphes conceptuels) mais n'est pas le seul que l'on pourrait choisir.

La relation de subsomption « est-un » (is-a) a un statut particulier car elle structure la hiérarchie ontologique. À ce titre, elle est implicite. Un concept C1 (concept père) subsume un concept C2 (concept fils) si toute propriété sémantique de C1 est également une propriété sémantique de

C2 et si C2 est plus spécifique que C1 (*cf.* figure 3.4). Ainsi, l'extension d'un concept est forcément plus réduite que celle de son concept père. Son intension est par contre plus riche. La relation de subsomption est définie dans la littérature de plusieurs manières :

- Définition intensionnelle : un concept C1 subsume un concept C2 si tout individu décrit par C2 l'est aussi par C1, autrement dit si l'ensemble des propriétés d'un individu dont la description est définie par C2 contient l'ensemble des propriétés spécifiées par C1. Par exemple, l'ensemble des propriétés associées au concept « scintigraphie » comprend l'ensemble des propriétés associées au concept « examenIsotopique ».
- Définition extensionnelle : un concept C1 subsume un concept C2 si l'ensemble des individus dénotés par C1 contient l'ensemble des individus dénotés par C2. Par exemple, le concept « pathologie » subsume le concept « pneumonie ».
- Définition logique : un concept C1 subsume un concept C2, si être un individu décrit par C2 implique être un individu décrit par C1.

La relation de subsomption n'est pas la seule relation qui permette de structurer la hiérarchie ontologique. Le domaine de la médecine, notamment la représentation des connaissances anatomiques, utilise souvent la relation de méronymie, « partie-tout » (part-of).

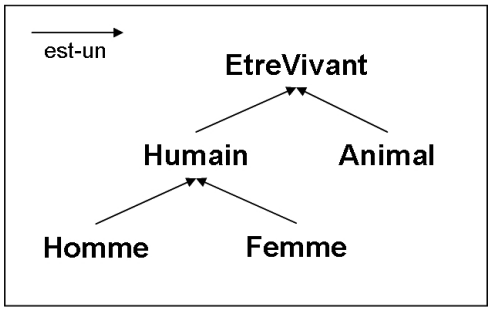


FIG. 3.4 – Exemple de la relation de subsomption.

1.6.1.4 Instances

Les instances d'un concept sont des éléments singuliers. Par exemple, la radiographie du membre inférieur droit de madame X est une instance du concept « radiographie ».

1.6.2 Différents types d'ontologies

Nous listons ci-dessous les différents types d'ontologies les plus courants dans la littérature (Gómez-Pérez *et al.*, 2004a). Nous reprenons, à cet effet, les travaux de R. Mizoguchi *et al.* (1995), G. van Heijst *et al.* (1997) et N. Guarino (1998). Cet état de l'art présente plusieurs typologies connues qui classifient les ontologies en fonction des objets qu'elles modélisent, du degré de granularité des connaissances représentées et du niveau de formalisme du modèle. Nous ne discuterons pas ici de la généricité des ontologies.

1. La classification faite par M. Uschold et M. Grüninger (1996) distingue les ontologies suivant le type de langage utilisé et donc en fonction du degré de formalisme de la représentation :
   - les ontologies hautement informelles sont des ontologies opérationnelles écrites en langage naturel ;
   - les ontologies semi-informelles utilisent un langage naturel structuré et limité ;
   - les ontologies semi-formelles définissent les concepts dans un langage artificiel et défini formellement ;
   - les ontologies rigoureusement formelles sont définies dans un langage contenant une sémantique formelle, des théorèmes et des preuves de propriétés telles que la robustesse et l'exhaustivité.
2. Comme le proposent A. Gómez-Pérez *et al.* (2004a) et N. Guarion (figure 3.5), la classification peut également se faire en fonction des objets que modélisent les ontologies pour répondre à un objectif précis :

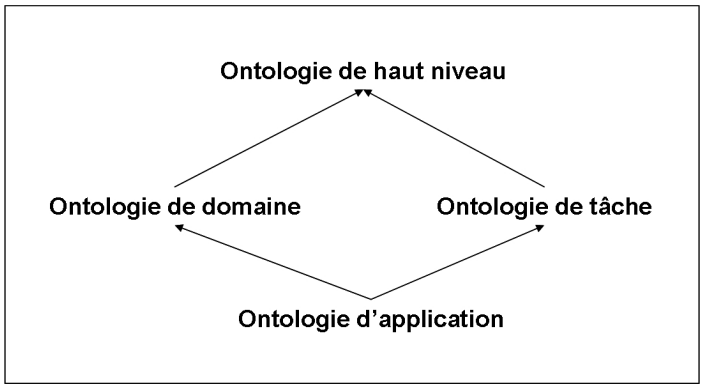


FIG. 3.5 – Classification des ontologies selon N. Guarino (1998).

   - Ontologies pour la représentation des connaissances (van Heijst *et al.*, 1997)

     Ce type d'ontologies regroupe les concepts utilisés pour formaliser les connaissances. Parmi les ontologies de représentation, on trouve des ontologies qui vont décrire les notions utilisées dans toutes les ontologies pour spécifier les connaissances, telles que les substances, les concepts, les relations ... La « Frame-Ontology » est un bon exemple d'ontologie de représentation. Elle définit, de manière formelle, les concepts utilisés principalement dans des langages à base de frames : classes, sous-classes, attributs, valeurs, relations et axiomes (Gruber, 1993). Selon N. Guarino (1994), les ontologies de représentation sont en fait indépendantes des différents domaines de connaissances, puisqu'elles décrivent des primitives cognitives communes aux différents domaines.
   - Ontologies de domaine (Mizoguchi *et al.*, 1995; van Heijst *et al.*, 1997)

Comme il a été dit précédemment, les ontologies vont permettre de spécifier les connaissances d'un domaine, de façon aussi indépendante que possible du type de manipulations qui vont être opérées sur ces connaissances. Ces ontologies sont appelées « ontologies de domaine », puisqu'elles sont construites sur un domaine particulier de la connaissance. Elles rendent compte du vocabulaire d'un domaine spécifique au travers de concepts et de relations qui modélisent les principales activités, les théories et les principes de base du domaine en question. De nombreuses ontologies de domaine existent déjà, telles que MENELAS dans le domaine médical (Zweigenbaum, 1999). Ce type d'ontologies est bien souvent raccroché à des ontologies de haut niveau de conceptualisation.

– Ontologies de haut niveau (Mizoguchi & Ikeda, 1997; Guarino, 1997; Guarino *et al.*, 1995)

  Une dinstinction est établie entre les ontologies de domaine portant sur des concepts renvoyant à des objets matériels ou à des concepts d'assez bas niveau (c'est-à-dire n'offrant que des possibilités limitées de raffinement) et les ontologies portant sur des concepts de haut niveau. Ces dernières décrivent des notions générales comme les notions d'objet, de propriété, d'état, de valeur, de moment, d'événement, d'action, de cause et d'effet (Sowa, 2000). En théorie, les ontologies de haut niveau doivent pouvoir être reliées au sommet des ontologies de domaine. Là encore, définir des critères précis permettant de structurer l'ontologie est problématique. En effet, comme le montre la figure 3.6, il existe autant de manières de classifier les objets qu'il y a d'ontologies de haut niveau différentes.

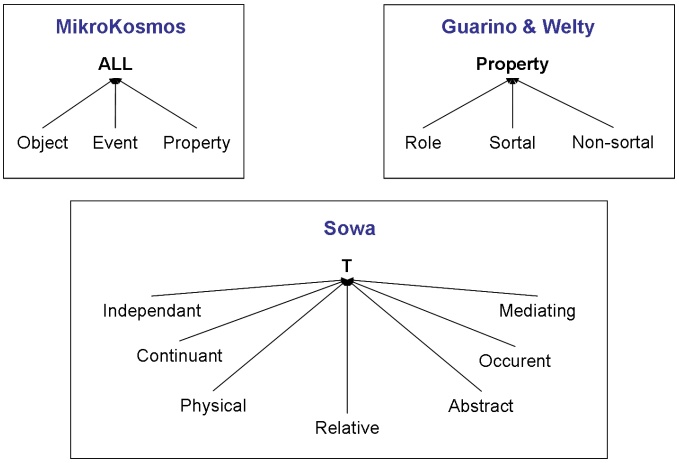


FIG. 3.6 – Exemple d'ontologies de haut niveau.

– Ontologies génériques (Mizoguchi *et al.*, 1995; van Heijst *et al.*, 1997)

  L'ontologie générique, aussi appelée méta-ontologie ou noyau d'ontologie, véhicule des connaissances génériques qui, bien que moins abstraites que celles modélisées dans l'ontologie de haut niveau, doivent être assez générales pour être réutilisées

dans différents domaines. Elle organise des connaissances factuelles ou des connaissances visant à résoudre des problèmes génériques d'un ou de plusieurs domaines. Ce type d'ontologies modélise des connaissances se rapportant aux choses, évènements, temps, espace, causalité … L'ontologie méréologique (*Mereology Ontology*) est un exemple classique d'ontologie générique (Borst, 1997).

- Ontologie de tâches (Mizoguchi *et al.*, 1995; Guarino, 1998)

  L'ontologie de tâche décrit les connaissances portant sur des tâches et/ou des activités particulières (faire un diagnostic, planifier une activité …). Ces ontologies fournissent un ensemble de termes au moyen desquels on peut décrire, au niveau générique, comment résoudre un type de problème. Selon R. Mizoguchi, cette ontologie caractérise l'architecture computationnelle d'un système à base de connaissances qui réalise une tâche. On peut les assimiler à l'inventaire des rôles, tâches et méthodes utiles à la description de la résolution de problèmes, à savoir les niveaux d'inférence et de tâches de CommonKADS.

- Ontologies d'application (van Heijst *et al.*, 1997)

  Ce sont les ontologies les plus spécifiques, elles contiennent les connaissances requises pour une application particulière. Les ontologies d'application étendent et spécialisent les connaissances contenues dans l'ontologie de domaine et dans l'ontologie de tâche pour une application donnée. Selon A. Maedche et S. Staab (2001), les concepts dans l'ontologie d'application correspondent souvent aux rôles joués par les objets du domaine tout en exécutant une certaine activité, par exemple : *hypothèse*, *signe*, *diagnostic* … L'ontologie d'application telle que définie par A. Maedche et S. Staab équivaut à l'ontologie de tâches pour N. Guarino.

3. Enfin, la dernière classification que nous présentons ici est en fonction du niveau de granularité, c'est-à-dire du niveau de détail des objets de la conceptualisation. En fonction de l'objectif opérationnel, une connaissance plus ou moins fine du domaine est nécessaire et des propriétés considérées comme accessoires dans certains contextes peuvent se révéler indispensables pour d'autres applications :

   - Granularité fine : correspondant à des ontologies très détaillées, possédant ainsi un vocabulaire plus riche capable d'assurer une description détaillée des concepts pertinents d'un domaine ou d'une tâche.

   - Granularité large : correspondant à un vocabulaire moins détaillé. Les ontologies de haut niveau ont une granularité large, du fait que les notions sur lesquelles elles portent peuvent être raffinées par des notions plus spécifiques (Fürst, 2004a).

Il est évident qu'il est difficile de faire la différence, pour une ontologie considérée, entre ces différentes classifications et de choisir celle qui correspond. Ainsi, l'ontologie de la pneumologie réalisée dans le cadre de notre doctorat participe à la fois de l'ontologie de domaine, de l'ontologie de tâche (activité de codage) et de l'ontologie d'application puisqu'elle est destinée à être le cœur d'un outil d'aide au codage.

1.6.3 Cycle de vie d'une ontologie

Puisque les ontologies sont destinées à être utilisées comme des composants logiciels dans des systèmes répondant à des objectifs opérationnels différents, leur développement doit s'appuyer sur les mêmes principes que ceux appliqués en génie logiciel. Ainsi, les ontologies doivent être considérées comme des objets techniques évolutifs et possédant un cycle de vie qui nécessite d'être précisé. Dans ce contexte, les activités liées aux ontologies sont, d'une part, des activités de gestion de projet (planification, contrôle, assurance qualité), et, d'autre part, des activités de développement (spécification, conceptualisation, formalisation) ; s'y ajoutent des activités transversales de support telles que l'évaluation, la documentation, la gestion de la configuration (Blasquez *et al.*, 1998).

Un cycle de vie inspiré du génie logiciel est proposé dans (Dieng *et al.*, 2001) et (Gandon, 2006). Nous l'avons adapté à nos besoins et proposons notre vision du cycle de vie d'une ontologie (*cf.* figure 3.7). Il comprend une étape initiale de détection et de spécification des besoins qui permet notamment de circonscrire précisément le domaine de connaissances, une étape de conception qui se subdivise en trois phases qui seront détaillées dans la section 1, une étape de déploiement et de diffusion, une étape d'utilisation, une étape, incontournable, d'évaluation, et enfin, une sixième étape consacrée à l'évolution et à la maintenance du modèle. Après chaque utilisation significative, l'ontologie et les besoins doivent être réévalués et l'ontologie peut être étendue et, si nécessaire, en partie reconstruite. La validation du modèle de connaissances est au centre du processus et se fait de manière itérative.

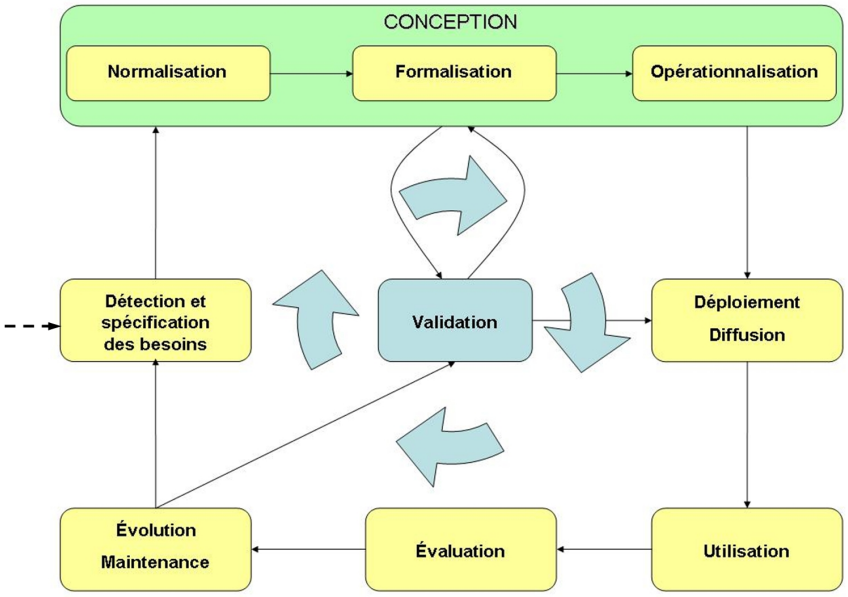


FIG. 3.7 – Cycle de vie d'une ontologie.

M. Fernandez (1997) insiste sur le fait que les activités de documentation et d'évaluation sont nécessaires à chaque étape du processus de construction, l'évaluation précoce permettant de

limiter la propagation d'erreurs. Le processus de construction peut et doit être intégré au cycle de vie d'une ontologie comme indiqué en figure 3.7.

2 Ressources terminologiques et ontologiques en médecine

Il existe dans le domaine médical un grand nombre de ressources terminologiques et ontologiques (RTO) construites pour répondre à des besoins précis et divers : la CIM (*cf.* section 2.1) et la CCAM (*cf.* section 2.2) sont des classifications utilisées pour le codage médico-économique des dossiers patients à des fins statistiques et budgétaires, le thésaurus MeSH (*cf.* section 2.3) vise à indexer les connaissances médicales pour la recherche d'information dans des bases documentaires, le catalogue CISMeF (*cf.* section 2.4) utilise pour partie les termes du MeSH ainsi que d'autres, la SNOMED (*cf.* section 2.5) est une nomenclature dédiée au codage des dossiers électroniques des patients mais avec une granularité plus fine, l'UMLS (*cf.* section 2.6) a pour objectif d'améliorer l'accès à l'information médicale à partir de sources diverses, le FMA (*cf.* section 2.7) est une ontologie de l'anatomie, DOLCE (*cf.* section 2.8) est une ontologie de haut niveau construite pour être reliée à des ontologies de domaine, GALEN (*cf.* section 2.9) est une ontologie médicale à visée généraliste, MENELAS (*cf.* section 2.10) est une ontologie ciblée sur les pathologies coronariennes ... Ces ressources font toutes partie, bien qu'à des degrés divers, de notre univers de recherche. Nous les décrivons dans les sections suivantes indépendamment les unes des autres.

2.1 CIM

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. Elle est publiée par l'Organisation Mondiale de la Santé 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 ont pris ces dernières années une part croissante. Elle bénéficie d'une remise à niveau régulière, la version la plus récente étant la 10e révision (CIM-10 [9], publiée en 1993). Il s'agit d'une classification monoaxiale avec 21 chapitres principaux dont 17 concernent des maladies et 4 concernent les signes et résultats anormaux, les causes de traumatismes, d'empoisonnement ou de morbidité, l'état de santé et les facteurs de recours aux soins. Les catégories de maladies sont définies en fonction d'un caractère commun qui peut être l'étiologie (1 = maladies infectieuses, lettres A et B), la topographie (9 = maladies de l'appareil circulatoire, lettre I), la physiologie (15 = grossesse et accouchement, lettre O) ou la pathologie

[9] http://www.med.univ-rennes1.fr/noment/cim10/

(II = tumeurs). Les affections (symptômes, maladies, lésions traumatiques, empoisonnements) et les autres motifs de recours aux soins sont répertoriés dans la CIM avec une précision qui dépend de leur importance, c'est-à-dire de leur fréquence et de l'intensité du problème de santé publique qu'ils posent. Par exemple, le chapitre des maladies infectieuses est le plus gros et le plus détaillé parce que ces maladies sont la première cause mondiale de morbidité et de mortalité. La classification aboutit par subdivisions successives à un code à trois caractères (une lettre correspondant au chapitre puis deux chiffres) pour les maladies définies à un niveau général, décliné par l'ajout d'un quatrième chiffre (derrière un point) pour désigner les diagnostics précis et les formes cliniques ; le sous-code 9 désignant l'absence de précision (SAI = sans autre indication) et le sous-code 8 les autres formes non précédemment définies. Par exemple, le code B25.0 désigne une pneumopatie à cytomégalovirus, le code B25.8 désigne les autres maladies à cytomégalovirus et le code B25.9 désigne une maladie à cytomégalovirus, sans précision. Dans certains cas, un cinquième chiffre a été rajouté afin d'améliorer la finesse de description. Par exemple, le code M66.58 désigne une déchirure spontanée d'un tendon, sans précision, dont le siège n'est pas précisé. La CIM-10 a introduit la notion de troubles iatrogènes[10]. Elle compte au total 16 390 entrées (date de dernière actualisation, janvier 2006) et comprend trois volumes.

La contrainte d'avoir un seul arbre hiérarchique implique qu'une entité pathologique soit représentée une seule fois dans la classification, ce qui pose des difficultés. Ainsi les tumeurs sont extraites de leur chapitre d'appareil et regroupées dans un chapitre spécial. Parfois cependant, une même maladie peut apparaître en deux places distinctes (avec deux codes). C'est le cas lorsqu'une maladie appartient à un processus pathologique initial général (code associé à une dague), par exemple la tuberculose, et correspond à des manifestations localisées à un appareil (code associé à un astérisque), par exemple une tuberculose rachidienne. De plus, le principe de différenciation n'est pas constant. La classification de l'Organisation Mondiale de la Santé sert, en France, au codage des causes de décès ainsi qu'au regroupement des séjours hospitaliers en groupes homogènes de malades dans le cadre du PMSI.

2.2 CCAM

Avant la Classification Commune des Actes Médicaux[11] (CCAM) existaient en France deux nomenclatures des actes médicaux : la Nomenclature Générale des Actes Professionnels (NGAP) et le Catalogue Des Actes Médicaux (CDAM). La NGAP est la nomenclature de la médecine ambulatoire. Elle permet la tarification des actes de médecine libérale. Il s'agit d'une liste de libellés d'actes assortis de cotations qui fixent les honoraires des professionnels – médecins, dentistes, sages-femmes et auxiliaires médicaux – du secteur libéral. Le CDAM permet de décrire l'ensemble des actes réalisés lors de l'hospitalisation d'un patient. Plus récent que la NGAP, le CDAM a vu le jour lors de la mise en place du programme de médicalisation des systèmes d'information, en 1985. Il s'agit d'une nomenclature des actes médicaux comportant pour chaque acte un code, un libellé, un indice de coût relatif et, le cas échéant, la lettre Y indiquant le groupe

[10] Les troubles iatrogènes sont occasionnés par le traitement médical qu'il y ait ou non erreur dans le traitement prescrit.

[11] `http://www.cnamts.fr/san/ccam/somccam.htm`

homogène de malades.

La réglementation contraint les établissements de soins et les professionnels à utiliser simultanément ces deux nomenclatures conçues pour des objectifs différents. Dans ce contexte, un groupe de travail s'est réuni d'avril 1994 à fin 1995 pour étudier la faisabilité d'une classification commune des actes médicaux c'est-à-dire une liste unique de libellés et de codes dont le principe serait étendu, à terme, à l'ensemble des professions de santé. L'élaboration de la CCAM par le Pôle Nomenclature de la CNAMTS et par le Pôle d'Expertise et de Référence National des Nomenclatures de Santé (PERNNS) a été lancée en 1996. Elle remplace maintenant les deux nomenclatures NGAP et CDAM. Elle sert à la fois au programme de médicalisation des systèmes d'information dans tous les établissements publics et privés, et aux praticiens du secteur libéral (médecins et dentistes) pour leurs honoraires. La CCAM est classée par grands appareils et non par spécialités. Elle comprend 17 chapitres, tels que : le système nerveux central, périphérique et autonome, les oreilles, le système cardiaque et vasculaire, le système immunitaire et hématopoïétique, le système respiratoire, le système digestif ... Chaque libellé comporte la mention de deux axes obligatoires - l'action et la topographie - et de deux axes facultatifs - la voie d'abord et la technique utilisée, par exemple : *Biopsie / du rein / par voie transcutanée / sans guidage*. La CCAM a été construite avec l'ontologie GALEN (*cf.* section 2.9)

2.3 MeSH

Le MeSH est un thésaurus médical qui compte, dans sa version 2005, 22 995 mots clés (ou descripteurs), 83 qualificatifs et environ 57 000 synonymes. Il a été conçu à la National Library of Medicine aux Etats-Unis comme support de l'Index Medicus, répertoire des principales publications scientifiques, et est utilisé par les systèmes de recherche bibliographique Medlars et MEDLINE [12]. Il est traduit en français par l'INSERM [13] et sert aussi de thésaurus au site CISMeF (*cf.* section 2.4). Le MeSH est organisé en deux parties : une liste alphabétique de termes (lexique) et une structure multiaxiale. Les 200 000 termes du lexique sont distribués selon 15 axes, allant de l'anatomie à la géographie. Les termes équivalents sont rapportés à celui des 20 000 termes principaux (descripteurs) qui exprime le mieux le concept, termes auxquels sont associés un code alphanumérique. Les descripteurs s'organisent selon une structure hiérarchique et associative qui permet, par exemple, de répondre à une requête sur les virus en proposant aussi des documents sur les antiviraux ou les vaccins antiviraux. Le MeSH comprend jusqu'à neuf niveaux de profondeur. Ces principaux composants sont les *Headings* (MH pour *Main Headings* par la suite), les *Subheading* et les *Supplementary Concept Records*. En outre, des connecteurs permettant des références explicites entre termes expriment les relations de synonymie, de voisinage ou d'association tandis que des qualificatifs permettent de considérer les différentes facettes d'un concept (par exemple : *tuberculose/traitement*). Les MH respectent un certain nombre de propriétés (Nelson *et al.*, 2001) : d'abord, ils couvrent tout le champ de la médecine qu'on veut bien leur faire couvrir et ne se recoupent pas les uns les autres. Ils forment une partition du domaine.

[12] Accessible grâce au moteur de recherche Pubmed sur le site
`http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?`

[13] `http://dicdoc.kb.inserm.fr:2010/basismesh/mesh.html`

Les seuls recouvrements acceptés sont ceux de généralisation (*broader-than*) et de spécialisation (*Narrower-than*) mis en œuvre dans leur structure hiérarchique. Ces arbres proposent des hiérarchies selon plusieurs points de vue et partagent les mêmes MH. Il est alors évident que ceux-ci ne peuvent être des concepts : ils représentent un ou plusieurs concepts et constituent des classes de descripteurs (appelés par la suite, simplement, descripteurs). On voit, figure 3.8, le descripteur « nez » impliqué dans trois hiérarchies de l'anatomie, une liée aux régions du corps [A01], une autre au système respiratoire [A04] et une dernière aux organes sensitifs [A09]. Chaque descripteur a pour label un terme préférentiel pris parmi les termes préférentiels de chacun de ses concepts (dans notre exemple, le nez dans les trois hiérarchies différentes). Les relations de subsomption des hiérarchies sont principalement des hyperonymies et des partinomies mais, dans le domaine de la recherche d'informations, qui est celui du MeSH, on trouve des relations liées au sujet d'intérêt comme dans l'exemple de la hiérarchie des accidents qui subsume « prévention des accidents ». V. Malaisé (2005) souligne les limites du MeSH et des systèmes documentaires dans lesquels il est utilisé. Alors qu'un certain nombre de propriétés du MeSH lui donnent un statut proche de celui des ontologies (objets conceptuels, partitions, arbres), les auteurs lui ont conservé une orientation vers la recherche d'informations, avec des liens de sujets d'intérêts ou des conceptualisations reflétant une vue de la littérature des usagers (*cf.* figure 3.8) et pas une conceptualisation destinée à permettre des inférences médicales. Stuart J. Nelson *et al.* l'affirment dans le paragraphe suivant :

> *« Many individuals have to tried to use Me*SH *as a concept representation language with only modest succes. That the relationships in the Me*SH *tree structure were designed with a different view, and with a different (and not formal) meaning of « broader-than », has frustated their efforts. The Me*SH *hierarchical structure was designed to reflect a view of the literature for a user.[…] The trees thus indicate what appears to be a useful set of relationships, based on the perceived needs of searchers*[14] *»* (Nelson *et al.*, 2001).

2.4 CISMeF

Le site du Catalogue et Index des Sites Médicaux Francophones[15], CISMeF a pour but « d'assister les professionnels de santé dans leur quête d'informations » (Thirion *et al.*, 1999). Il s'agit d'un catalogue et d'un index spécialisé référençant les sites médicaux francophones répondant à un critère de « qualité de l'information de santé sur l'Internet (NetScoring) ». Ce catalogue est le fruit d'un projet initié en février 1995 par le Centre Hospitalier Universitaire de Rouen.

Il existe d'autres catalogues médicaux, mais nous ne présenterons ici que CISMeF parce qu'il s'agit du catalogue médical francophone le plus fréquemment renvoyé lors de requêtes

[14] « Un grand nombre d'individus ont essayé d'employer le MeSH comme langage de représentation conceptuel et n'ont obtenu que des succès modestes. Les relations dans la structure arborescente du MeSH ont été conçues selon une vue différente, avec une signification différente (non formelle) de plus « large-que », cela a contrarié leurs efforts. La structure hiérarchique du MeSH a été conçue pour refléter une vue de la littérature pour un utilisateur.[…] Les arbres indiquent ainsi ce qui semble être un ensemble utile de relations, basé sur les besoins apparents des chercheurs. »

[15] `http://www.chu-rouen.fr/cismef/`

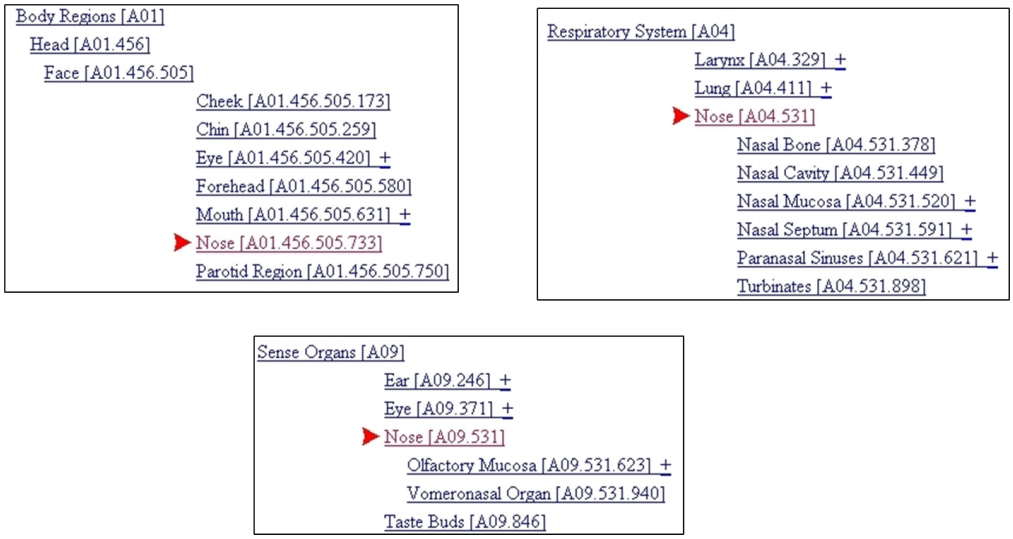


FIG. 3.8 – Extrait du MeSH pour le descripteur du nez.

génériques portant sur le domaine médical, à partir de moteurs de requête eux aussi génériques (Zweigenbaum *et al.*, 2002).

Les bases documentaires de CISMeF sont indexées manuellement[16] : quand une page web est cataloguée, elle est indexée pour pouvoir être retrouvée et reproposée aux utilisateurs. Le langage RDF (*cf.* section 5.2) est utilisé pour décrire des informations bibliographiques à l'aide des balises de métadonnées définies par le Dublin Core[17] (Darmoni & Thirion, 2000) : créateur de la ressource (`<dc:Creator>`), langue utilisée dans le document (`<dc:Language>`), sujets abordés (`<dc:Subject>`), type de lecteur souhaité (`<dc:Audience>`) ... À chaque document est associé un certain nombre de descripteurs caractérisant son genre (cours, conférence, séminaire, etc.), son url d'accès et les sujets qu'il aborde. Ces derniers sont extraits du thésaurus MeSH, ou de sa version française traduite par l'INSERM. CISMeF utilise le MeSH mais propose des méta-termes. Les méta-termes correspondent à des spécialités biologiques ou médicales concernées par un ou plusieurs mots clés (ou arborescences de mots clés), qualificatifs, ou types de ressources. Par exemple pour le terme *cancérologie*, on trouve les arborescences, mots clés et qualificatifs du thesaurus MeSH concernant cette spécialité : *antinéoplasiques (arb)*, *marqueur biologique tumeur (arb)*, *oncologie médicale (arb)*, *secondaire (qualificatif)*, *service oncologie hôpital (MeSH)*, *service oncologie hôpital (type de ressource)*, *tumeurs (arb)*. Le catalogue CISMeF se consulte par l'intermédiaire de l'interface[18] Doc'CISMeF.

[16] Concernant les modes d'indexation, le lecteur peut se reporter aux articles suivants (Darmoni *et al.*, 2000) et (Darmoni & Thirion, 1996).

[17] `http://dublincore.org/`

[18] `http://doccismef.chu-rouen.fr/`

2.5 SNOMED

La SNOMED (*Systematized Nomenclature of Medicine*[19]) combine une nomenclature de plus de 50 000 termes et une classification multiaxiale et multi-domaines comportant à l'origine 7 axes : topographie, morphologie, étiologie, altération fonctionnelle, nosologie, actes médicaux (College of American Pathologists, 1993). La 3e édition compte désormais 200 000 termes et 11 axes définis par une lettre (par exemple, T pour topographie, E pour étiologie). À l'intérieur de chaque axe, les éléments sont organisés suivant une structure hiérarchique. La classification d'un terme repose sur une décomposition de celui-ci en combinaison de termes appartenant à différents axes. Ainsi un diagnostic est traduit par plus d'un élément signifiant, mais chaque axe ne doit pas être obligatoirement validé. Par exemple, la juxtaposition : *T2856 (lobe supérieur du poumon gauche) / M4100 (inflammation) / F0300 (fièvre) / E2012 (pneumocoque)* correspond à la phrase *« Pneumonie fébrile à pneumocoque du lobe supérieur gauche »*. L'ajout de connecteurs concernant notamment les liens de causalité permet de décrire un fait complexe en plusieurs phrases. La SNOMED est une des classifications médicales les plus complètes mais un même concept peut y être décrit de différentes façons et rien n'empêche de créer par combinaison des concepts inconsistants (le Bozec, 2001). Ce modèle pose encore des problèmes, par exemple : les termes des différents axes ne sont pas complètement indépendants entre eux, l'axe Maladie fait souvent double emploi, certains concepts peuvent apparaître dans plusieurs axes.

En raison de ce type de défaut, la SNOMED a évolué en SNOMED-RT (pour *Root Procedure*) puis en SNOMED-CT (pour *Clinical terms*), fusion de la SNOMED-RT et d'une terminologie britannique *Clinical Terms* de la NHS (services de santé britanniques). Dans cette nouvelle configuration, la SNOMED-RT respecte un certain nombre de principes : structure hiérarchique de concepts, définitions de types ou de rôles pour des concepts, consistance, exploitation dans le cadre d'une logique de description (Dolin *et al.*, 2001) qui en font une ontologie formelle. La principale difficulté qui semble apparaître dans la transformation de la SNOMED en une ontologie est le choix qui a été fait de conserver les termes de la classification comme concepts et labels des concepts de l'ontologie (Spackman *et al.*, 2002) : la volonté de transformer la SNOMED en une réelle ontologie formelle se heurte à la « nécessité » que s'imposent les auteurs, de conserver la classification – presque – telle quelle avec les souplesses et les imprécisions d'un paradigme de construction pour partie linguistique. En l'état, cette transformation n'est pas totalement assumée, en particulier au niveau de la normalisation.

La SNOMED-CT se veut une terminologie des soins de santé cliniques, dynamique et valide scientifiquement. Son objectif est de rendre les connaissances de soins de santé plus accessibles pour l'ensemble des spécialités médicales. La terminologie Core SNOMED-CT contient plus de 361 800 concepts de soins et comporte également 975 000 descriptions et près de 1,47 million de relations sémantiques.

Pour des informations plus complètes sur le MeSH, la SNOMED et CISMeF dans un cadre de recherche d'informations, nous conseillons la lecture de la thèse de doctorat d'A. Névéol, (2005).

[19] http://www.snomed.org/

2.6 UMLS

UMLS a été mis en place dans le but d'améliorer l'accès à l'information médicale à partir de sources diverses : bases de données bibliographiques, bases de données d'enregistrements cliniques et bases de connaissances médicales (Lindberg & Humphreys, 1990). Un des moyens d'UMLS est alors de définir un vocabulaire médical de base, un « métathésaurus » qui reprend et dédoublonne les termes de l'ensemble des 95 ressources terminologiques qu'il inclut (MeSH, SNOMED ...)[20]. Ce métathésaurus propose une description hiérarchique des connaissances médicales utilisées dans divers documents et systèmes à base de connaissances. De plus, un réseau sémantique de 134 types sémantiques environ permet de typer tous les termes du métathésaurus. L'intérêt d'UMLS réside dans sa grande couverture du domaine médical (1 276 301 concepts dans la version A du 1^{er} trimestre 2006) et dans sa disponibilité. Le métathésaurus n'est pas une ontologie : il n'a pas été fait dans ce but et une tentative de réutilisation comme une ontologie (l'utilisation du métathésaurus pour construire l'ontologie de MENELAS) s'est soldée par un échec (Charlet *et al.*, 1996). Le réseau sémantique a une structure beaucoup plus proche d'une ontologie mais n'en a pas la précision : il recense 134 types plus ou moins généraux en médecine qui servent à typer chaque terme du métathésaurus.

Pour le traitement du vocabulaire médicale en anglais, l'UMLS constitue un outil informatique puissant qui permet d'accéder à de nombreuses informations puisqu'il gère les variations des termes et les relations qu'ils entretiennent entre eux. Cependant, ce métathésaurus concerne essentiellement des terminologies anglophones. S'appuyant sur la dynamique et les acquis de l'UMLS, le projet VUMEF[21] (Darmoni *et al.*, 2003) a pour objectif d'augmenter la part du français dans l'UMLS, afin de consolider les ressources terminologiques francophones du domaine médical. Ce projet a pour intérêts 1) l'amélioration des traductions existantes pour certaines terminologies, en particulier le MeSH ; 2) la traduction de nouvelles terminologie, en particulier la SNOMED ; 3) l'intégration de vocabulaires spécifiquement français comme le Catalogue Commun des Actes Médicaux. Par ailleurs, le consortium VUMEF a également pour objectif de fournir des outils et des méthodes permettant de mettre en correspondance des expressions libres et des vocabulaires contrôlés, et d'évaluer ces tâches. Le projet prévoit, par la suite, de mettre en œuvre les ressources développées pour l'aide au codage dans les dossiers patients et l'indexation automatique de sites web.

Mené en parallèle, le projet UMLF[22] (Zweigenbaum *et al.*, 2003) se donne pour tâche d'effectuer la collecte, la synthèse, la complétion et la validation de ressources lexicales pour le français médical. Par une approche monolingue, il vise à produire un lexique contenant les variantes flexionnelles et dérivationnelles des mots du domaine. Ces informations doivent être encodées dans un format informatique standard afin de favoriser leur intégration dans des systèmes de Traitement automatique de la langue médicale.

[20] http://www.nlm.nih.gov/research/umls/
[21] http://www.vidal.fr/vumef/
[22] http://www-test.biomath.jussieu.fr/umlf/

2.7 FMA

Le Foundational Model of Anatomy[23] (FMA) est une ontologie de référence pour le domaine de l'anatomie. C'est une représentation de toutes les entités anatomiques et les relations nécessaires pour la modélisation symbolique de la structure phénotypique du corps humain dans une forme qui soit compréhensible par l'homme et qui soit également traitable par une machine (Rosse & Mejino, 2003). Cette ontologie permet de décrire les localisations des tumeurs ou des métastases, ce qui permet par exemple de déduire qu'une tumeur principale située dans le lobe supérieur du poumon gauche est également située dans le poumon gauche (car le lobe supérieur est une partie du poumon). Le FMA est mis à la disposition d'utilisateurs qui peuvent récupérer des parties de la modélisation pour les intégrer dans leur propre ontologie. Cette ontologie suscite de réels espoirs dans les communautés de l'Ingénierie des connaissances et de l'Informatique médicale.

2.8 DOLCE

L'ontologie DOLCE[24] (Descriptive Ontology for Linguistic and Cognitive Engineering), élaborée par l'équipe de Nicola Guarino (LOA, Trento, Italie), constitue un des résultats du projet européen WonderWeb Fundation Ontologies Library. DOLCE est une ontologie de haut niveau dont la vocation est d'être utilisée pour concevoir des ontologies de domaine. Sa structure repose sur la distinction philosophique entre entités perdurantes et endurantes[25].

2.9 GALEN

Le projet GALEN [26], développé à l'université de Manchester, vise à mettre en place un serveur de terminologie en médecine. Développé depuis 1992 au sein de projets européens successifs, il est centré sur un *common reference model*, une ontologie de la médecine telle que nous l'entendons ici (*cf.* section 1.6). Cette ontologie respecte une structure arborescente au niveau de ses types primitifs et est le cœur du système et des services qu'il propose (Rector, 1998). GALEN utilise un formalisme appelé GRAIL (Galen Representation and Integration Language) qui permet de saisir la connaissance terminologique dans le domaine médical (Rector *et al.*, 1992). Ce formalisme est hautement génératif et permet de définir des concepts complexes (ou définis), composés de concepts plus élémentaires (ou primitifs). Tous les concepts, et les relations qui les lient, sont représentés indépendamment du langage dans lequel ils sont exprimés.

Nous n'allons pas décrire le système par le menu[27] et nous allons plutôt nous intéresser au *common reference model* et à sa mise en œuvre. L'ontologie de la médecine ne couvre que les domaines dans lesquels le projet s'est développé, où des opportunités se sont créées. Ainsi, le

[23] `http://sig.biostr.washington.edu/projects/fm/index.html`

[24] `http://www.loa-cnr.it/DOLCE.html`

[25] Pour plus d'informations sur la distinction entre les notions d'entités perdurantes et endurantes dans les top-ontologies, le lecteur peut se reporter à l'article de I. Johansson (2005).

[26] `http://www.opengalen.org/`

[27] Nous renvoyons le lecteur intéressé à, par exemple, (Rogers *et al.*, 2001).

département de santé publique et d'informatique médicale de Saint-Étienne participe au développement de la classification commune des actes médicaux (CCAM) et justifie l'utilisation d'un tel outil pour le développement cohérent d'une terminologie médicale (Rodrigues *et al.*, 1998, 1999). Les promoteurs du projet sont confrontés au problème de l'évolution rapide de la médecine et donc des ontologies attenantes[28], au point qu'il est difficile d'avoir une vue complète et cohérente d'une ontologie d'un domaine précis. Confronté à la difficulté de compréhension des ontologies, le projet GALEN y a répondu, même si ça n'était pas au début le but d'un tel module, par l'utilisation d'un générateur de langage naturel permettant de valider les représentations proposées avec les praticiens. Enfin, les promoteurs du projet, s'ils construisent des ontologies, ne proposent pas réellement de méthode argumentée et constructive.

2.10 MENELAS

MENELAS est un projet européen piloté de 1992 à 1995 par le DIAM/SIM/DSI/AP-HP. Le but du projet MENELAS était la conception et l'implémentation d'un système pilote capable d'accéder à des rapports médicaux rédigés en langage naturel dans trois langues : l'anglais, le français et le néerlandais. Ce système devait pouvoir analyser le contenu de rapports médicaux (comptes rendus d'hospitalisation ou CRH) et l'archiver dans une base de données sous la forme d'un ensemble de structures conceptuelles (graphes conceptuels de J. Sowa (1984)). Ces structures, qui constituent la représentation de chaque CRH, devaient pouvoir ensuite être consultées pour accéder à des informations spécifiques contenues dans le CRH. Une partie des informations était encodée à l'aide de nomenclatures internationales, ce qui permettait leur échange à partir de CRH écrits en différentes langues (Zweigenbaum *et al.*, 1995). Le projet a été confronté aux problèmes habituels de la compréhension de textes en langage naturel. Il s'agit bien sûr des problèmes inhérents au langage, comme la paraphrase, l'ambiguïté, la métonymie et de façon plus générale la description et la mise en œuvre de connaissances syntaxiques et sémantiques adéquates et d'une couverture suffisante. Il s'agit aussi des problèmes généraux de représentation du sens des énoncés et des connaissances à fournir au système. On rejoint alors des problématiques classiques en Intelligence artificielle : acquisition, représentation, mise en œuvre, validation de connaissances complexes.

MENELAS repose sur l'hypothèse que la compréhension d'un CRH consiste à construire une représentation conceptuelle de la situation du monde décrit dans le texte. Cette hypothèse peut être justifiée par le fait que nous nous intéressons à des rapports techniques qui décrivent ce qui est arrivé au patient durant son hospitalisation. Le sous-système d'analyse du langage naturel inclus dans MENELAS utilise un analyseur morpho-syntaxique, un analyseur sémantique et un analyseur « pragmatique ». L'analyseur sémantique produit une représentation du sens sous forme de graphes conceptuels (Sowa, 1984). Cette représentation correspond au sens littéral des phrases ; elle est construite à partir d'une phrase en associant des concepts à des mots grâce à un lexique sémantique : on passe « du mot au concept ». La compréhension d'un texte repose sur l'utilisation de connaissances médicales et de connaissances de sens commun qui permettent d'inférer de nombreuses informations implicites. Ces informations correspondent à des infé-

[28] Nous avons également rencontré ce problème, *cf.* chapitre 6.

rences effectuées naturellement par un spécialiste du domaine lorsqu'il lit un CRH. L'analyseur pragmatique a pour tâche d'obtenir un niveau de compréhension plus profond en construisant un modèle de la situation décrite : il va « du concept au concept » (Zweigenbaum *et al.*, 1995).

La question de la construction de l'ontologie[29] de MENELAS a été abordée de façon approfondie dans (Bouaud *et al.*, 1994; Charlet *et al.*, 1996). Elle a amené la mise au point de la méthodologie ARCHONTE (*cf.* section 1), principalement par B. Bachimont, et l'implémentation d'un système opérationnel, principalement par J. Bouaud (1992), se servant de cette ontologie. Finalement, l'ontologie construite comporte plus de 1 800 types et 300 relations.

2.11 Synthèse sur les RTO en médecine

À la fin de ce panorama, on peut constater que peu parmi les produits terminologiques étudiés, en particulier en médecine, sont des ontologies. Ce n'est pas étonnant, les besoins de la médecine ayant d'abord été focalisés sur des problèmes de terminologie médicale, les produits construits avaient les caractéristiques correspondantes. Les ontologies n'étant pas sans rapport avec les terminologies, on peut trouver dans ces thésaurus, en particulier UMLS et pourquoi pas le MESH, des ressources pour amorcer une ontologie. Cela dit, il convient de faire très attention à ne pas confondre les objectifs auxquels ces différents types de ressources peuvent apporter une réponse. Après avoir situé les différents types de RTO les uns par rapport aux autres (*cf.* section 1) et avoir présenté un panorama des RTO disponibles en médecine et appartenant à notre domaine de recherche, nous allons maintenant nous intéresser de plus près à la question de leur élaboration.

3 Formalismes pour la représentation des connaissances

Représenter des connaissances propres à un domaine consiste à décrire et à coder les éléments de ce domaine pour qu'une machine puisse les traiter, raisonner et résoudre des problèmes particuliers (Kayser, 1997). Il faut donc savoir : (1) exprimer ces connaissances à l'aide d'un langage formel de description des connaissances et (2) les manipuler, c'est-à-dire être capable de faire un certain nombre d'opérations dessus (modifier, compléter, déduire de nouvelles connaissances ...) à l'aide de mécanismes définis opérant sur les différents éléments de la représentation.

Il existe un certain nombre de formalismes dans le domaine de la représentation des connaissances ; nous ne présentons ici que ceux qui nous semblent les plus intéressants pour nos travaux : les graphes conceptuels et les logiques de description. L'un comme l'autre permettent de représenter des ontologies, des propriétés, et de mettre en œuvre des mécanismes d'inférence : inférences propres aux structures de graphes comme la jointure ou la projection pour les graphes conceptuels, classifications dans des structures arborescentes pour les logiques de description.

> *« In a logic-based approach, the representation language is usually a variant of the first-order predicate calculus, and reasoning amounts to verifying logical consequence. In the non-logical approaches, often based on the use of graphical interfaces, knowledge is represented by the means of some ad hoc data structure, and*

[29] http://estime.spim.jussieu.fr/Menelas/Ontologie/html/

> *reasoning is accomplished by similarly ad hoc procedures that manipulate the structure*[30] *» (Baader* et al.*, 2003b).*

Chacun de ces formalismes de représentation est implémenté dans un ou plusieurs langages, en particulier adaptés au web et utilisant généralement la syntaxe XML. Nous aborderons ces langages dans la section 5.

3.1 Graphes conceptuels

Les graphes conceptuels sont introduits par J. Sowa (1984) et appartiennent à la famille des réseaux sémantiques. Ces réseaux modélisent les connaissances sous forme de graphes orientés et étiquetés (ou, plus précisément, de multi-graphes, car rien n'exclut que deux nœuds du graphe soient reliés par plusieurs arcs), dans lesquels les nœuds sont associés à des concepts et les arêtes à des relations. Le système de représentation proposé par J. Sowa allie une certaine souplesse, par laquelle il se rapproche de l'efficacité descriptive du langage naturel, et de la rigueur, grâce à laquelle il est possible de mettre en œuvre des procédures inférentielles.

> *« Conceptual graphs (CGs) are a system of logic based on the existential graphs of Charles Sanders Peirce and the semantic networks of artificial intelligence. They express meaning in a form that is logically precise, humanly readable, and computationally tractable. [. . .] conceptual graphs serve as an intermediate language for translating computer-oriented formalisms to and from natural languages. With their graphic representation, they serve as a readable, but formal design and specification language*[31] *» (J. Sowa*[32]*).*

Comme nous le verrons dans le chapitre 5, section 1, à propos de notre application, le modèle des graphes conceptuels se prête bien à des présentations graphiques des connaissances. La brève présentation que nous faisons ici est inspirée du livre de D. Kayser (1997) qui s'inspire lui-même de l'exposé très complet de M. Chein et M.-L. Mugnier (1992).

Le modèle des graphes conceptuels se décompose en deux parties :

1. Une partie terminologique dédiée au vocabulaire conceptuel des connaissances à représenter, c'est-à-dire les types de concepts, les types de relations et les instances des types de concepts. Il faut noter que le terme *concept* est utilisé dans ce formalisme pour désigner

[30] *« Dans une approche orientée "logique", le langage de représentation est souvent une variante de la logique du premier ordre et raisonner revient à contrôler la conséquence logique. Dans les approches non-logiques, souvent fondées sur l'utilisation d'interfaces graphiques, les connaissances sont représentées au moyen de structures ad hoc et le raisonnement se fait par des processus également ad hoc manipulant ces structures. »*

[31] *« Les graphes conceptuels (CGs) forment un système logique fondé sur les graphes existentiels de Charles Sanders Peirce et sur les réseaux sémantiques de l'intelligence artificielle. Ils expriment le sens sous une forme précise d'un point de vue logique, lisible par un humain et manipulable par une machine. [. . .] les graphes conceptuels servent de langage intermédiaire pour traduire des formalismes liés à l'utilisation d'un système informatique d'après les langues naturelles. Avec leur représentation graphique, ils servent de langage de modélisation et de spécification à la fois lisible et formel. »*

[32] Texte extrait du site web de J. Sowa dédié aux graphes conceptuels, `http://www.jfsowa.com/cg/`.

les instances des types de concepts, qui correspondent à la notion classique de concept. Cette partie concernant la terminologie correspond à la représentation du modèle conceptuel mais intègre également des connaissances sur la hiérarchisation des types de concepts et de relations.

Le niveau terminologique du modèle comprend trois ensembles disjoints : l'ensemble des types de concepts (noté $\mathsf{T_c}$), l'ensemble des types de relations (noté $\mathsf{T_r}$) et l'ensemble des marqueurs individuels (noté M).

- Un type de concepts rassemble les caractéristiques communes à plusieurs concepts. Les concepts sont des instances de leur type de concepts. Par exemple, ce qui peut être vu comme le concept *« Patient »* dans une ontologie sera représenté par un type de concepts *« Patient »* dans le modèle des graphes conceptuels, et la patiente désignée sous le nom de *Madame X* sera une instance du type de concepts *« Patient »*. Une relation de spécialisation peut être définie sur l'ensemble des types de concepts. Ainsi, le type de concepts *« Prélèvement »* peut être vu comme un sous-type du concept *« InvestigationBiologique »*. Il existe donc une hiérarchie $\mathsf{T_c}$ de types de concepts représentée par une relation de spécificité organisée en treillis[33].
- Les types de relations représentent les relations qui peuvent exister entre les différents concepts, c'est pourquoi à chaque type de relations est associé une signature qui spécifie les types de concepts définis dans l'ensemble des types de concepts existants. Formellement, il faut définir, pour chaque type de relations, la signature qui lui est associée, c'est-à-dire le n-uplet des types de concepts les plus généraux pouvant être liés par ce type de relations. L'arité de la signature est également l'arité du type de relations. Par exemple, le type de relations *« diagnostiquer »* est d'arité 2 et sa signature est *(PersonnelMedical ;Pathologie)*. Les types de relations sont également hiérarchisés par une relation de spécialisation.
- L'ensemble des marqueurs individuels permet d'identifier les concepts, c'est-à-dire les instances des types de concepts. Par exemple, la patiente Martine Dupont sera identifiée par le marqueur *« Martine Dupont »* associé au type de concepts *« Patient »*. Chaque marqueur est associé au type le plus spécifique de l'instance qu'il désigne.

Ces trois ensembles constituent ce qu'on appelle le support qui va régir l'ensemble des graphes conceptuels portant sur un même domaine de la connaissance. Un support se définit ainsi : $\mathsf{S} = (\mathsf{T_c}; \mathsf{T_r}; \mathsf{M})$, les 3 ensembles $\mathsf{T_c}$, $\mathsf{T_r}$ et M étant disjoints et les ensembles $\mathsf{T_c}$ et $\mathsf{T_r}$ étant partiellement ordonnés. Chaque nœud de relation est étiqueté par un élément de l'ensemble $\mathsf{T_r}$. Chaque nœud de concept reçoit donc un double étiquetage : un élément de l'ensemble $\mathsf{T_c}$ et un marqueur pris dans l'ensemble M.

2. Une partie assertionnelle dédiée à la représentation des assertions du domaine de connaissances étudié. Les faits, représentés au niveau assertionnel, utilisent le vocabulaire décrit par le support S et sont présentés sous la forme de graphes particuliers : les graphes conceptuels.

[33] Un treillis est défini comme étant le graphe d'une relation R sur un ensemble T, telle que pour toute paire t' et t'' d'éléments de T, il existe deux éléments de T notés respectivement $t' \wedge t''$ et $t' \vee t''$, tels que $(t' \wedge t'')\mathsf{R}t'$, $(t' \wedge t'')\mathsf{R}t''$, $t'\mathsf{R}(t' \vee t'')$, $t''\mathsf{R}(t' \vee t'')$.

Un graphe conceptuel est défini comme étant un multi-graphe fini, non-orienté et biparti[34], composé de sommets « concepts » et « relations ».

– Chaque sommet concept est composé d'un type de concepts et d'un marqueur individuel. Un nouveau marqueur, le marqueur générique[35] (noté $\star$) est ajouté à l'ensemble des marqueurs M.
– Chaque sommet relation a un type de relations. Un graphe conceptuel est forcément défini sur un support S donné et les arêtes du graphe ne peuvent relier qu'un sommet concept à un sommet relation.

La figure 3.9 exprime la phrase *« La patiente Martine Dupont est atteinte de diabète et elle est soignée par un médecin de l'hôpital. »* dans le formalisme des graphes conceptuels. Par convention, les rectangles symbolisent les concepts et les ovoïdes, les relations. Les chiffres indiqués sur la figure représentent l'arité de la relation.

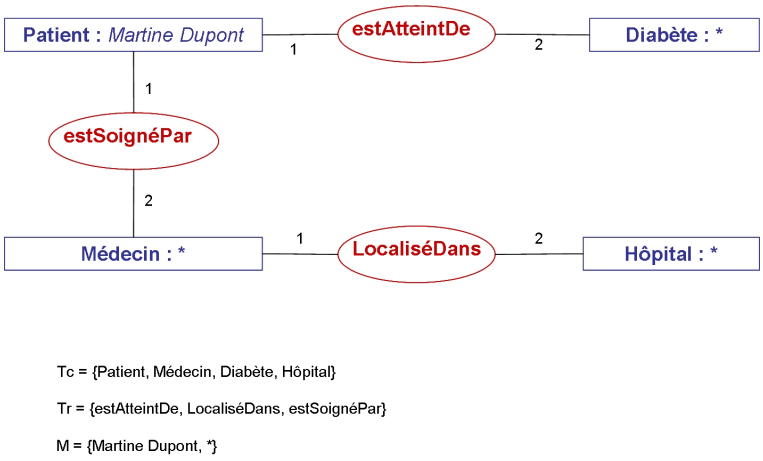


FIG. 3.9 – Exemple de graphe conceptuel.

Le processus de raisonnement lié aux graphes conceptuels s'appelle la projection[36]. Il s'agit d'une opération qui permet de déterminer si un graphe est plus spécialisé ou plus général qu'un autre. M. Chein et M.-L. Mugnier (1992) démontrent qu'il existe une projection d'un graphe conceptuel G dans un graphe conceptuel H si et seulement si H est une spécialisation de G tel que $(H \leq G)$. La figure 3.10 représente un exemple simplifié de cette opération de projection.

[34]Un graphe biparti a ses sommets répartis en deux groupes et chacune de ses arêtes lie un sommet de chacun des deux groupes. J. Sowa propose un certain nombre d'exemples de graphes conceptuels à l'adresse web suivante : `http://www.jfsowa.com/cg/cgexampw.htm`.

[35]Le marqueur est dit générique lorsqu'on connaît son type mais pas son identifiant. Par exemple, le sommet concept « Hôpital : $\star$ » dans la figure 3.9.

[36]Le lecteur intéressé peut notamment se reporter aux travaux détaillés de M. Chein et M.-L. Mugnier (1992).

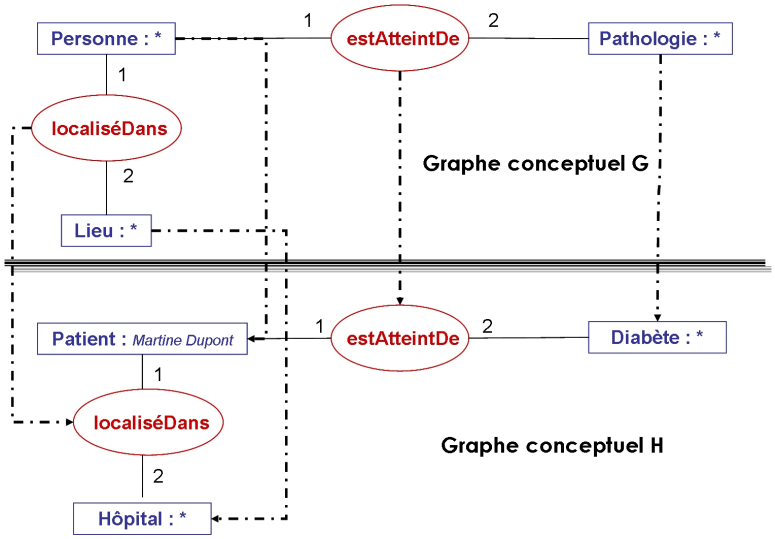


FIG. 3.10 – Exemple de projection du graphe conceptuel G dans le graphe conceptuel H, ($H \leq G$).

J. Sowa définit un opérateur, noté ϕ, qui transforme chaque élément du formalisme des graphes conceptuels en un élément équivalent de la logique des prédicats du premier ordre. Nous ne détaillerons pas ici ce processus. Simplement, tout graphe conceptuel bien formé G peut être transformé en une formule bien formée $\phi(G)$ appartenant à la logique des prédicats par l'application de règles. Par exemple, le graphe de la figure 3.9 est équivalent à la formule suivante :
$\exists x \exists y \exists z(\mathsf{Patient}(\mathsf{MartineDupont}) \wedge \mathsf{Diabete}(x) \wedge \mathsf{Medecin}(y) \wedge \mathsf{Hopital}(z) \wedge$
$\mathsf{estAtteintDe}(\mathsf{MartineDupont}, x) \wedge \mathsf{estSoignePar}(\mathsf{MartineDupont}, y) \wedge \mathsf{localiseDans}(y, z))$.

La correspondance entre les opérations de généralisation sur les graphes conceptuels et l'inférence logique montre qu'étant donné deux graphes G et H, si $G \leq H$, alors $\phi(S), \phi(G) \vdash \phi(H)$. En outre, si G et H sont deux graphes conceptuels bien formés et sous forme normale, c'est-à-dire que deux sommets concepts n'ont jamais le même marqueur individuel, et si $\phi(S), \phi(G) \vdash \phi(H)$, alors $G \leq H$. Ainsi, le formalisme des graphes conceptuels est doté d'une sémantique logique complète. Il est également doté d'une sémantique ensembliste détaillée par M.-L. Mugnier et M. Chein (1996)

D. Kayser (1997) reprend une critique courante de la représentation des connaissances à l'aide de graphes. Les graphes conceptuels n'expliquent pas avec assez de précision ce que représentent véritablement leurs nœuds et leurs arcs. Il paraît facile, en effet, de tracer quelques ronds et rectangles reliés et d'affirmer que cela représente les connaissances d'un certain domaine. Dans les cas simples, la signification de ces figures semble claire et l'on peut en donner une traduction logique. Mais, ce à quoi l'on parvient ainsi est, d'après D. Kayser, relativement rudimentaire. À partir de ces considérations, il est tentant d'introduire des connaissances plus précises et raffinées, en ajoutant de nouvelles conventions d'expression. Cependant, il paraît évident que les procédures inférentielles se complexifient à mesure qu'on augmente le niveau d'expressivité des représentations. De plus, il faut également gérer les problèmes de décidabilité

et de temps de calcul qui peuvent rapidement devenir prohibitifs.

3.2 Logiques de description

Les logiques de description[37] sont des langages formels permettant de représenter des propriétés pour des ensembles d'objets et exploitent, en général, des sous-ensembles décidables[38] de la logique du premier ordre. Elles sont nées du besoin de pallier le manque de sémantique formelle des réseaux sémantiques et des systèmes à base de frames et ont été largement étudiées et utilisées dans plusieurs systèmes à base de connaissances (Nardi & Brachman, 2003).

Ces logiques de description sont également une réponse à la neutralité ontologique de la logique du premier ordre et à son indécidabilité, comme le souligne O. Dameron (2003). En effet, les logiques de description utilisent une approche ontologique, c'est-à-dire que pour décrire les individus d'un domaine, elles requièrent tout d'abord la définition des catégories générales d'individus et les relations logiques que les individus ou catégories peuvent entretenir entre eux. Cette approche ontologique est naturelle pour le raisonnement puisque même si la majorité des interactions se déroulent au niveau des individus, la plus grande partie du raisonnement se produit au niveau des catégories (Russell & Norvig, 2002).

Le développement des logiques de description fut fortement influencé par les travaux sur la logique des prédicats, les schémas (*frames* en anglais) (Minsky, 1981) et les réseaux sémantiques. Des correspondances existent entre ces logiques et ces formalismes (Sattler *et al.*, 2003; Baader & Nutt, 2003). Comme l'explique D. Kayser (1997), la présence de catégories générales d'objets et de relations fait d'ailleurs partie de l'héritage conceptuel des schémas et des réseaux sémantiques.

Les premiers travaux sur les logiques de description commencent au début des années 1980 avec des systèmes à base de connaissances tels que KL-ONE, BACK et LOOM (Baader *et al.*, 2003a; Nardi & Brachman, 2003). Ces premières implantations résolvent des problèmes d'inférence en temps souvent polynomial, par le biais d'une catégorie d'algorithmes de vérification de subsomption de type normalisation/comparaison (*structural subsumption algorithms*). Ces algorithmes ne s'appliquent qu'à des logiques de description peu expressives, sans quoi ils sont incomplets, c'est-à-dire qu'ils sont incapables de prouver certaines formules vraies. Dans les années 1990, une nouvelle classe d'algorithmes apparaît : les algorithmes de vérication de satisfiabilité à base de tableaux (*tableau-based algorithms*). Ces algorithmes raisonnent sur des logiques de description dites expressives ou très expressives, mais en temps exponentiel. Cependant, en pratique, le comportement des algorithmes est souvent acceptable (Baader *et al.*, 2003a). L'expressivité accrue a ouvert la porte à de nouvelles applications telles que le Web sémantique (Horrocks *et al.*, 2003). Le terme « logiques de description expressives » désigne l'ensemble des logiques de description qui ont émergé pendant cette période.

Il existe plusieurs types de logiques de description, parmi lesquelles : FL (Brachman & Levesque, 1984), PL_1 et PL_2 (Donini *et al.*, 1991), ... Nous ne souhaitons pas consacrer cette

[37] Le lecteur intéressé peut consulter la page web suivante concernant les logiques de description : `http://www.ida.liu.se/labs/iislab/people/patla/DL/index.html`

[38] Pour une logique, un problème de raisonnement est décidable si une machine de Turing peut le résoudre en un nombre fini d'étapes.

section à un type de logique particulier, par conséquent, nous adopterons ici un point de vue global.

La modélisation des connaissances d'un domaine à l'aide des logiques de description comporte deux niveaux, la T-box et la A-box (*cf.* figure 3.11), et combine des représentations intentionnelles et extensionnelles des connaissances (Baader *et al.*, 2003b) :

- la T-box (T pour terminologique) décrit les connaissances générales d'un domaine et contient les déclarations des primitives conceptuelles organisées en concepts et relations. Ces déclarations décrivent les propriétés des concepts et des relations et constituent donc une définition intentionnelle des connaissances ;
- une A-box (A pour assertionnel) décrit les connaissances factuelles d'un domaine et représente une configuration précise. Elle contient les déclarations d'individus, instances des concepts qui ont été définis dans la T-box. Plusieurs A-box peuvent être associées à une même T-box ; chacune représente une configuration constituée d'individus, et utilise les concepts et rôles de la T-box pour l'exprimer.

TBox	*ABox*
$Femelle \sqsubseteq \top \sqcap \neg Mâle$	$Humain(Anne)$
$Mâle \sqsubseteq \top \sqcap \neg Femelle$	$Femelle(Anne)$
$Animal \equiv Mâle \sqcup Femelle$	$Femme(Sophie)$
$Humain \sqsubseteq Animal$	$Humain(Robert)$
$Femme \equiv Humain \sqcap Femelle$	$\neg Femelle(Robert)$
$Homme \equiv Humain \sqcap \neg Femelle$	$Homme(David)$
$Mère \equiv Femme \sqcap \exists relationParentEnfant$	$relationParentEnfant(Sophie, Anne)$
$Père \equiv Homme \sqcap \exists relationParentEnfant$	$relationParentEnfant(Robert, David)$
$MèreSansFille \equiv Mère \sqcap \forall relationParentEnfant.\neg Femme$	
$relationParentEnfant \sqsubseteq \top_R$	

FIG. 3.11 – Base de connaissances composée d'une T-box et d'une A-box, (Fournier-Viger, 2005).

Les concepts et les rôles (c'est-à-dire des relations entre concepts) atomiques constituent les entités élémentaires d'une T-box. Le côté gauche de la figure 3.11 présente un exemple de T-box dans laquelle les noms commençant par une lettre majuscule, comme Humain, Animal, Femelle ou Male, désignent des concepts et ceux commençant par une minuscule, comme relationParentEnfant, désignent des rôles.

Les concepts et rôles atomiques peuvent être combinés au moyen de constructeurs pour former des concepts et des rôles composés. La figure 3.12 montre ces constructeurs et leur traduction en logique du premier ordre. Par exemple, le concept composé Male $\sqcap$ Femelle est le résultat de l'utilisation du constructeur $\sqcap$ sur les concepts atomiques Male et Femelle. L'interprétation du concept ainsi composé est « l'ensemble des individus qui appartiennent à la fois au concept Male et au concept Femelle ».

Une T-box contient des axiomes de la forme $C \sqsubseteq D$ ou bien $C \equiv D$, sachant que C et D sont des concepts composés. Le premier axiome exprime une relation d'inclusion et le second

Constructor	Syntax	FOL
Concept	C	$C(x)$
Role name	R	$R(x,y)$
Instance	$\{ a_1 \}$	$\top(a_1)$
Bottom	$\bot$	$\forall x \neg(\bot(x))$
Top	$\top$	$\forall x \top(x)$
Concept hierarchy	C $\sqsubseteq$ D	$\forall x\ C(x) \Rightarrow D(x)$
Role hierarchy	R $\sqsubseteq$ S	$\forall x,y\ R(x,y) \Rightarrow S(x,y)$
Intersection	C $\sqcap$ D	$\{x \mid C(x) \wedge D(x)\}$
Union	C $\sqcup$ D	$\{x \mid C(x) \vee D(x)\}$
Universal restriction	$\forall$ R.C	$\{x \mid \forall y\ R(x,y) \Rightarrow C(y)\}$
Existential restriction	$\exists$ R.C	$\{x \mid \exists y\ R(x,y) \wedge C(y)\}$
Negation	$\neg$ C	$\{x \mid \neg C(x)\}$
Enumeration	$\{ a_1 ... a_n \}$	$\forall x\ C(x) \Rightarrow (x = a_1) \vee ... \vee (x = a_n)$
Role inverse	R^-	$\forall x,y\ R(x,y) \Leftrightarrow R^-(x,y)$
Minimum cardinality	($\geq$ n R.C)	$\exists y_1 ... y_n\ \bigwedge_{1 \leq i \leq n}(R(x,y_i) \wedge C(y_i)) \wedge \bigwedge_{1 \leq i \leq n, i<j \leq n}\ y_i \neq y_j$
Maximum cardinality	($\leq$ n R.C)	$\forall y_1 ... y_n (\ \bigwedge_{1<i<n}(R(x,y_i) \wedge C(y_i))) \Rightarrow \bigvee_{1<i<n, i<j<n}\ y_i = y_j$

FIG. 3.12 – Exemple de constructeurs en logique de description et leur traduction en logique du premier ordre (FOL), (Fürst, 2004a).

exprime une relation d'équivalence. Une interprétation $\mathcal{I}$ se compose d'un domaine d'interprétation $\triangle^{\mathcal{I}}$ et d'une fonction d'interprétation $^{\mathcal{I}}$. Le domaine d'interprétation regroupe un ensemble d'individus. La fonction d'interprétation $^{\mathcal{I}}$ assigne à chaque concept atomique A, un ensemble tel que $\mathsf{A}^{\mathcal{I}} \subseteq \triangle^{\mathcal{I}}$, et à chaque rôle atomique R, une relation binaire $\mathsf{R}^{\mathcal{I}} \subseteq \triangle^{\mathcal{I}} \times \triangle^{\mathcal{I}}$. Donc, Une interprétation $\mathcal{I}$ satisfait un axiome tel que $\mathsf{C} \sqsubseteq \mathsf{D}$ si et seulement si $\mathsf{C}^{\mathcal{I}} \subseteq \mathsf{D}^{\mathcal{I}}$. Plus généralement, on peut dire qu'une interprétation satisfait une T-box si et seulement si l'interprétation satisfait tous les axiomes de la T-box .

Une A-box contient un ensemble d'assertions sur les individus qui se composent d'assertions dites d'appartenance et d'assertions de rôle. Comme nous l'avons dit dans les premiers paragraphes de cette section, chaque A-box (*cf.* partie droite de la figure 3.11) doit être associée à une T-box car les assertions dont il est question s'expriment en terme de concepts et de rôles définis dans la T-box. Une A-box désigne des individus par les noms qu'elle leur donne, par exemple Anne, Sophie ou David dans la figure 3.11. Une fonction d'interprétation $^{\mathcal{I}}$ assigne à chacun de ces noms a, un individu désigné par $\mathsf{a}^{\mathcal{I}}$, tel que $\mathsf{a}^{\mathcal{I}} \in \triangle^{\mathcal{I}}$. Chaque assertion d'appartenance d'une A-box, notée $\mathsf{C}(\mathsf{a})$, exprime que pour cette A-box, il existe un individu nommé a, membre du concept C appartenant à la T-box associée. On dit qu'une interprétation satisfait une assertion d'appartenance $\mathsf{C}(\mathsf{a})$, si et seulement si $\mathsf{a}^{\mathcal{I}} \in \mathsf{C}^{\mathcal{I}}$. Une assertion de rôle s'écrit $\mathsf{R}(\mathsf{a},\mathsf{b})$ et indique que, pour une A-box donnée, il existe un individu a en relation avec un individu nommé b par le rôle R qui est défini dans la T-box associée. Une interprétation satisfait cette assertion $\mathsf{R}(\mathsf{a},\mathsf{b})$ si et seulement si $(\mathsf{a}^{\mathcal{I}}, \mathsf{b}^{\mathcal{I}}) \in \mathsf{R}^{\mathcal{I}}$.

L'inférence se fait au niveau terminologique (T-box) ou factuel (A-box). Pour chacun de ces niveaux, F. Baader et W. Nutt (2003) soulignent quatre problèmes d'inférence que les moteurs d'inférence essaient de résoudre. Nous ne détaillerons pas ici cet aspect, le lecteur intéressé peut

se reporter à l'ouvrage très complet cité ci-dessus.

– Au niveau terminologique :
    1. Satisfiabilité
    2. Subsomption
    3. Équivalence
    4. Disjonction
– Au niveau factuel :
    1. Cohérence entre une A-box et une T-box
    2. Vérification d'instance
    3. Vérification de rôle
    4. Problème de récupération

Plusieurs types de logiques de description peuvent être utilisés en fonction des constructeurs nécessaires à la représentation des connaissances souhaitée. Autrement dit, les logiques de description se distinguent par la richesse des constructeurs qu'elles proposent. Le fait d'utiliser tous les constructeurs ou seulement une partie d'entre eux joue sur la décidabilité et la complexité des calculs de raisonnement. Ainsi, plus les logiques de description sont expressives, plus il y a des problèmes d'inférence qui tendent vers l'indécidabilité et des complexités élevées. À contrario, les logiques trop peu expressives ne vont pas pouvoir représenter des domaines de connaissances complexes.

3.3 Synthèse sur les formalismes

Les deux formalismes que nous venons de présenter ont un certain nombre de points communs. Tout d'abord, ils sont tous deux dotés d'une sémantique formelle qui permet de raisonner sur les représentations produites. Ensuite, ils séparent également nettement les connaissances de type ontologique (Support S pour les formalismes à base de graphes et T-box pour les logiques de description) et de type assertionnel (graphe et A-box). Ils présentent également quelques différences. Ainsi, les graphes conceptuels décrivent le support avec une expressivité réduite et ont tendance à privilégier la description de connaissances assertionnelles. Les logiques de description sont plus orientées vers la description des connaissances ontologiques. Les ontologies construites suivant le formalisme des logiques de description diffèrent dans leur approche. En effet, plutôt que de créer manuellement une hiérarchie et d'assigner ensuite des propriétés aux concepts, le processus de construction de l'ontologie est inverse. Ainsi, l'utilisateur commence par fournir une définition logique de chaque concept en fonction de laquelle est déduite la classification. Il y a plus d'une manière de classifier un ensemble de concepts et cette approche permet de produire différentes classifications pour différents objectifs nécessitant une même connaissance terminologique fondamentale. Par contre, ce processus ne nécessite pas de la part de l'auteur de l'ontologie une connaissance ou même un aperçu général du domaine modélisé.

Les outils tels que les raisonneurs peuvent être employés pour maintenir des hiérarchies et pour détecter les contradictions logiques dans des descriptions de concept. Les logiques de description fournissent une grammaire sémantique formelle avec laquelle des concepts complexes

peuvent être établis à partir d'éléments plus simples. Les auteurs des ontologies exprimées en logiques de description ou à l'aide de graphes conceptuels doivent fournir les concepts et fixer les contraintes sur la façon dont ils peuvent se composer. L'exigence de donner des définitions formelles aux concepts force les auteurs à être beaucoup plus explicites et précis au sujet de la signification attribuée à chaque concept. Cela facilite l'interprétation des connaissances modélisées par d'autres utilisateurs et également par des machines.

4 Méthodes et méthodologies de construction d'ontologies

Nous avons précisé, en section 1.6, ce que nous entendions par ontologie. Rappelons qu'il s'agit d'un objet informatique qui résulte d'une modélisation des connaissances d'un domaine particulier et qui a pour objectif de répondre à un problème spécifique. Jusqu'à présent, l'élaboration, tout comme la validation, l'évaluation et la maintenance d'ontologies relève le plus souvent d'un savoir-faire que d'une démarche d'ingénierie. Ainsi, les développements d'applications utilisant des ontologies, la visibilité et le travail collaboratif au sein de la communauté d'Ingénierie des connaissances sont retardés.

La plupart des équipes de recherche dans le domaine travaillent de manière *ad hoc*. Bien que la plupart des méthodologies initient le processus de construction par l'identification, puis l'organisation et la structuration des concepts et des relations à représenter, les ontologies réalisées sont très différentes les unes des autres. Faut-il faire l'hypothèse qu'il y ait autant de manières de représenter les connaissances d'un domaine qu'il y a d'ontologues ? Auquel cas, le principal problème de la construction d'ontologies est celui de l'organisation des concepts et des relations en taxinomies.

Nous nous intéressons à la construction d'ontologies *ex nihilo*, ce qui pose trois problèmes : 1) comment fournir un ensemble de termes généraux pour décrire des classes et des relations qui seront employées pour caractériser le domaine lui même ; 2) comment organiser les termes en taxinomies de classes en utilisant la relation « est un » ; enfin, 3) comment exprimer de manière explicite les contraintes qui donnent du sens à la hiérarchie ontologique.

Cette section présente une modeste revue des principales méthodologies de construction d'ontologies et tente d'apporter des éléments de réponse aux problèmes soulevés ci-dessus. Pour dresser cet état de l'art, nous nous sommes appuyés sur les travaux de M. Cristiani et R. Cuel (2005a; 2005b) et de A. Gómez-Pérez *et al.* (2002; 2004a).

4.1 Stratégies descendantes et ascendantes

D'un point de vue méthodologique, il existe, en Ingénierie des connaissances, deux stratégies pour élaborer le modèle conceptuel qui sous-tend la représentation des connaissances d'un domaine : une approche descendante et une approche ascendante. Nous distinguons ci-dessous leur fonctionnement pour le raisonnement et pour l'analyse de textes.

Stratégies pour le raisonnement

Les approches descendantes pour le raisonnement reposent sur des méthodes de modélisation des connaissances telles que CommonKADS (Schreiber *et al.*, 2000) et MKSM (Ermine, 2000) et sont guidées par les méthodes de résolution de problèmes. Ces approches sont censées permettre la construction du schéma du modèle conceptuel en sélectionnant sa description dans une bibliothèque de modèles qui vont être ainsi réutilisés. Le mode de représentation des connaissances du domaine est alors contrait par ce modèle de raisonnement. Ces approches n'ont pas été initialement prévues pour modéliser les objets d'un domaine car elles ont été pensées avant qu'on s'intéresse aux ontologies d'un point de vue informatique. Dans le paradigme de l'Intelligence artificielle de cette période, ces approches s'intéressaient d'abord aux raisonnements et les ontologies n'y ont été modélisées qu'*a posteriori*. Les derniers développements réintègrent la question des connaissances du domaine comme étant un point fondamental de la méthodologie.

Les approches ascendantes pour le raisonnement sont guidées par les données pour abstraire un modèle du domaine et une méthode de résolution de problèmes *ad hoc*. Ces approches permettent d'identifier en premier les concepts les plus spécifiques puis de s'intéresser à des concepts plus abstraits en adoptant un processus de généralisation. MACAO est un exemple d'outil proposant une telle méthode (Aussenac, 1989; Aussenac-Gilles, 2005). Ces stratégies ne nécessitent que des connaissances très locales qui peuvent être connues a priori pour construire des unités d'informations plus grandes.

Stratégies pour l'analyse de textes

Les approches descendantes pour l'analyse de textes sont guidées par les concepts (le modèle) et permettent la modélisation des informations recherchées en utilisant les structures syntaxiques et sémantiques des langues. Les concepts les plus abstraits sont reconnus en premier, puis par un processus de spécialisation, les concepts plus spécifiques sont identifiés. Ces approches sont essentiellement utilisées en acquisition des connaissances.

Les approches ascendantes pour l'analyse de textes sont guidées par les données, c'est-à-dire par l'usage de la langue dans les textes. Elles permettent de faire remonter des informations du corpus, par exemple les syntagmes nominaux désignant des notions du domaine.

Cette dernière question, mobilisée dans nos recherches, sera développée dans la suite du mémoire.

4.2 Les travaux de M. Uschold et M. Grüninger

La méthodologie de M. Uschold et M. Grüninger (1996) propose plusieurs étapes pour construire des ontologies *via* un processus complètement manuel :

1. identifier l'objectif souhaité, spécifier le domaine concerné ;
2. construire l'ontologie et pour cela distinguer, d'une part, les concepts et les relations clés, et, d'autre part, produire des définitions textuelles précises et non ambigües de ces concepts et de ces relations ; coder les connaissances en les intégrant à d'autres ontologies qui préexistent ;

3. évaluer le résultat ;
4. documenter le modèle et éditer des recommandations précises pour chaque étape.

La seconde étape est celle de la construction de l'ontologie. Elle aboutit à une ontologie formalisée. Le modèle de connaissances n'est pas construit directement car les auteurs suggèrent de passer par l'identification d'un ensemble de « questions de compétences » qui constitue une étape intermédiaire de la phase de construction. Ces questions sont une aide pour catégoriser les connaissances de l'ontologie et facilitent l'intégration d'autres ontologies. Le principal inconvénient de la méthodologie proposée est le manque de conceptualisation avant l'implémentation de l'ontologie. L'objectif de la phase de conceptualisation est de donner un modèle du domaine qui serait moins formel que celui issu de la phase d'implémentation, mais plus formel qu'en étant seulement défini en langage naturel. D'autres problèmes sont dus à l'absence d'une phase de conceptualisation. En effet, d'une part, les experts du domaine, les utilisateurs humains, les ontologues ont de grandes difficultés à comprendre les ontologies implémentées dans des langages spécialisés, et, d'autre part, les experts du domaine ne peuvent d'eux-mêmes construire des ontologies modélisant leur propre domaine d'expertise.

Selon M. Uschold et M. Grüninger, les étapes du processus listées ci-dessus ne sont pas suffisantes pour parler de méthodologie. Toute méthodologie devrait inclure un ensemble de techniques, de méthodes et de principes qui préciseraient chacune des quatre étapes et devraient indiquer les relations et les liens entre ces étapes (leur ordre, les *inputs* et les *outputs* ...).

4.3 METHONTOLOGY

METHONTOLOGY (Fernandez *et al.*, 1997; Lopez *et al.*, 1999) est une méthodologie[39] mise au point par l'équipe du LAI de l'Université polytechnique de Madrid. Comme le montre la figure 3.13, cette méthodologie intègre la construction d'ontologies dans un processus de gestion de projet complet, comprenant aussi bien les étapes de spécification des besoins et de planification que celles, par exemple, de la réalisation, de la maintenance et de la documentation. Le processus de construction d'ontologie est composé des dix étapes suivantes :

1. construire le glossaire des termes qui seront inclus dans l'ontologie, préciser leur définition en langage naturel, identifier leurs synonymes et leurs acronymes ;
2. construire des taxinomies de concepts pour les classifier ;
3. construire des diagrammes de relations binaires *ad hoc* pour identifier des relations *ad hoc* entre les concepts d'une même ontologie et également entre les concepts d'ontologies différentes ;
4. construire le dictionnaire de concepts qui inclut, pour chaque concept, ses attributs d'instance, ses attributs de classe et ses relations *ad hoc* ;
5. décrire en détail chaque relation binaire *ad hoc* qui apparaît dans le diagramme de relations binaires *ad hoc* et dans le dictionnaire de concepts ;
6. décrire en détail chaque attribut d'instance qui apparaît dans le dictionnaire de concepts ;

[39] Nous renvoyons le lecteur intéressé à la lecture des travaux de A. Gómez-Pérez *et al.* (2004b).

7. décrire en détail chaque attribut de classe qui apparaît dans le dictionnaire de concepts ;
8. décrire en détail chaque constante (les constantes donnent des informations sur le domaine de connaissances) ;
9. décrire les axiomes formels ;
10. décrire les règles utilisées pour contraindre le contrôle et pour inférer des valeurs aux attributs.

METHONTOLOGY permet de caractériser les ontologies au niveau des connaissances et insiste sur la nécessité de travailler à partir de représentations intermédiaires des connaissances lors de la phase de conceptualisation. METHONTOLOGY est partiellement soutenu par l'environnement WebODE (*cf.* section 6.4).

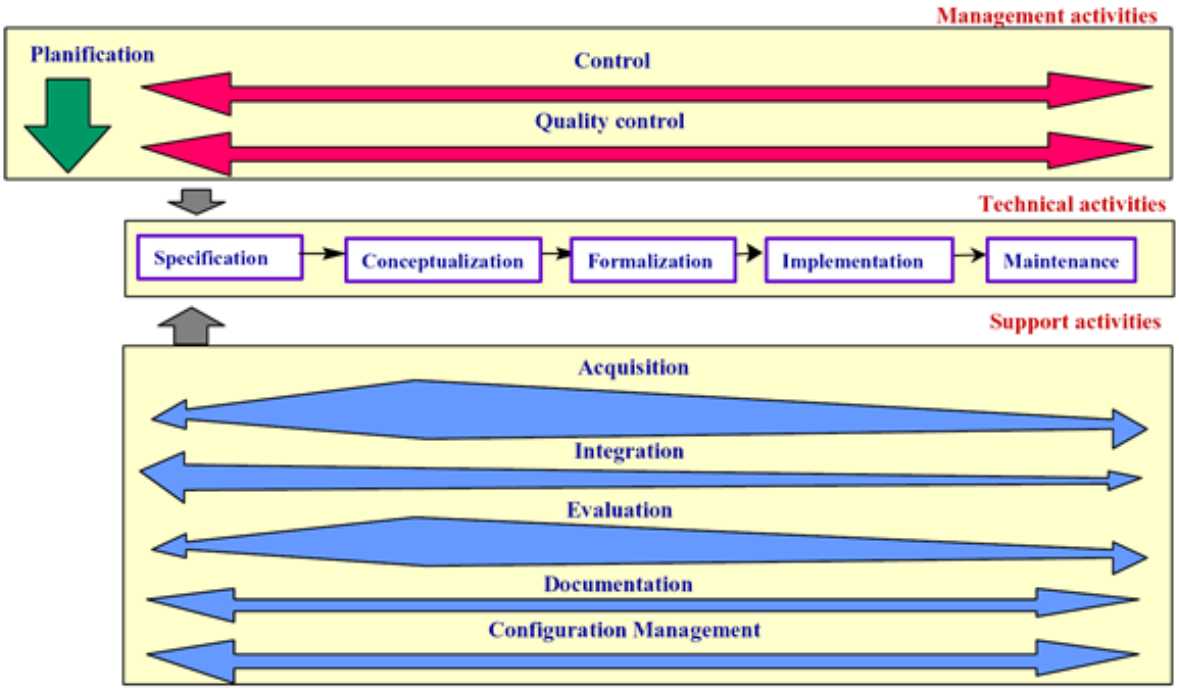


FIG. 3.13 – Processus de développement d'ontologie de METHONTOLOGY, (Corcho *et al.*, 2005).

4.4 Les travaux de N. Guarino et C. Welty

La méthodologie mise au point par N. Guarino et C. Welty (2000b) permet de corriger les hiérarchies taxinomiques construites de manière plus ou moins anarchique. Les auteurs ne proposent pas de guide méthodologique à proprement parler mais une étape de vérification et de correction à intégrer dans le cycle de vie de l'ontologie. Cette méthode repose sur des distinctions

entre concepts (ex : type, rôle, attribution) résultant de la combinaison de différentes propriétés formelles (ex : rigidité, dépendance, identité, ...) et sur des principes de cohérence logique pour les liens de subsomption qui découlent de ces distinctions (Welty & Guarino, 2001). Ces distinctions prennent la forme d'une ontologie de propriétés (Guarino & Welty, 2000a) que l'on peut assimiler à une ontologie « formelle », par opposition à « matérielle » (Bachimont, 2001). Ces propriétés formelles permettent à l'ingénieur des connaissances de vérifier la cohérence et le respect des règles de subsomption. Le module ODEClean fonctionnant avec l'outil WebODE (*cf.* section 6.4) est une implémentation de cette méthodologie.

4.5 OntoSpec

OntoSpec propose une méthode de spécification semi-informelle d'ontologies conceptuelles en langue naturelle fortement structurée et contrôlée. Cette méthode conduit à définir les entités conceptuelles, les concepts et leurs relations, comme des ensembles structurés de propriétés. Chaque propriété est présentée comme une proposition logique et des règles typographiques sont posées. L'objectif d'OntoSpec est de pousser l'ontologue à passer par une phase initiale de modélisation des connaissances. L'ontologie conceptuelle en résultant sert de base à un travail collaboratif entre l'ontologue et l'expert. G. Kassel (2002) note dès lors que l'introduction d'une ontologie formelle spécifiée en langage de représentation, comme KIF ou RDFS, est optionnelle pour le développement d'une ontologie computationnelle. La suite logique de la définition d'une ontologie conceptuelle est la définition d'une ontologie computationnelle qui puisse être utilisée en tant que telle dans un système informatique. OntoSpec s'attache donc à ce que l'élaboration d'une ontologie computationnelle passe par celle d'une ontologie conceptuelle porteuse d'une modélisation précise et rigoureuse des connaissances. Cette ontologie conceptuelle est spécifiée en langue naturelle structurée et contrôlée, ce qui la rend lisible par un non cogniticien. L'objectif d'OntoSpec est donc de prévoir à la fois une modélisation précise des connaissances de l'ontologie, et une lisibilité de l'ontologie conceptuelle qui en résulte.

4.6 ARCHONTE

La méthode ARCHONTE (ARCHitecture for ONTological Elaborating) proposée par B. Bachimont pour construire des ontologies s'appuie sur la sémantique différentielle (Bachimont, 2000; Bachimont *et al.*, 2002). La construction d'une ontologie comporte trois étapes :

1. choisir les termes pertinents du domaine et normaliser leur sens puis justifier la place de chaque concept dans la hiérarchie ontologique en précisant les relations de similarités et de différences que chaque concept entretient avec ses concepts frères et son concept père ;
2. formaliser les connaissances, ce qui implique par exemple d'ajouter des propriétés à des concepts, des axiomes, de contraindre les domaines d'une relation ...
3. l'opérationnalisation dans un langage de représentation des connaissances.

DOE (*cf.* section 6.5) implémente partiellement ARCHONTE. Cette méthodologie est celle sur laquelle nous travaillons, aussi nous proposons un descriptif plus complet au chapitre 4, section 1.

4.7 Conclusion

L'exposé des méthodes et méthodologies ci-dessus permet de distinguer deux grandes phases : 1) une modélisation pour donner du sens, autrement dit, une modélisation des connaissances ontologiques conduisant à la définition d'une ontologie conceptuelle ; 2) une modélisation pour implémenter un système conduisant à une ontologie computationnelle.

D'autres auteurs proposent un certain nombre de critères et d'étapes pertinents pour la construction d'une ontologie dynamique, interopérable, facilement maintenable et indépendante du contexte. Il n'y a pas une seule et unique manière de modéliser un domaine de connaissances, il n'y a que des alternatives plus ou moins réussies. La plupart du temps, le choix d'une méthodologie adéquate dépend des objectifs et des buts poursuivis ainsi que de l'outil de construction utilisé.

5 Langages pour exploiter des ontologies

Une des principales décisions à prendre dans le procédé de développement d'ontologies consiste à choisir le langage (ou l'ensemble de langages) dans lequel l'ontologie sera exprimée et utilisée. L'ingénieur des connaissances a des exigences concernant ces langages : (1) la lisibilité : le langage doit être compréhensible pour un utilisateur humain et doit donc avoir une certaine continuité avec le langage naturel ; (2) la portabilité : le langage choisi doit être le plus standard possible afin de pouvoir être réutilisé dans d'autres systèmes et (3) la possibilité de faire des inférences : le langage doit permettre le traitement informatique des données en vue de calculer les déductions logiques possibles.

Par ailleurs, dans le cadre de ses travaux sur le Web sémantique, le W3C a mis en place en 2002 un groupe de travail dédié au développement de langages standards pour modéliser des ontologies utilisables et échangeables sur le Web (*cf.* figure 3.14). S'inspirant de langages précédents comme DAML+OIL et des fondements théoriques des logiques de description, ce groupe a publié en 2004 une recommandation définissant le langage OWL (Web Ontology Language), fondé sur le standard RDF et en spécifiant une syntaxe XML. Plus expressif que son prédécesseur RDFS, OWL a rapidement pris une place prépondérante dans le paysage des ontologies et est désormais, de facto, le standard le plus utilisé.

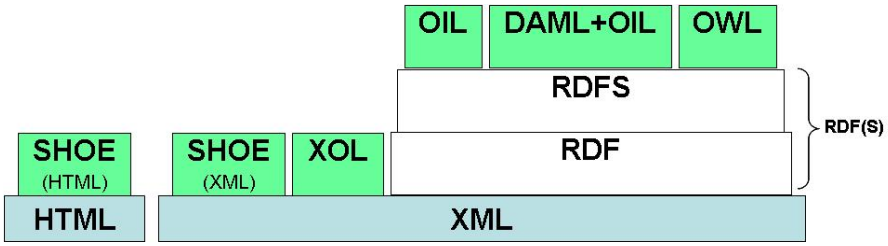


FIG. 3.14 – Ontology markup languages, (Gómez-Pérez, 2004).

5.1 XML

XML (eXtensible Markup Language)[40] est un métalangage dérivé de SGML (Standart General Markup Language) utilisé pour définir des langages de marquage comme XHTML. Ces langages de marquage permettent la structuration des documents du Web non plus sur la base d'une structure figée, comme le permettait HTML, mais en laissant la possibilité au concepteur de distinguer les données selon leur sens et leur contenu. Un document XML se présente sous la forme de données *taggées* par un ensemble de balises, chacune pouvant comporter des attributs et des valeurs. Il n'y a pas de définition figée des balises. Donc, leur grammaire peut être consignée dans une DTD séparée (Document Type Definition). Cette DTD n'a aucune vocation à traiter de sémantique et ne sert qu'à définir des conventions de syntaxe. Les schémas XML, normalisé par le W3C en 2001, permettent de la même façon que les DTD, de modéliser et de valider les documents et leurs données. Ces schémas XML sont cependant beaucoup plus précis que les DTD car ils permettent de donner des contraintes plus strictes aux valeurs.

Une des contradictions fondamentales entre XML et les ontologies réside dans le fait qu'XML ne traduit finalement que des grammaires tandis que les ontologies s'attachent à représenter la sémantique des objets d'un domaine en déterminant les concepts et les relations qui le peuplent. De la même façon et concernant les schémas XML, la logique de classe, nécessaire à la représentation d'une ontologie, ne peut être traduite. Seul un programme effectuant des tâches sur des données XML est donc en mesure d'en extraire une quelconque sémantique, et celle-ci n'est pas formelle. Par conséquent, XML est un bon moyen de décrire des données et de les stocker mais ne permet pas d'interpréter ces mêmes données. Il ne peut donc être utilisé en tant que tel pour décrire une ontologie.

Un langage a pourtant été défini sur la base de XML avec pour principal objectif l'échange d'ontologies. XOL (XML-based Ontology exchange Language)[41] tente de répondre à deux impératifs, d'une part, la nécessité d'avoir un langage de représentation des connaissances orienté objet, et d'autre part, la nécessité d'avoir une syntaxe XML permettant l'interopérabilité et intégrant les capacités des parsers XML. Il faut noter que, l'objet même de XOL n'est pas la représentation d'ontologies mais bien leur échange. XOL peut toutefois être utilisé comme langage intermédiaire lors du transfert d'ontologies dans différents systèmes de bases de données (Corcho & Gómez-Pérez, 2000).

5.2 RDF

RDF (Ressource Description Framework)[42] est un modèle de représentation sémantique des informations du Web qui utilise la syntaxe d'XML. Ces représentations comportent des méta-données sur les ressources du Web comme les auteurs de pages Web, leur date de création ... Les ressources du Web sont l'élément de base de RDF. Chaque ressource est pourvue d'un identifiant uniforme de ressource (*Uniform Resource Identifier, URI*). Initialement recommandé par le W3C dans le but de standardiser les définitions et les usages des méta-données, RDF est éga-

[40] `http://www.w3.org/XML/`
[41] `ftp://smi.stanford.edu/pub/bio-ontology/xol.doc`
[42] `http://www.w3.org/RDF/`

lement utile à la représentation de données en elles-mêmes. Les éléments principaux de RDF sont les objets, leurs attributs et les valeurs de ces attributs. L'intérêt principal de RDF est de définir un mécanisme permettant de décrire des données indépendamment de tout domaine et de toute spécificité. De même qu'avec XML, RDF ne permet pas la déclaration de propriétés particulières ; leur définition est totalement libre.

Les schémas RDF (RDFS)[43] permettent de définir le vocabulaire utilisé dans les descriptions RDF. Ils confèrent un formalisme de représentation riche, incluant des classes, sous-classes, propriétés, sous-propriétés, des règles d'héritage de propriétés ..., mais ne normalisent pas les inférences que l'on pourrait faire avec. La structure objet-classe des RDFS permet de représenter un modèle du domaine en définissant des objets du domaine et leurs relations pour rendre compte d'une ontologie.

5.3 OIL

OIL (Ontology Inference Layer)[44] est un langage de représentation d'ontologie qui dérive de RDF. Les principaux fondements du langage OIL sont les langages de frame (tels que OKBC, XOL ou RDF) et les logiques de descriptions. OIL a été défini dans l'objectif de permettre la spécification et l'échange d'ontologies. Le premier objectif prévalant à la création du langue OIL, projet financé par la Communauté Européenne, était d'en faire un langage largement intuitif pour l'homme. Le langage devait également avoir une sémantique formelle bien définie et enfin, pour des raisons d'interopérabilité, être lié aux langages existants tels que XML ou RDF. Pour résumer, le défi des concepteurs du langage OIL était de réunir dans un même langage :

- la richesse épistémologique de la modélisation objet (frame-based logic),
- la sémantique formelle et un support de raisonnement efficace porté par la logique de description,
- et enfin les standards d'échanges de la communauté Web.

OIL permet la description d'ontologie sur la base des éléments de la logique de frame (Chabert-Ranwez, 2000) : les classes, les propriétés de classes, les relations entre les classes etc. Les classes s'organisent en hiérarchie incluant des classes et des sous-classes. Elles peuvent se combiner par le biais d'expressions logiques : intersection, union, etc. Les propriétés se déclarent de la même façon que les axiomes logiques. Elles peuvent être fonctionnelles (functional), c'est-à-dire n'avoir qu'une seule et même valeur, transitives ou symétriques. Des restrictions peuvent être apportées sur le type des propriétés ainsi que sur le nombre de valeurs qu'elles peuvent, le cas échéant, prendre.

Un des aspects fondamentaux de OIL est sa sémantique formelle, Ainsi, chaque classe est organisée en ensemble d'objets et chaque propriété en ensemble de paires d'objets, cette organisation devant bien évidemment se conformer aux contraintes spécifiées lors de la définition des classes et des propriétés.

[43] http://www.w3.org/TR/rdf-schema/
[44] http://www.ontoknowledge.org/oil/

OIL est organisé en couches hiérarchisées (*ever-increasing layers*) de différents sous-langages. Chaque couche, ou niveau, ajoute des fonctionnalités et partant, de la complexité au niveau précédent. Un agent, humain ou machine, traitant spécifiquement d'une ontologie définie par un sous-langage d'un certain niveau est ainsi en mesure de comprendre partiellement une ontologie définie dans un sous-langage de niveau supérieur.

Les différentes couches de OIL sont les suivantes :

- Standart OIL : permet d'exprimer les principales primitives de modélisation de façon suffisamment expressive et compréhensible. La sémantique est spécifiée, ainsi que les règles d'inférence.
- Instance OIL : permet de spécifier une modélisation plus profonde. Cette couche rend possible la spécification complète d'une base de données.
- Heavy OIL : est en cours de développement et devrait posséder des capacités encore supérieures de représentation et de raisonnement notamment la définition des règles et de méta-classes.

Le choix entre les différents sous-langages OIL de description d'ontologies dépend de la complexité même de l'ontologie en question. Il ne sert à rien d'exprimer une ontologie dans un sous-langage OIL comportant plus d'expressivité et de complexité que nécessaire.

Une ontologie exprimée en OIL est elle-même annotée avec des données propres telles que le titre de l'ontologie, son ou ses créateurs, sa date de création, sa date de réactualisation, etc. OIL suit les spécifications du W3C Dublin Core Standard pour exprimer ces méta-données sur l'ontologie elle-même.

L'objectif d'OIL est de permettre la définition d'environnements complets de gestion des connaissances (*knowledge management*) dans le cadre d'intranets.

5.4 DAML et DAML+OIL

DAML (DARPA Agent Markup Language)[45] est un langage permettant la représentation d'ontologies. Il a été développé par la DARPA aux États-Unis dans le but de développer des langages et des outils permettant de rendre les contenus de documents accessibles et exploitables par des machines. DAML est une combinaison de XML et de RDF permettant de spécifier des objets mais également les relations entre ces objets. La dernière version de DAML se combine avec OIL (DAML+OIL)[46]. Ce nouveau langage supporte désormais les types de données primitifs (tels qu'on les trouve dans la norme XML Schéma) et la définition d'un certain nombre d'axiomes comme l'équivalence de classes ou de propriétés.

5.5 OWL

OWL (Ontology Web Language), créé en 2001 par le W3C, hérite du langage DAML+OIL et doit permettre de représenter des ontologies sur le Web. OWL est destiné à être utilisé lorsque les

[45] `http://www.daml.org/`
[46] `http://www.w3.org/TR/daml+oil-reference`

informations contenues dans les documents doivent être traitées par des applications logicielles, c'est-à-dire lorsqu'elles ne sont pas simplement « montrées » à l'utilisateur. Une ontologie OWL est composée d'un en-tête (métadonnées), d'axiomes et de faits. Les axiomes concernent la définition complète ou partielle de concepts et de relations (ou priorités), la spécification de propriétés sur les relations (propriétés algébriques) et la définition d'axiomes sur les classes et les relations (équivalence, expression booléenne). Parmi les relations, on distingue celles dont le domaine de valeur sera de type primitif (attribut) de celles dont le domaine de valeur sera un autre concept (relation). Les faits concernent des individus pour lesquels on donne des valeurs aux propriétés des classes dont ils sont les instances. OWL fournit en fait trois sous-langages, d'expressivité croissante, nommés OWL Lite, OWL DL et OWL Full.

- Le langage OWL Lite peut être vu comme une extension du langage RDFS, mais auquel on aurait enlevé certaines fonctionnalités. Le principal intérêt de ce langage est de permettre la modélisation d'ontologies simples, d'une complexité formelle peu élevée, de sorte qu'il soit facile d'implémenter des raisonneurs corrects et complets.
- Le langage OWL DL contient des constructeurs supplémentaires, mais il ne peut être utilisé qu'avec certaines restrictions. Par exemple, une classe ne peut pas être une instance d'une autre classe. Il en résulte un langage un peu plus expressif mais toujours décidable, c'est-à-dire que les conséquences sont toujours calculables en un temps fini.
- Le langage OWL Full dispose des mêmes constructeurs que OWL DL mais il les interprète de manière plus large. Ainsi, une classe peut cette fois être vue comme un ensemble d'individus (définition extensionnelle) ou comme un individu à lui tout seul (définition intensionnelle) qui pourra, par exemple, donner une valeur à une propriété. À ce titre OWL Full devient clairement un sur-ensemble de RDF. Cette expressivité accrue est gagnée au détriment de la complexité : le langage OWL Full n'est plus décidable.

La sémantique de OWL est basée sur une hypothèse de monde ouvert qui est particulièrement bien adaptée au Web. Concrètement, cela signifie que le système est capable d'effectuer des raisonnements même s'il n'a pas une connaissance complète du monde, puisque tout fait absent de la base n'est pas systématiquement considéré comme faux. À l'instar de DAML + OIL , le langage dispose d'une sémantique en théorie des modèles (compatible avec celle de RDFS pour OWL Full) et d'une axiomatique (une traduction en logique du premier ordre est possible).

La vision d'un Web sémantique, dans lequel l'information serait accessible et manipulable automatiquement par la machine, s'appuie sur une pile de langages jouant chacun un rôle particulier (*cf.* figure 3.15).

- XML fournit une manière de représenter des documents structurés, mais il n'impose aucune contrainte sémantique sur les documents produits ;
- XML Schéma permet de contraindre la structure des documents XML ;
- RDF est un modèle de données simple, fondé sur des ressources et des relations entre ces ressources, équipé d'une sémantique et qui peut se représenter en XML ;
- RDF Schema permet de définir le vocabulaire pour décrire des classes et des propriétés, hiérarchisées en taxinomies ;

– OWL fournit d'avantage de primitives de modélisation pour décrire des ontologies plus riches sur le web.

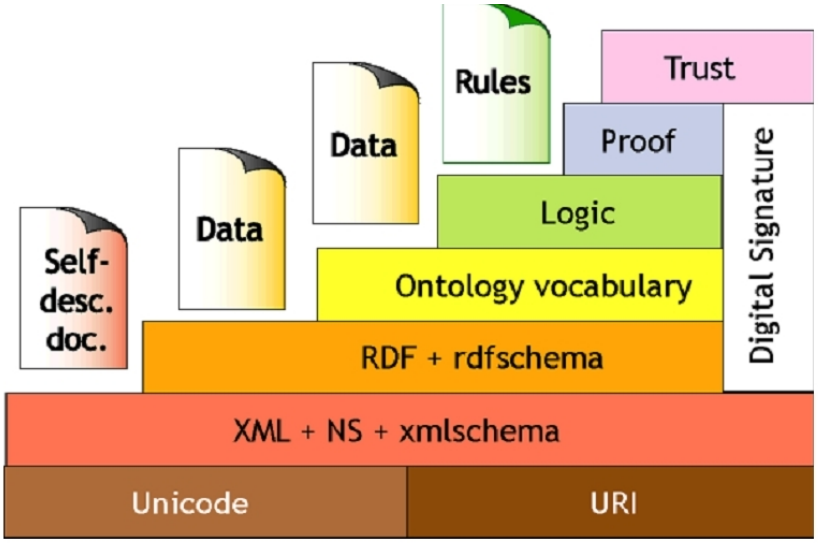


FIG. 3.15 – Le « cake » de Tim Berners Lee.

Comme nous venons de le voir, l'éventail des choix possibles concernant les langages de représentation et de spécification d'ontologies est très large. Cependant, OWL tend à s'imposer. Le choix du langage est fondé principalement sur des critères techniques, notamment en fonction des degrés d'expressivité et de formalisme voulus, mais également dans un souci de portabilité, en fonction de la diffusion, de la visibilité et de l'intérêt que l'on veut succiter au sein d'une communauté de collègues et d'utilisateurs.

6 Éditeurs d'ontologies

Il existe un certain nombre d'outils permettant de construire des ontologies. Nous allons en premier présenter des outils d'ingénierie ontologique qui permettent à l'utilisateur de créer la ressource « à la volée », de manière indépendante des langages de représentation, et de prendre en charge la phase d'opérationnalisation de l'ontologie en l'exportant dans des langages informatisés standards. Nous présentons ensuite des outils et plateformes qui mettent l'accent sur l'importance du texte comme source privilégiée des connaissances du domaine à modéliser en vue de construire l'ontologie.

6.1 PROTÉGÉ

PROTÉGÉ [47] a été développé par le *Stanford Medical Informatics* de l'université de médecine de Stanford depuis 1995. Il est construit autour d'un modèle de connaissances inspiré par le paradigme des frames : classes, slots (attributs) et facets (contraintes sur les attributs) sont les primitives de modélisation proposées. Ce modèle autorise une liberté de conception importante, puisque le contenu des formulaires de spécification des classes peut être modifié suivant les besoins, via un système de méta-classes, qui constituent des sortes de « patrons » pour les classes du modèle du domaine. Il est adapté à la construction d'ontologies depuis la version PROTÉGÉ 2000. L'interface très complète ainsi que l'architecture logicielle bien pensée permettant l'insertion de pluggins, notamment des pluggins pour gérer les représentations sous forme graphique, par exemple OWLViz[48] (*cf.* figure 3.16), ont grandement contribué au succès de PROTÉGÉ. En quelques années, cet éditeur s'est imposé comme la référence, avec une communauté d'utilisateurs extrêmement importante et active. Ses nombreuses extensions lui permettent en particulier de gérer des langages standards comme RDF et surtout OWL (Knublauch *et al.*, 2004), de créer des axiomes formels de manière intuitive, d'accéder aux ontologies par des interfaces graphiques évoluées, de comparer et fusionner des ontologies avec la suite PROMPT [49] (Noy & Musen, 2003) …. Il est également possible de faire fonctionner des raisonneurs, comme RACER[50] (*Renamed ABox and Concept Expression Reasoner*) pour le langage OWL par exemple, pour vérifier la cohérence et la consistance de la structure ontologique.

La prédominance de PROTÉGÉ ne pourra qu'être renforcée par le lancement de l'initiative CO-ODE[51] (*Collaborative Open Ontology Development Environment project*) en septembre 2003, qui a pour objectif la création d'outils d'assistance à la création d'ontologies OWL riches et cohérentes[52], et qui se concentrent sur cet éditeur.

6.2 OILED

L'éditeur OILED [53] a été développé par l'université de Manchester pour éditer des ontologies dans les langages de représentation OIL , puis DAML+OIL , les précurseurs de OWL (Bechhofer *et al.*, 2001). Il est donc explicitement orienté vers la représentation en logique de description expressive et, à ce titre, fournit tous les éléments d'interface permettant de spécifier des hiérarchies de concepts et de rôles, ainsi que la construction des expressions complexes définissant ces entités.

[47] Auparavant appelé PROTÉGÉ 2000, cet éditeur a repris le nom de l'outil d'acquisition des connaissances qui l'a précédé. PROTÉGÉ est disponible à l'adresse suivante : `http://protege.stanford.edu/`.

[48] OWLViz est téléchargeable à l'adresse suivante : `http://www.co-ode.org/downloads/binaries/OWLViz_Build_17.zip`.

[49] La suite PROMPT est disponible à l'adresse suivante : `http://protege.cim3.net/cgi-bin/wiki.pl?Prompt`.

[50] `http://www.racer-systems.com/products/racerpro/index.phtml`.

[51] `http://www.co-ode.org/`

[52] Un certain nombre d'ontologies créées à l'initiative de CO-ODE sont disponible à l'adresse suivante : `http://www.co-ode.org/ontologies/`.

[53] `http://oiled.man.ac.uk/`.

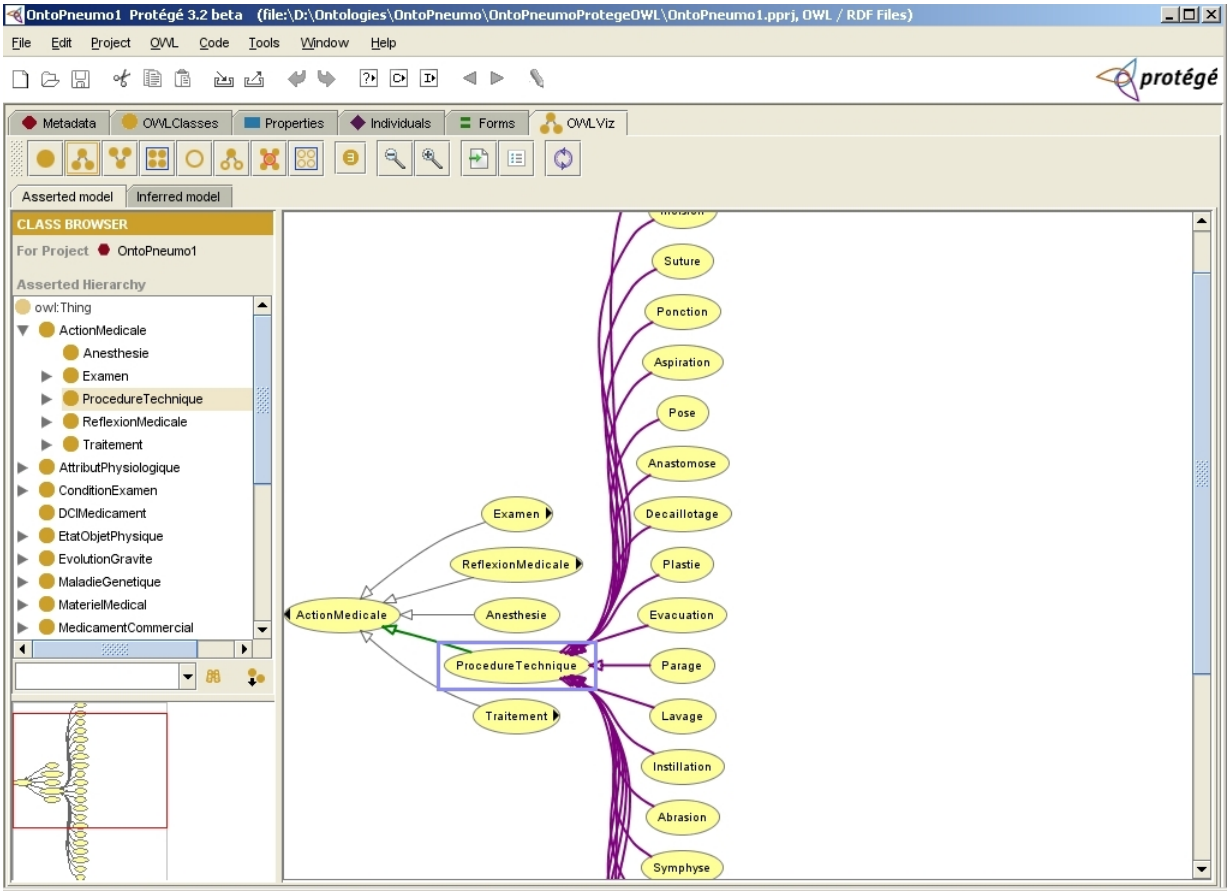


FIG. 3.16 – Extrait de l'ontologie de la pneumologie vue avec le pluggin OWLViz de PROTÉGÉ centré sur le concept *ProcedureTechnique*.

À l'origine, il n'a pas d'autre ambition que d'illustrer les vertus du langage pour lequel il a été créé. Les versions disponibles de OILED ne constituent pas un environnement complet pour le développement d'ontologies d'envergure. En effet, cet outil n'implémente pas la migration et l'intégration d'ontologies, ne gère pas les différentes versions et autres activités impliquées dans la construction d'ontologies. Néanmoins, la simplicité, la robustesse de cet outil et la présence d'un raisonneur de logique de description FACT [54], capable de tester la satisfiabilité des ontologies construites ou d'expliciter de nouvelles relations de subsomption entre concepts complexes, en font un outil de référence relativement populaire avec plus de 2 000 téléchargements. Comme le soulignent les concepteurs, il s'agit plutôt d'un « bloc-notes » offrant assez de fonctionnalités pour permettre à des utilisateurs de construire des ontologies et de démontrer comment il est possible d'employer le raisonneur de FACT pour examiner les ontologies et en assurer l'uniformité.

A. Rector *et al.* met à disposition des utilisateurs intéressés un tutoriel[55] pour la création

[54] FAst Classification of Terminologies, `http://www.cs.man.ac.uk/horrocks`.

[55] Le tutoriel d'A. Rector *et al.*, intitulé *OilEd Normalised Ontology Tutorial - Biomedical Ver-*

d'ontologies biomédicales avec OILEd. La figure 3.17 illustre la hiérarchie des concepts obtenue à la fin de ce tutoriel.

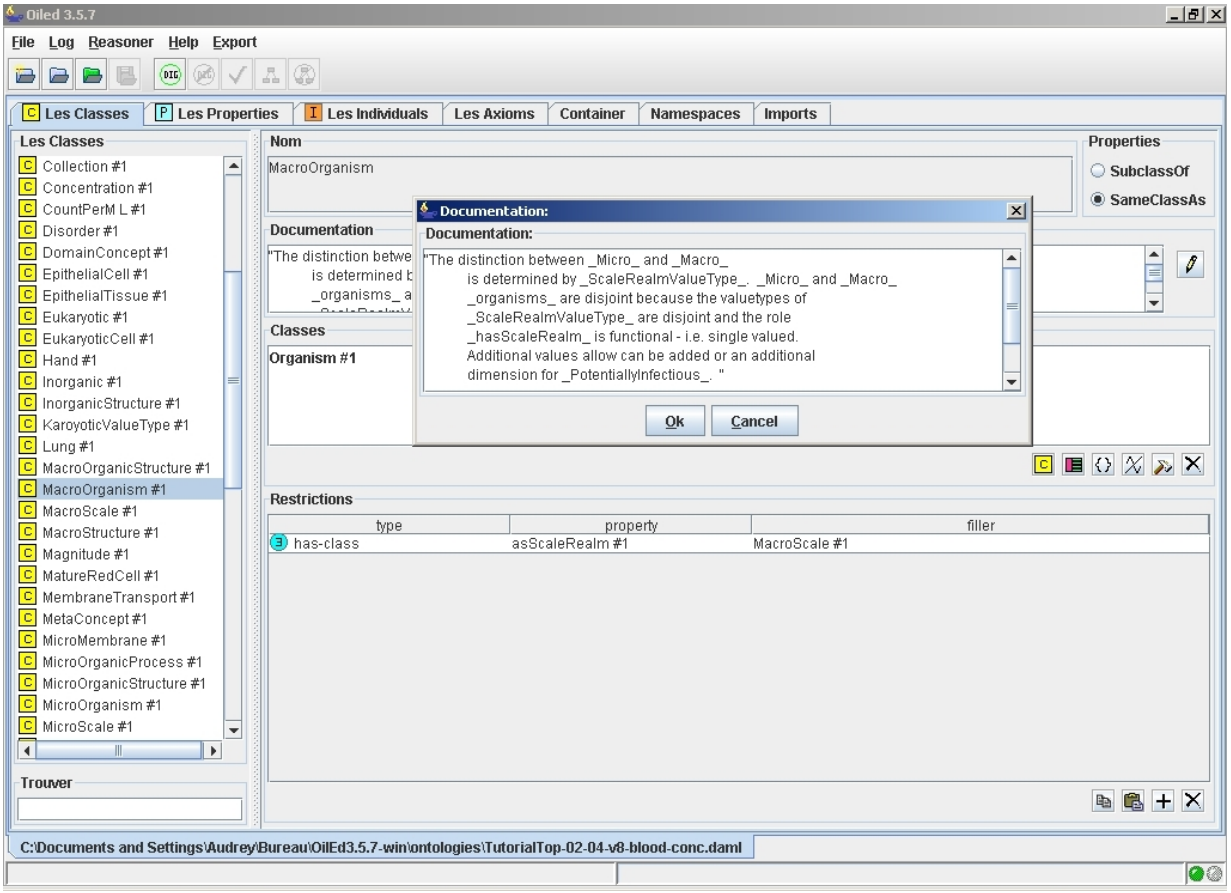


FIG. 3.17 – Extrait de l'ontologie biomédicale issue du tutoriel d'A. Rector *et al.* vue avec OILEd centré sur le concept *MacroOrganism*.

OILEd permet d'exporter les ontologies construites dans des langages standards tels que DAML+OIL, RDFS ou OWL.

6.3 ONTOEDIT

ONTOEDIT est un outil mis au point par l'institut AIFB de l'université de Karlsruhe et qui est maintenant commercialisé par la société Ontoprise GmbH. Il s'inspire de l'approche par *frames* mais gère de nombreux formats libres de la communauté Web sémantique (FLogic, DAML+OIL , RDFS) et a le mérite de s'appuyer sur une réflexion méthodologique significative, celle d'On-To-Knowledge (Sure *et al.*, 2002). Il s'est en effet le premier intéressé à la modélisation « intuitive »

sion, est disponible à l'adresse suivante : `http://www.cs.man.ac.uk/mig/ontology-tutorial/oiled-biomedical-ontology-tutorial.zip`.

des axiomes, indépendamment d'un formalisme ou d'un autre, pour faciliter la traduction d'un langage de représentation à un autre. Cet outil met à disposition de l'utilisateur plusieurs vues graphiques correspondant aux différentes phases de conception de l'ontologie. Il permet d'éditer une hiérarchie de concepts ou de classes. Ces concepts peuvent être abstraits ou concrets, ce qui indique s'il est permis d'instancier le concept en question.

S'efforçant de mettre en œuvre les propositions du Projet On-To-Knowledge, il propose également une gestion originale des « questions de compétences ». Il s'agit des questions auxquelles les connaissances ontologiques doivent apporter une réponse. Un petit outil compare au niveau lexical les termes extraits des différentes questions pour en déduire automatiquement d'éventuelles subsomptions (*cf.* figure 3.18). Comme le fait remarquer R. Troncy (2002), ce procédé semble très peu fiable car il repose sur l'identification du nom du concept dans ses spécialisations. De plus, ONTOEDIT gère la synonymie en admettant plusieurs noms pour un même concept. Visiblement aucune distinction n'est faite entre le terme désignant le concept et ceux désignant les connaissances qu'il recouvre. ONTOEDIT est un des seuls éditeurs que nous connaissons, avec DOE, à s'attaquer au problème de la synonymie. La solution proposée par DOE nous semble plus intéressante (*cf.* section 6.5 de ce chapitre). ONTOEDIT permet d'exporter les ontologies construites dans différents langages : RDF(S), OXML[56], DAML+OIL et FLogic.

6.4 WebODE

WebODE [57] est une plateforme en ligne développée par le groupe *Ontological Engineering* du département d'Intelligence artificielle de la faculté d'Informatique de l'université polytechnique de Madrid (Corcho *et al.*, 2002). Elle se place au niveau méthodologique dans la lignée d'ODE, un éditeur qui assurait le support de METHONTOLOGY, la méthodologie proposée par ce laboratoire. L'ambition nouvelle de WebODE par rapport à ODE est de considérer que les ontologies doivent être construites et mises à dispositon via le web pour faciliter le développement d'application du Web sémantique. WebODE est composée de plusieurs modules : un éditeur d'ontologie (*cf.* figure 3.19) qui intègre la plupart des services nécessaires à la construction d'ontologies (édition, navigation, comparaison, fusion, raisonnement ...), un système de gestion des connaissances à base ontologique, un générateur automatique de portail du Web sémantique, un outil pour annoter les ressources du web et un éditeur de services pour le Web sémantique. La plateforme WebODE met l'accent sur la possibilité d'un travail collaboratif et sur la possibilité, comme dans PROTÉGÉ, d'étendre la plateforme à l'aide de modules complémentaires, comme un moteur d'inférences ou bien l'outil ODEClean, intégration dans WebODE de la méthode mise au point par C. Welty et N. Guarino. WebODE, similaire en cela aux autres éditeurs, accepte l'export et l'import d'ontologies en RDFS, DAML+OIL et OWL.

[56] ONTOEDIT's XML-based Ontology representation Language

[57] `http://webode.dia.fi.upm.es/WebODEWeb/index.html`

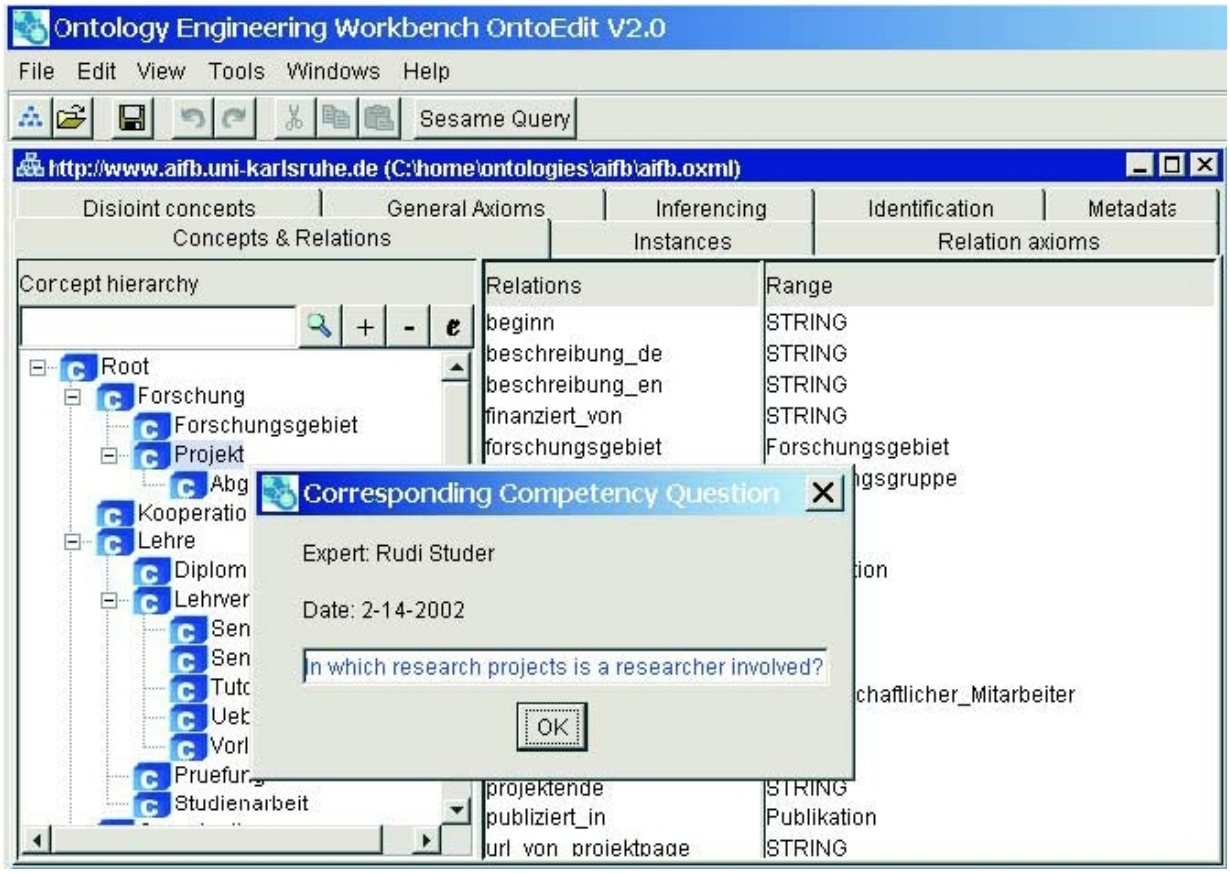


FIG. 3.18 – ONTOEDIT fait le lien entre la hiérarchie de concepts et les « questions de compétences ».

6.5 DOE

DOE [58] (*Differential Ontologies Editor*) a été développé à l'Institut National de l'Audiovisuel par R. Troncy et A. Isaac (2002). Si tous les outils précédemment présentés peuvent être considérés, en tout cas dans une première approche, comme satisfaisants en matière d'expressivité ou d'interface, il existe toujours un certain vide méthodologique. La structuration des taxinomies produites est, en particulier, très peu prise en charge : on ne peut pas dire que les outils existants guident réellement l'utilisateur lors de cette étape primordiale. La piste d'un traitement plus complet des informations véhiculées par le langage ne semble pas non plus pouvoir être suivie grâce à ces environnements, puisque les commentaires demeurent toujours accessoires.

L'éditeur DOE offre des interfaces de création, modification et suppression de concepts et de relations, une représentation graphique de l'arbre ontologique, et des fonctionnalités de recherche et de navigation dans la structure créée. Il propose une interface ergonomique pour associer à

[58] DOE est un logiciel libre sous licence GPL disponible à l'adresse suivante : `http://opales.ina.fr/public`.

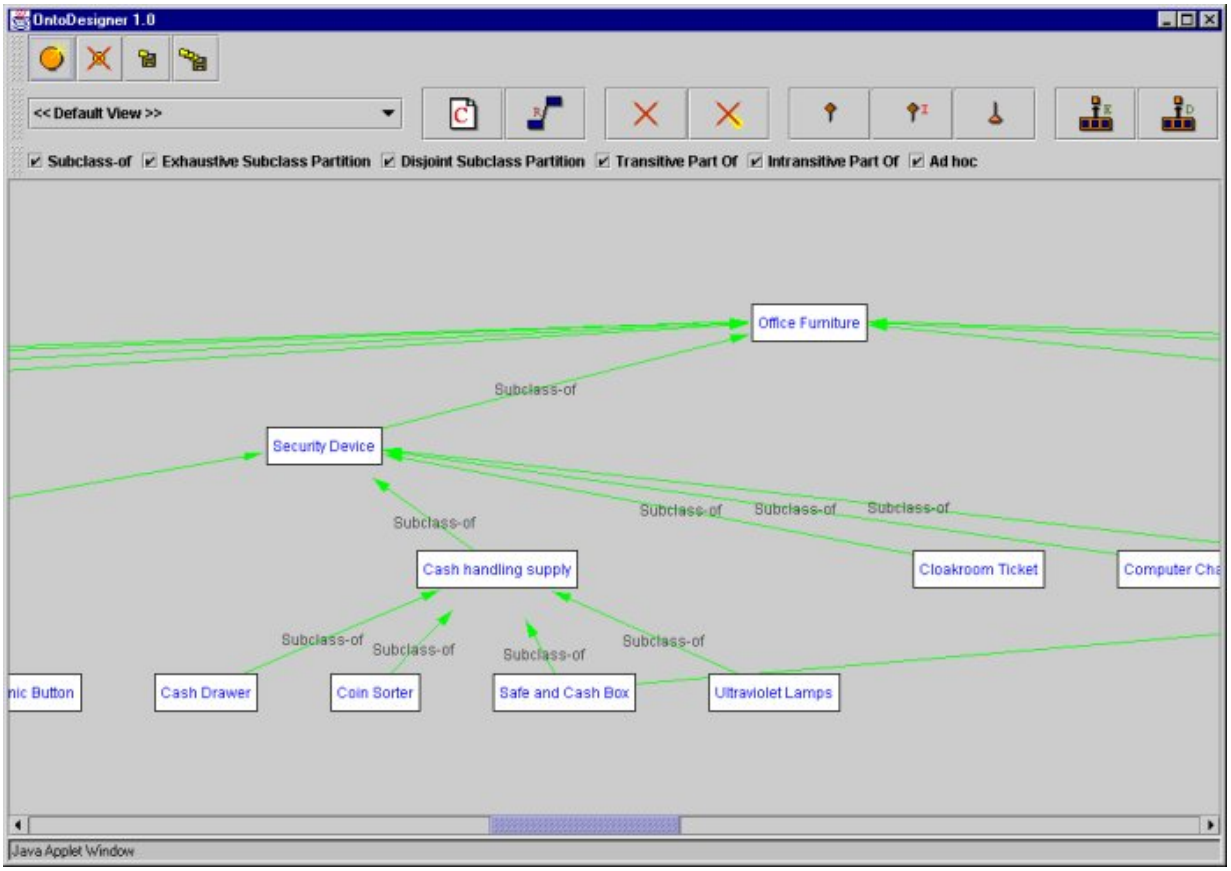


FIG. 3.19 – Représentation graphique d'une ontologie vue avec OntoDesigner, l'éditeur d'ontologie de WebODE.

chaque terme sa définition encyclopédique, ses éventuels synonymes[59] et ses principes différentiels et cela en plusieurs langues (*cf.* figure 3.20). Cet outil n'a pas pour ambition de concurrencer les grands environnements existants, mais plutôt d'implémenter la méthodologie de structuration différentielle ARCHONTE en permettant à l'ontologue de bien faire la distinction entre ontologie différentielle et référentielle (plus de précisions seront données en section 4.6 de ce chapitre). L'outil assiste également la saisie des principes différentiels issus de la méthodologie en automatisant partiellement cette tâche. Le modèle de représentation de l'ontologie est finalement proche de celui du langage RDFS, à ceci près qu'il autorise la modélisation de relations n-aires. Dans le domaine formel, l'éditeur est capable de faire quelques inférences en vérifiant la consistance de l'ontologie (propagation de l'arité le long de la hiérarchie des relations et héritage des domaines par exemple).

Cet éditeur permet également d'ajouter des individus à l'ontologie. Finalement, le passage à une ontologie computationnelle s'effectue par un export de l'ontologie formelle dans un certain nombre de langages opérationnels tels que RDFS, DAML+OIL , OWL et CGXML. Cette traduc-

[59] Sur la question de la synonymie, voir aussi l'outil ONTOEDIT, section 6.3 de ce chapitre.

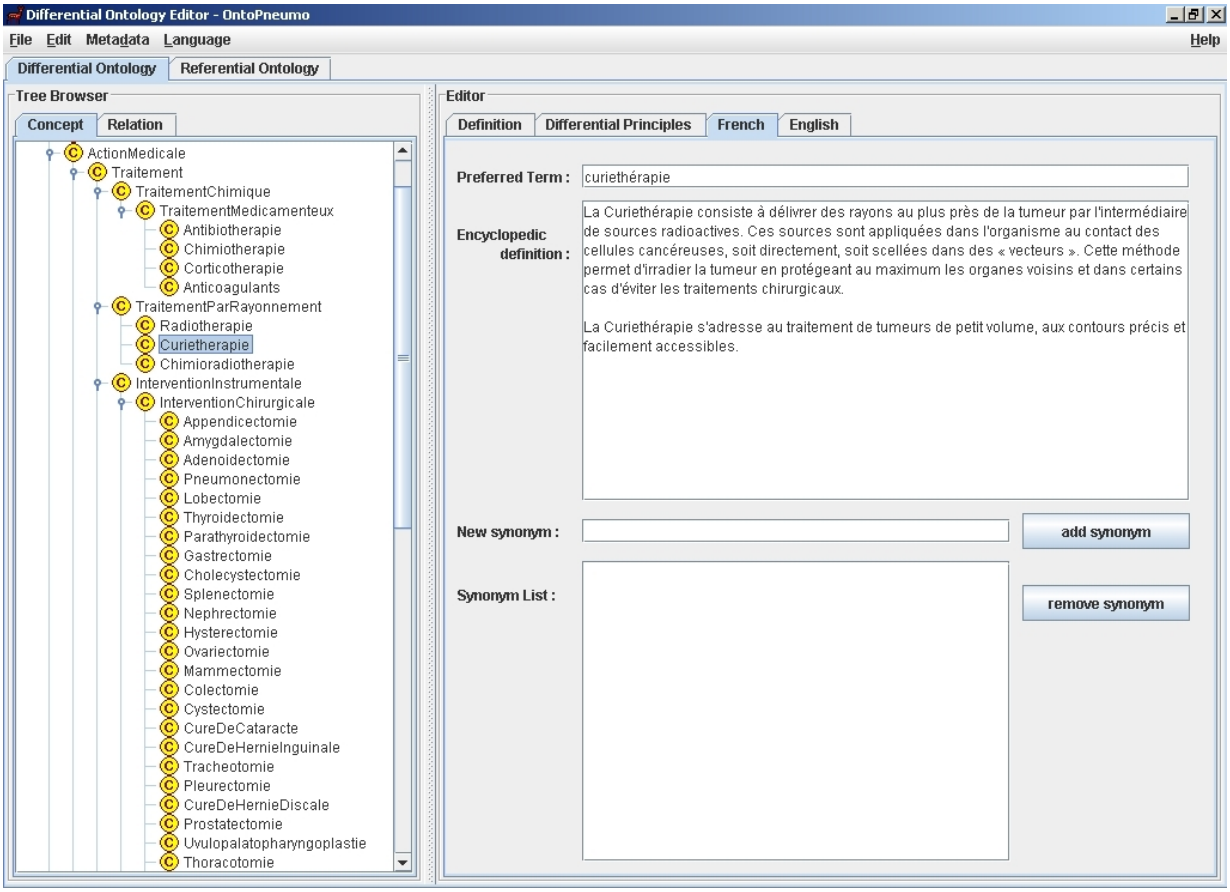


FIG. 3.20 – Extrait de l'ontologie de la pneumologie vue avec DOE, centré sur le concept *Curietherapie*.

tion s'effectue grâce à des feuilles de style XSLT appliquées au format de sauvegarde XML de l'éditeur. De la même façon, DOE peut importer des ontologies modélisées dans d'autres outils grâce à des feuilles XSLT dédiées. Pour être tout à fait complet, il faut noter que l'éditeur permet aussi de sauvegarder un certain nombre de métadonnées concernant l'ontologie elle-même. Il suit ici les propositions du Dublin Core.

L'interopérabilité de DOE avec d'autres outils, tel que PROTÉGÉ par exemple, est nécessaire dans la mesure où il ne met pas en œuvre les fonctionnalités d'expressivité formelle prises en charge par les autres environnements (Troncy *et al.*, 2003). Cet outil ne constitue pas, pour l'instant, un environnement de développement d'ontologies complet (*cf.* chapitre 7, section 3) mais se situe en amont d'autres éditeurs implémentant, par exemple, les logiques de description. Il met l'accent sur la structuration de taxinomies.

7 Outils d'ingénierie ontologique à partir de textes

Ces outils sont essentiellement dédiés à l'extraction, à partir de documents, des concepts du domaine et des relations existant entre eux, mais offrent également des fonctionnalités de structuration permettant de bâtir de véritables ontologies. Nous présentons dans cette section trois outils : deux suites logicielles outillant une méthodologie globale et un outil d'acquisition de termes à partir d'analyse linguistique. TERMINAE est un exemple d'outil de conceptualisation, qui a évolué en intégrant des fonctionnalités liées à l'ontologisation et même au raisonnement. Text-To-Onto est une application maintenant intégrée à un véritable atelier d'ingénierie logicielle nommé KAON. Ces deux outils utilisent des techniques de traitement des langues naturelles pour identifier, dans un corpus, les concepts et relations du domaine. SYNTEX- UPERY est un outil composé de deux modules, l'un propose une analyse syntaxique complète du corpus et l'autre propose une approche distributionnelle pour aider à l'organisation des termes d'un domaine. l'interface permettant de visualiser les résultats de ces analyses s'appelle TERMONTO.

7.1 TERMINAE

TERMINAE , développée au LIPN de l'Université Paris-Nord II, est une méthode et une plateforme logicielle d'aide à l'élaboration de ressources terminologiques et ontologiques à partir de textes (Aussenac-Gilles *et al.*, 2005; Szulman *et al.*, 2002). Cet outil intègre un environnement d'étude terminologique, un environnement d'aide à la conceptualisation et un système de gestion d'ontologies. L'un des points marquants de TERMINAE est qu'il conserve le lien vers le corpus. Ainsi, l'utilisateur peut, à tout moment, consulter le contexte dans lequel une notion apparaît. Ce lien permet de rendre compte de phénomènes linguistiques tels que, par exemple, la polysémie ou la synonymie, et de concerver une trace des choix de l'ontologue quant à l'organisation de la hiérarchie ontologique. L'environnement d'étude terminologique permet :

- de dépouiller les résultats des extracteurs de candidats termes comme LEXTER (Bourigault, 1994b) et maintenant SYNTEX (Bourigault *et al.*, 2005) ou YaTeA (*Yet another Term extrActor*) ;
- d'explorer un corpus à l'aide de patrons lexicaux et/ou syntaxiques grâce au module LINGUAE ;
- de déterminer d'éventuels synonymes grâce à l'outil SynoTerm[60].

Le formalisme de représentation mis en œuvre dans TERMINAE est proche des logiques de description (*cf.* section 3.2). Un export dans le langage OIL est possible pour valider l'ontologie à l'aide du raisonneur FaCT (intégré à l'outil OILEd présenté section 6.2). Le modèle des données de TERMINAE fait la différence entre les termes extraits du corpus, les notions décrites dans des fiches terminologiques, les concepts décrits dans des fiches de modélisation et enfin les concepts formels caractérisés à l'aide des logiques de description. Chacun de ces niveaux de représentation rend compte d'une étape dans le processus de modélisation. La figure 3.21 montre l'interface de gestion d'un concept dans TERMINAE. Le système de gestion d'ontologie est accessible via un éditeur d'ontologie. Celui-ci permet de créer, modifier et visualiser les concepts ainsi que

[60] `http://www-lipn.univ-paris13.fr/~hamon/SynoTerm/SynoTerm.html`

leurs relations. Il vérifie le respect des contraintes imposées par la sémantique du langage tout en acceptant que l'ontologie ne soit pas valide. L'interface permet de visualiser l'ontologie sous forme graphique ou bien sous forme hiérarchique. Les ontologies sont sauvegardées à priori en XML mais les imports et exports peuvent se faire en OWL ou en RDFS.

FIG. 3.21 – Interface « concept » de TERMINAE.

La démarche suivie dans TERMINAE est peu automatisée, l'ontologue décide du contenu de l'ontologie, du niveau de granularité des connaissances représentées et de leur organisation. Cet outil aide à l'interprétation des textes et à l'extraction des connaissances mais laisse à l'ontologue un rôle fondamental.

7.2 Text-To-Onto et KAON

Text-To-Onto, développé à l'institut AIFB de l'Université de Karlsruhe, est un module de création d'ontologie intégré à la plateforme KAON (*KArlsruhe ONtology*). Il implémente des algorithmes de fouille de textes sur des corpus textuels pour construire des ontologies de manière semi-automatique. Il inclut plusieurs traitements comme, par exemple, l'extraction des termes en utilisant soit des calculs statistiques, soit des expressions régulières, l'identification des relations à l'aide de patrons lexico-syntaxiques ou de calculs de proximités … Egalement originaire de l'institut AIFB de l'Université de Karlsruhe, la plateforme KAON est un environnement de création d'ontologie qui comprend plusieurs modules, chacun correspondant à une étape nécessaire du processus de développement : création, stockage, exploitation … L'ontologue peut travailler

avec des ontologies exprimées dans différents langages de représentation aussi longtemps que les primitives de représentation sont définies pour être sémantiquement équivalentes aux primitives du langage de KAON qui fonctionne avec les schémas RDF et les ontologies en DAML+OIL . Cette plateforme est associée à OntoEdit (*cf.* section 6.3) qui peut prendre en entrée la sortie du module Text-To-Onto. KAON utilise le modèle de connaissance de RDFS et est orienté vers l'utilisation des ontologies sur le web, l'application KAON Portal permettant la recherche et le parcours d'ontologie via un navigateur Web.

7.3 SYNTEX- UPERY

SYNTEX- UPERY est un outil composé de deux modules développé par D. Bourigault à l'ERSS (Toulouse) (Bourigault *et al.*, 2005). SYNTEX est un module d'analyse syntaxique développé à l'origine pour remplacer le logiciel LEXTER. Il prend en entrée un corpus de textes étiquetés (à chaque mot du texte est affecté une catégorie grammaticale par l'outil Treetagger, développé à l'Université de Stuttgart) en français ou en anglais, effectue l'analyse syntaxique de chacune des phrases du corpus et produit comme résultat un réseau de mots (noms, adjectifs, verbes ...) et de syntagmes nominaux, adjectivaux ou verbaux extraits du corpus. Dans le réseau terminologique construit, chaque syntagme est relié d'une part à sa tête et d'autre part à son (ses) expansion(s). SYNTEX fournit un certain nombre d'informations numériques associées à chacun des candidats termes, en particulier la fréquence (le nombre de fois que le candidat terme a été repéré dans le corpus), la productivité (le nombre de contextes dans lesquels le candidat terme apparaît) et la répartition (le nombre d'articles différents dans lesquels le candidat terme a été repéré). La figure 3.22 représente le formulaire d'entrée du module SYNTEX dans l'interface TERMONTO.

UPERY est un module d'analyse distributionnelle (Bourigault, 2002). Il exploite l'ensemble des données présentes dans le réseau terminologique construit par SYNTEX pour effectuer un calcul des proximités distributionnelles entre ces unités. UPERY rapproche deux à deux des candidats termes qui se retrouvent dans les mêmes configurations syntaxiques, c'est-à-dire qui sont exprimés en tête d'un même ensemble de termes.

Nous avons utilisé cet outil dans notre travail de doctorat aussi nous préciserons certains points méthodologiques dans le chapitre 4 consacré à nos expérimentations.

7.4 Conclusion

En dehors des caractéristiques techniques, les critères utilisés pour évaluer un outil d'ingénierie ontologique portent sur le modèle de représentation de connaissances qu'il utilise, les fonctionnalités de raisonnements qu'il offre, son interopérabilité avec les autres outils, et la facilité d'usage qu'il propose. La plupart des outils utilisent pour paradigme de représentation de connaissances le modèle des *frames* enrichi de formules de la logique du premier ordre ou celui des logiques de description.

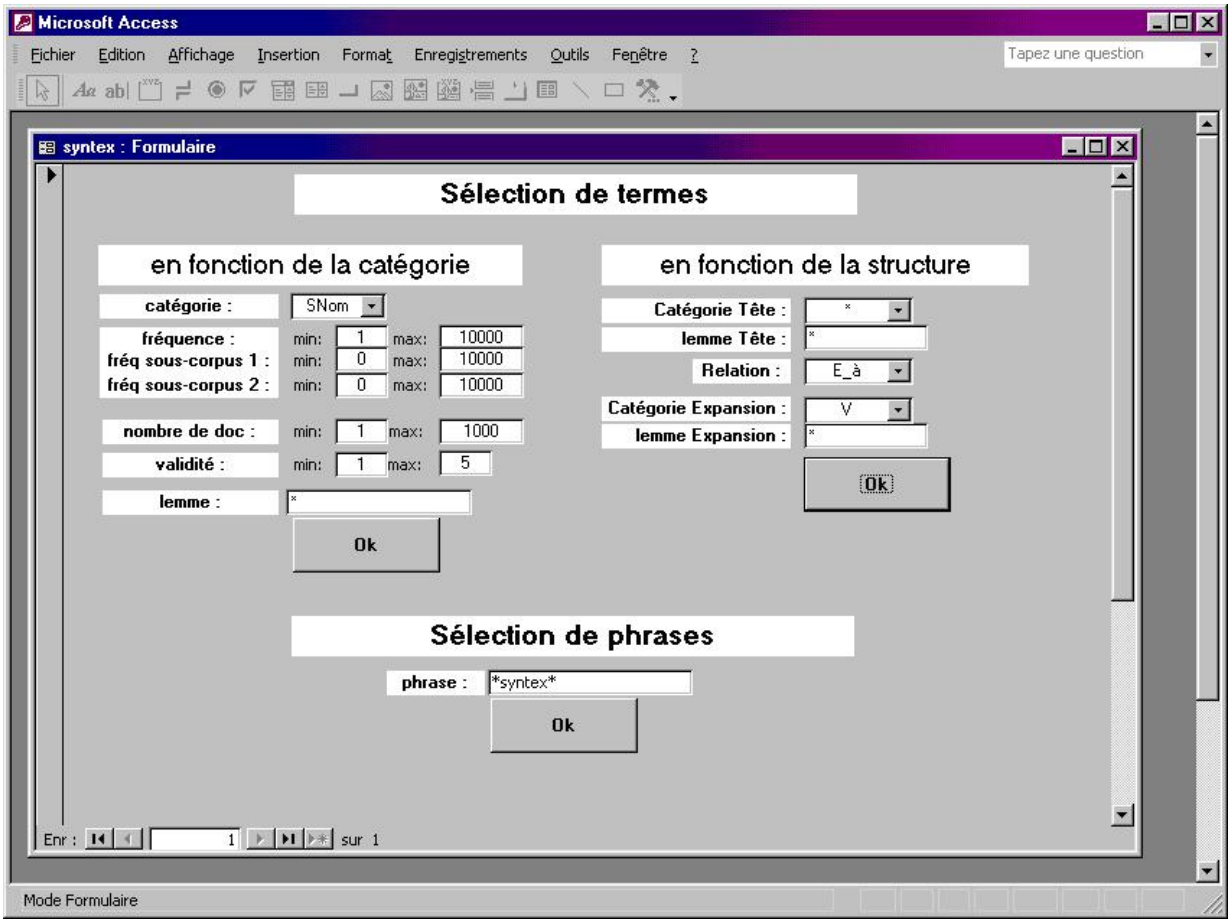


FIG. 3.22 – Formulaire d'entrée du module SYNTEX dans l'interface TERMONTO.

CHAPITRE 4

Construction d'une ontologie dans le domaine de la pneumologie

« L'examen de ces propriétés forme cette branche de la philosophie dont toutes les autres empruntent en partie leurs principes : on la nomme l'ontologie ou science de l'être, ou métaphysique générale. »
Jean le Rond d'Alembert
Discours préliminaire à l'Encyclopédie (1751)

Nous construisons une ontologie régionale de la pneumologie à partir de ressources textuelles. Pour cela, nous nous sommes dotés d'un certain nombre de principes de bonne modélisation. Ces critères sont également utilisés pour évaluer le modèle achevé, aussi nous les détaillons au chapitre 6, sections 2.2 et 2.2.2. Pour développer les corpus nécessaires à la structuration de l'ontologie, nous appliquons sur ces ressources des techniques appartenant au domaine du Traitement automatique du langage. La méthode ARCHONTE *que nous employons a été mise au point par B. Bachimont au sein du groupe Terminologie et Intelligence Artificielle et est fondée sur, entre autres, les principes de la sémantique différentielle (Bachimont, 2000). Notre principale hypothèse de recherche concerne l'utilisation en parallèle de deux méthodes pour enrichir le travail de construction de l'ontologie et particulièrement la mise en œuvre des principes différentiels : a) une méthode éprouvée qui consiste à construire des ressources termino-ontologiques par analyse distributionnelle (Bourigault & Lame, 2002), et b) une méthode fondée sur la définition* a priori *d'une relation sémantique, puis sur l'observation de séquences en corpus qui véhiculent la relation souhaitée (Séguéla, 2001). L'expérimentation menée avec cette seconde méthode (b) a été faite en collaboration avec V. Malaisé*[1]*. Sachant qu'aucune ontologie*

[1]Nous avons utilisé les patrons lexico-syntaxiques développés lors d'un précédent travail par V. Malaisé ainsi que

ne couvre le domaine de la pneumologie, notre objectif était double : il s'agissait, d'une part, de construire l'ontologie de la pneumologie et, d'autre part, d'apporter des précisions sur les premières étapes de la méthode. Nous proposons notre propre expérimentation de construction d'ontologie médicale dans la même optique que le travail de Le Moigno et al. *(2002b). Un point de vue quelque peu différent a cependant été adopté puisque l'ontologie est construite par un ingénieur des connaissances et non par un expert du domaine médical comme dans ce précédent travail. Rappelons que l'intérêt consiste à mettre au point un processus méthodologique précis, destiné à l'ingénieur des connaissances, de manière à ne faire appel à l'expert médical que pour des moments particuliers de validation.*

*Nous apportons dans la section 1 des précisions sur les principes méthodologiques d'*ARCHONTE *et soulignons l'originalité de cette méthode. La section 2 s'intéresse au corpus de référence, à sa définition et aux ressources qui le constitue. La section 3 décrit en détails les différents traitements que nous appliquons sur nos corpus. La section 4 s'intéresse au choix des connaissances à modéliser, aux opérations d'extraction, de filtrage et de sélection de ces connaissances. La section 5 compare les résultats obtenus avec l'analyse distributionnelle et le repérage par patrons lexico-syntaxiques sur nos corpus et l'apport de cette expérience sur la définition des principes différentiels. La section 7 résume les étapes de formalisation et d'opérationalisation de l'ontologie de la pneumologie. Nous présentons ensuite brièvement, en section 6, notre tentative de « raccrochage » de cette ontologie avec l'ontologie de haut niveau issue du projet* MENELAS. *Puis, en section 8, nous résumons les étapes successives de la construction de la hiérarchie de l'ontologie. Enfin, nous concluons ce chapitre, en section 9, en discutant les résultats obtenus.*

1 Méthode ARCHONTE : principes et originalité

Le langage médical est caractérisé par un vocabulaire extrêmement riche et difficile à manipuler. Il n'y a pas de consensus établi sur la définition des termes employés. Les synonymes sont nombreux (plusieurs termes désignant le même objet) tandis que le même terme peut avoir plusieurs significations selon l'auteur ou le contexte (polysémie). Les textes médicaux sont donc souvent imprécis, ambigus d'autant qu'ils font un large usage d'abréviations et d'acronymes. Pour permettre une description et une communication efficaces et dépourvues d'ambiguïté, a fortiori un traitement automatique, un minimum de standardisation du langage est nécessaire. Comme nous l'avons vu au chapitre 3 section 4, peu de méthodologies proposent réellement de guider l'ingénieur des connaissances pour organiser les connaissances d'un domaine et, par la suite, les concepts entre eux. L'essentiel des démarches reposent sur une intuition quant à la ma-

son interface de validation des résultats. La préparation et la mise à disposition des corpus sur lesquels nous avons appliqué ces patrons sont de mon fait ainsi que tout le travail de validation des résultats que nous avons obtenus. Les dernières étapes d'analyse des résultats et de synthèse ont été faites en commun. Nous avons bien conscience qu'il existe d'autres travaux sur les patrons dont nous aurions pu nous inspirer – notamment pour améliorer nos résultats – mais il s'agissait là de tester avec l'existant.

nière de modéliser le domaine ou sur l'avis d'un expert. Aucune des méthodologies présentées, mis à part le système ARCHONTE et TERMINAE, ne définit de directives précises pour expliciter véritablement les concepts à l'aide du langage. L'utilisation d'ARCHONTE est une donnée de départ de mon sujet de thèse. B. Bachimont propose de contraindre l'ingénieur des connaissances à un « engagement sémantique », c'est-à-dire à expliciter clairement le sens de chacun des concepts de l'ontologie, en introduisant une « normalisation sémantique ».

> *« Les primitives nécessaires à la représentation des connaissances doivent être modélisées à partir des données empiriques dont on dispose, à savoir l'expression linguistique des connaissances. Le travail de modélisation doit s'effectuer à partir de documents attestés dans la pratique d'un domaine et rassemblés en un corpus. Le corpus est constitué de documents produits dans le contexte où le problème à résoudre se pose » (Bachimont, 2000).*

B. Bachimont considère que le corpus textuel est la source privilégiée permettant de caractériser les notions utiles à la modélisation ontologique et le contenu sémantique qui leur est associé. C'est pourquoi ARCHONTE permet de décrire les variations des sens des termes considérés en contexte.

Comme le montre la figure 4.1 reprise de (Troncy, 2004), ARCHONTE comporte initialement trois étapes : la normalisation, la formalisation et l'opérationalisation (Bachimont *et al.*, 2002). Nous détaillons ces étapes ci-dessous.

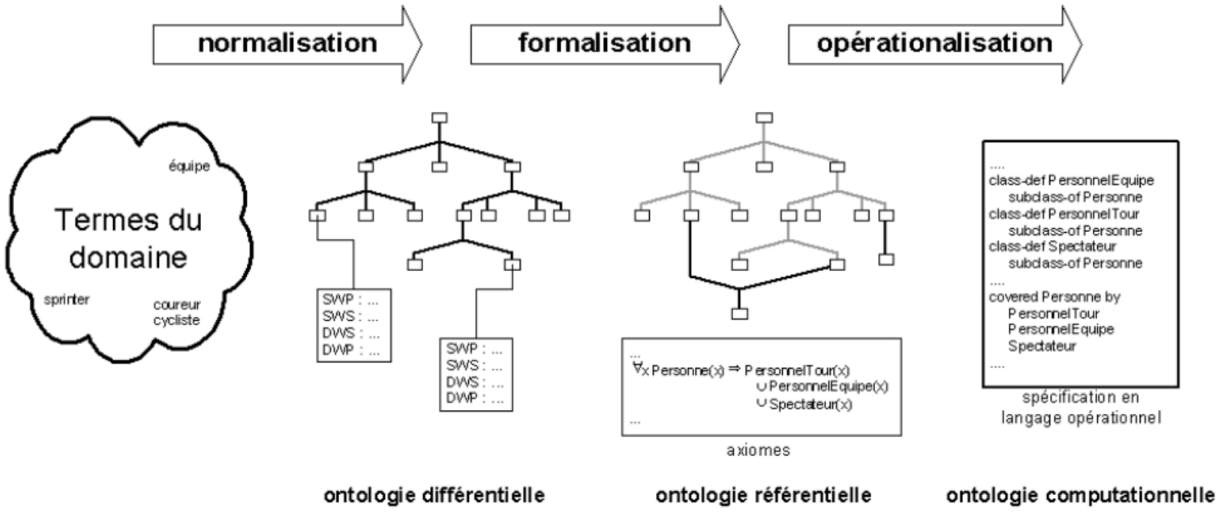


FIG. 4.1 – Les trois étapes d'ARCHONTE telles que proposées par B. Bachimont (2000).

1.1 Normalisation sémantique et engagement sémantique

L'étape de normalisation sémantique a pour objectif de rendre explicite le sens des expressions linguistiques. Il s'agit, par ce processus, d'en faire des primitives du domaine, c'est-à-dire

d'identifier les notions élémentaires à partir desquelles l'ensemble des connaissances du domaine sont construites. La langue est le média naturel de l'accès et de la diffusion des connaissances, aussi les outils d'extraction terminologique comme SYNTEX paraissent d'un usage évident. Ces outils permettent d'extraire des termes candidats dont les libellés doivent être normalisés pour être utilisés. B. Bachimont propose pour cela d'utiliser la sémantique différentielle présentée dans les travaux de F. Rastier (1994). Cette théorie attribue un sens aux termes grâce à la définition de traits sémantiques génériques et spécifiques. Ces traits permettent de fixer le cadre interprétatif, en fonction de l'objectif que s'est donné l'ingénieur des connaissances, et d'obtenir une primitive exploitable. Cela revient à associer aux termes une signification qui fasse abstraction des variations de sens liées aux différents contextes textuels dans lesquels ils peuvent apparaître. Les concepts sont donc normés puisqu'ils sont décrits selon un certain point de vue, en l'occurrence celui de la tâche à réaliser. En pratique, l'ingénieur des connaissances doit exprimer en langue naturelle les identités (traits sémantiques génériques) et les différences (traits sémantiques spécifiques en opposition les uns avec les autres) que chaque notion[2] entretient avec celles qui lui sont proches. La structuration de ces notions, en fonction des identités et des différences qu'elles partagent avec leurs notions mères et leurs notions sœurs dans un arbre, permet de passer à « l'ontologie différentielle ». B. Bachimont propose de définir quatre principes fondamentaux, les principes différentiels :

- le principe de communauté avec le père : il faut expliciter en quoi le fils est identique au père qui le subsume.
- le principe de différence avec le père : il faut expliciter en quoi le fils est différent du père qui le subsume. Puisque le fils existe, c'est donc qu'il est distinct du père.
- le principe de différence avec les frères : il faut expliciter la différence de la notion considérée avec chacune des notions sœurs car toute notion doit se distinguer de ses sœurs sinon il n'y aurait pas lieu de la définir.
- le principe de communauté avec les frères : il faut expliciter la communauté exitante entre la notion considérée et chacune des notions sœurs. Ce principe de communauté doit être différent du principe de communauté existant avec le parent. La communauté entre les notions filles doit permettre de définir des différences mutuellement exclusives entre les notions filles. B. Bachimont illustre très clairement ce principe par l'exemple suivant :

  > *« L'unité parente est « être humain », et les unités filles sont « homme » et « femme ». Ces unités partagent le fait d'être des humains. Mais cette propriété ne permet pas de définir en quoi les hommes et les femmes sont différents. On choisit alors comme principe de communauté la sexualité, où l'on peut attribuer à « homme » le trait masculin et à « femme » le trait féminin. Ces deux traits sont mutuellement exclusifs, car ce sont deux valeurs possibles d'une même propriété » (Bachimont, 2000).*

Le dernier principe justifie le troisième : il faut non seulement savoir caractériser les différences entre les notions filles mais également savoir en quoi ces notions filles sont semblables.

L'ingénieur des connaissances obtient à la fin de cette étape une taxinomie de notions. La signification de chacune s'obtient de manière compositionnelle en parcourant les identités et

[2]Une fois le processus de normalisation sémantique initié, on parle de « notion ».

les différences qui définissent l'ensemble des notions de l'arbre, allant de la plus générique (la notion racine) à la notion cible considérée. Autrement dit, la position d'un nœud dans l'arbre ontologique conditionne sa signification. Le processus de normalisation sémantique permet de passer d'un terme candidat à une notion dont le sens est invariable et, par conséquent, à une primitive représentant une connaissance du domaine à modéliser.

1.2 Formalisation des connaissances et engagement ontologique

La seconde étape d'ARCHONTE permet de formaliser les connaissances du domaine à représenter. Il s'agit de définir des concepts, et non plus des notions, selon une sémantique formelle et extensionnelle. Grâce à cela, les concepts pourront servir comme primitives dans un langage formel de représentation des connaissances. Il faut passer de la dimension linguistique et interprétative de la taxinomie à « l'ontologie référentielle » ou « l'ontologie formelle » composée de concepts dont le sens est décontextualisé. Selon la sémantique extensionnelle, les concepts sont liés à un ensemble de référents dans le monde, à un ensemble d'objets du domaine. Cette ensemble est appelé l'extension du concept. Désormais, l'ingénieur des connaissances peut mettre en œuvre des opérations ensemblistes, telles que la réunion, l'intersection …, qui vont lui permettre de composer de nouveaux sens et donc de nouveaux concepts formels. C'est ici qu'intervient l'idée d'« engagement ontologique » comme l'énonce B. Bachimont (2000) :

> *« Respecter le sens d'un concept, c'est s'engager à ce que lui corresponde une extension d'objets existants dans l'univers d'interprétation. Il s'agit donc bien d'un « engagement ontologique », puisque c'est l'existence d'objets qui est prescrite par le sens du concept. »*

La structure de la hiérarchie ontologique n'est plus alors un arbre mais un treillis puisque les extensions des concepts peuvent avoir un sous-ensemble commun. L'héritage multiple est désormais possible (*cf.* figures 4.1 et 4.2). Cette étape de la méthode permet également de formaliser les relations qui existent entre les concepts en définissant leur arité et les ensembles d'extensions de concepts qu'elles relient.

1.3 Opérationalisation

Cette dernière étape s'attache à informatiser l'ontologie référentielle dans un langage opérationnel de représentation des connaissances, adoptant le formalisme des graphes conceptuels (*cf.* chapitre 3 sous-section 3.1), ou celui des logiques de description (*cf.* chapitre 3 sous-section 3.2). En effet, un système informatique ne peut pas manipuler des concepts en fonction de leur interprétation sémantique. Il ne peut exploiter les concepts qu'en suivant les règles et les opérations qu'il peut leur associer. Ainsi, la sémantique qui permet à une machine d'utiliser des concepts réside dans la spécification informatique des opérations mathématiques que l'on peut faire sur ces concepts. Ces opérations peuvent être de plusieurs sortes en fonction du formalisme de représentation des connaissances choisi. Il s'agit ici de définir une « sémantique computationnelle » pour chaque concept de l'ontologie, c'est-à-dire que chaque concept est vu par la machine comme le

résultat d'un ensemble d'inférences et de calculs. Cette étape marque le passage à « l'ontologie computationnelle ».

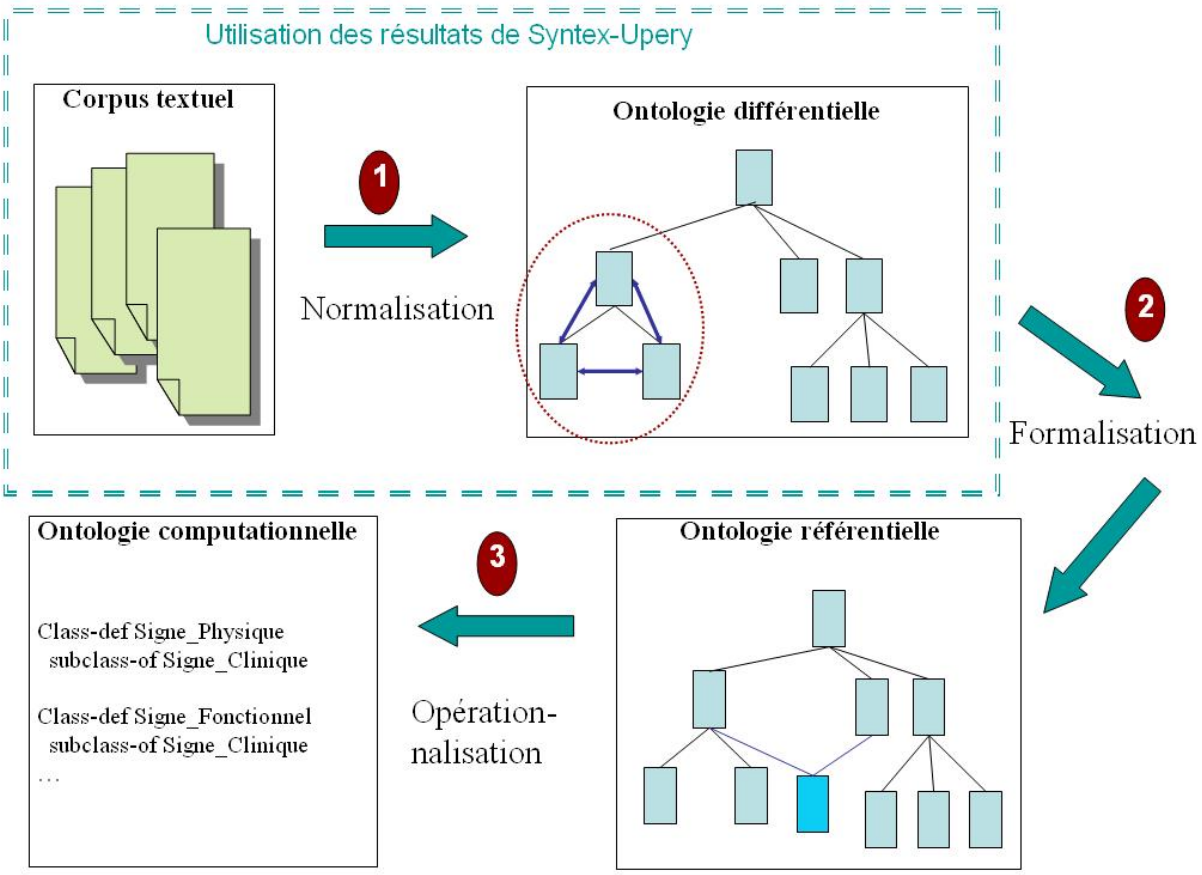


FIG. 4.2 – Une autre vue des étapes d'ARCHONTE d'après nos travaux.

Notre expérimentation dans le domaine de la pneumologie nous permet d'adapter et de préciser ces trois étapes et d'en ajouter une première consacrée à la constitution du corpus des connaissances et à son analyse par des outils de traitement automatique du langage (*cf.* section 2 et figure 4.2). En ce sens, notre méthode est proche de TERMINAE. La seule différence marquante semble concerner la partie référentielle.

2 Élaboration des corpus de référence

Dans le domaine de l'Ingénierie des connaissances, l'utilisation des corpus se veut une réponse au problème de l'accès aux connaissances d'un domaine pour un objectif particulier, lié à une application informatique. Ainsi, le corpus de référence est un ensemble de textes collectés pour être la base d'une démarche d'acquisition des connaissances. Ce corpus est utilisé par des

outils de Traitement automatique de la langue pour en extraire une majorité des termes et les relations du domaine à représenter. Les avantages d'une analyse de corpus sont nombreux :

- Les textes sont des sources de savoirs privilégiées car ils contiennent des connaissances explicites mises à disposition sur un support physique, par opposition aux connaissances des experts souvent tacites, difficilement exprimables et difficilement accessibles (Aussenac-Gilles & Condamines, 2003). Nous n'affirmons pas qu'il est souhaitable, voire même possible, de développer une modélisation de qualité sans l'intervention d'experts du domaine. Nous pensons, par contre, que leur participation au processus de construction peut être ponctuelle et limitée aux étapes de validation.
- L'analyse automatisée de corpus est un gain de temps par rapport aux entretiens avec des experts du domaine.
- Les textes sont des ressources fiables et stables puisqu'ils fixent par écrit, à un instant *t*, un certain nombre de connaissances sur le domaine.
- Les modèles sont *a priori* plus facilement compréhensibles et donc maintenables car le retour aux textes est possible.

Nous avons élaboré notre corpus de référence en collectant des documents épars préexistants dont l'objet premier n'était pas de servir de ressources de base pour l'élaboration d'une ontologie de la pneumologie. Nous avons essayé de réunir un corpus représentatif du domaine en rassemblant des comptes rendus d'hospitalisation (CRH) provenant de six hôpitaux différents de l'Assistance Publique-Hôpitaux de Paris. Nous pensons donc avoir une couverture du domaine et une représentativité de l'activité suffisantes. Il s'agit de ressources confidentielles qu'il nous a fallu anonymiser. Pour garder la cohérence interne des documents, nous n'avons pas supprimé les dates ou les noms des hôpitaux, des médecins et des patients mais nous les avons remplacées par des équivalents de notre cru. Les CRH se répartissent comme suit : Créteil : 326 CRH, Hôtel-Dieu : 97 CRH, Kremlin-Bicêtre : 125 CRH, Pitié-Salpêtrière : 57 CRH, Saint Antoine : 372 CRH, Tenon : 61 CRH. Ces 1038 CRH constituent notre premier corpus auquel nous ferons désormais référence sous le nom de corpus [CRH]. Ces ressources nous ont semblé adéquates car elles font état des pathologies de patients hospitalisés dans des services de pneumologie et notre objectif, à terme, est de représenter le domaine pour assister les pneumologues dans leur tâche de codage des pathologies. Ces CRH s'adressent à des professionnels de santé comme les infirmières, les pneumologues, ... et ne sont pas destinés à être communiqués au patient. Le niveau de langue et le vocabulaire utilisé sont donc précisément ceux que nous recherchons.

Les textes nous sont parvenus au format propriétaire Microsoft Word (*cf.* figure 4.3). Il a fallu les découper, les baliser, les anonymiser et les convertir dans un format proche de XML, compatible avec SYNTEX-UPERY (*cf.* figure 4.4). Ce premier corpus [CRH] compte environ 417 000 mots.

Nous disposons également d'un autre corpus constitué d'un livre de cours au format propriétaire Microsoft Word (*cf.* figure 4.5) (Housset, 1999). Cette ressource s'adresse initialement à des étudiants en première année de médecine. Tout comme le corpus [CRH], nous avons découpé, balisé et converti les fichiers pour créér une ressource disponible dans un format proche de XML, compatible avec SYNTEX-UPERY (*cf.* figure 4.6). Nous nous sommes appuyés sur la structure du document en chapitre, section, paragraphe ... pour réaliser le découpage. Ce second corpus, intitulé [LIVRE] compte environ à 823 000 mots.

MOTIF DE L'HOSPITALISATION

Pleurésie fébrile.

ANTECEDENTS

Médicaux : pneumonie à l'âge de 13 ans.

Familiaux : père décédé à l'âge de 68 ans d'un cancer rectal (dernière coloscopie du patient en juin 2001 normale).

MODE DE VIE

Marié, deux enfants en bonne santé.

Travail : chef de cuisine dans un restaurant universitaire.

Tabagisme : 20 PA sevré il y a un mois ;

HISTOIRE DE LA MALADIE

L'histoire remonte en fait à avril, mai 2001 avec une toux isolée productive sans altération de l'état général.

Le patient décide alors d'arrêter de fumer en octobre 2001 à cause de cette toux et de l'angoisse du cancer.

En novembre, apparaissent des douleurs thoraciques droites peu gênantes, sensibles aux antalgiques de palier I. Puis aggravation des douleurs motivant la consultation cher son médecin généraliste qui diagnostique alors une douleur d'origine musculaire. Survient alors le 14 novembre dans la nuit une douleur brutale thoracique droite augmentée par l'inspiration profonde et la toux accompagnée de fièvre. Le lendemain matin le patient retourne voir son médecin traitant qui diagnostique une névrite costale. Le 16 novembre apparaissent des frissons, une température à 38,4°C. La radiographie thoracique faite en ville à ce moment met en évidence un épanchement pleural droit. Son médecin le met alors sous Augmentin ce qui entraîne une amélioration clinique très progressive et une apyrexie. La radio de contrôle du 28 novembre montre une aggravation avec une pleurésie en voie d'enkystement. Le patient est alors adressé aux urgences.

Aux urgences : FR 16/mn, pas de sueur, pas de détresse respiratoire ni circulatoire, le patient est alors adressé dans le service.

EXAMEN CLINIQUE A L'ENTREE

Température 36,9°C (sous Diantalvic), pouls 81/mn, TA 16-9. Saturation en AA 96 %.

Auscultation pulmonaire : diminution du MV à droite, mais pas de crépitant. Pas de dyspnée, pas de douleur thoracique. Doute sur un hippocratisme digital.

Auscultation cardiaque : bruits du cœur réguliers, souffle aortique systolique non connu côté 2/6ème.

Examen buccodentaire : infection dentaire traitée par des antibiotiques et une dévitalisation en janvier 2001. Pas d'infection aigue dentaire évolutive.

Examen cutané : normal.

FIG. 4.3 – Extrait d'un CRH au format Microsoft Word.

Sachant que notre démarche consiste à utiliser des outils de Traitement automatique de la langue sur les documents constituants les corpus de référence, nous devions choisir un nombre de ressources raisonnables. Il fallait donc trouver un équilibre entre les problématiques d'exhaustivité, de représentativité et de taille du corpus, cette dernière étant directement liée à la question de la calculabilité. Les corpus de référence devaient alors présenter une densité terminologique relativement importante. Dans la droite ligne des travaux de S. Le Moigno *et al.* (2002b; 2002a), nous avons utilisé la loi de G. K. Zipf (1949). Nous présentons les résultats que nous obtenons en nous intéressant à la question de l'évaluation des corpus dans le chapitre 6, section 2.1.

Nous souhaitons construire une ontologie pour l'aide au codage. Dans cette optique, il est important de mettre à disposition du pneumologue un vocabulaire qu'il emploie couramment. Le corpus [LIVRE] est alors moins intéressant car les connaissances qu'il contient sont destinées à une personne non spécialiste du domaine et exprimées dans un souci évident de pédagogie.

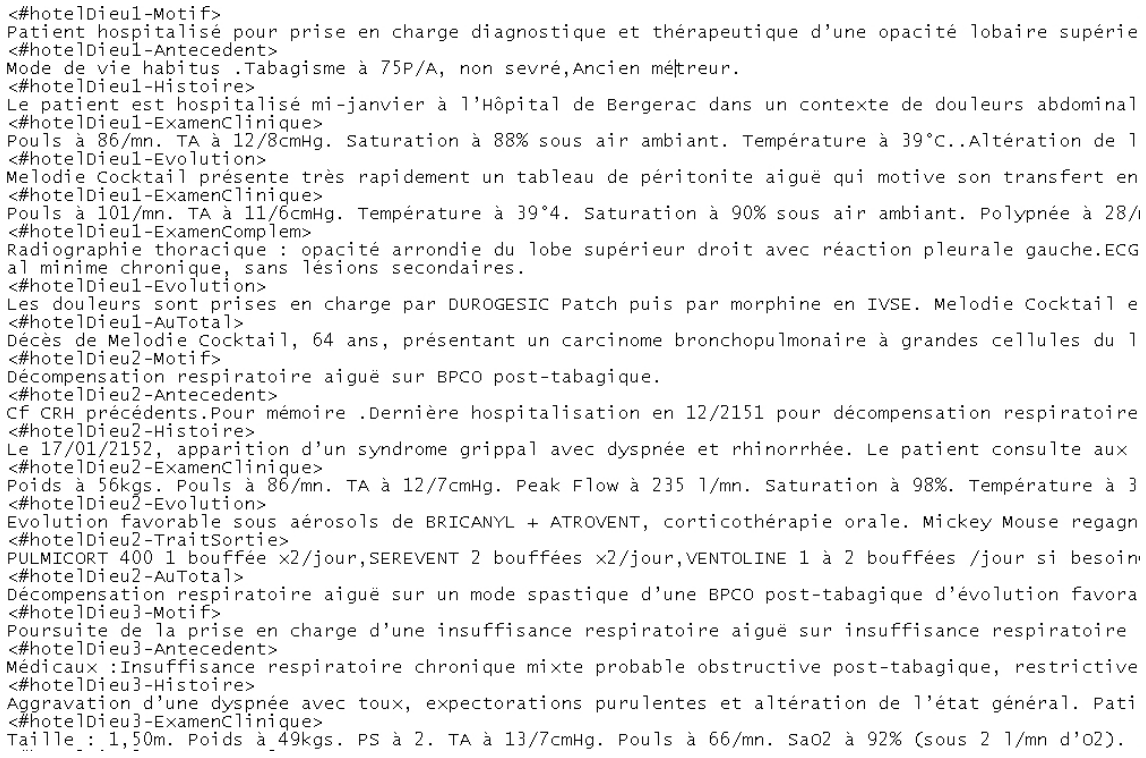
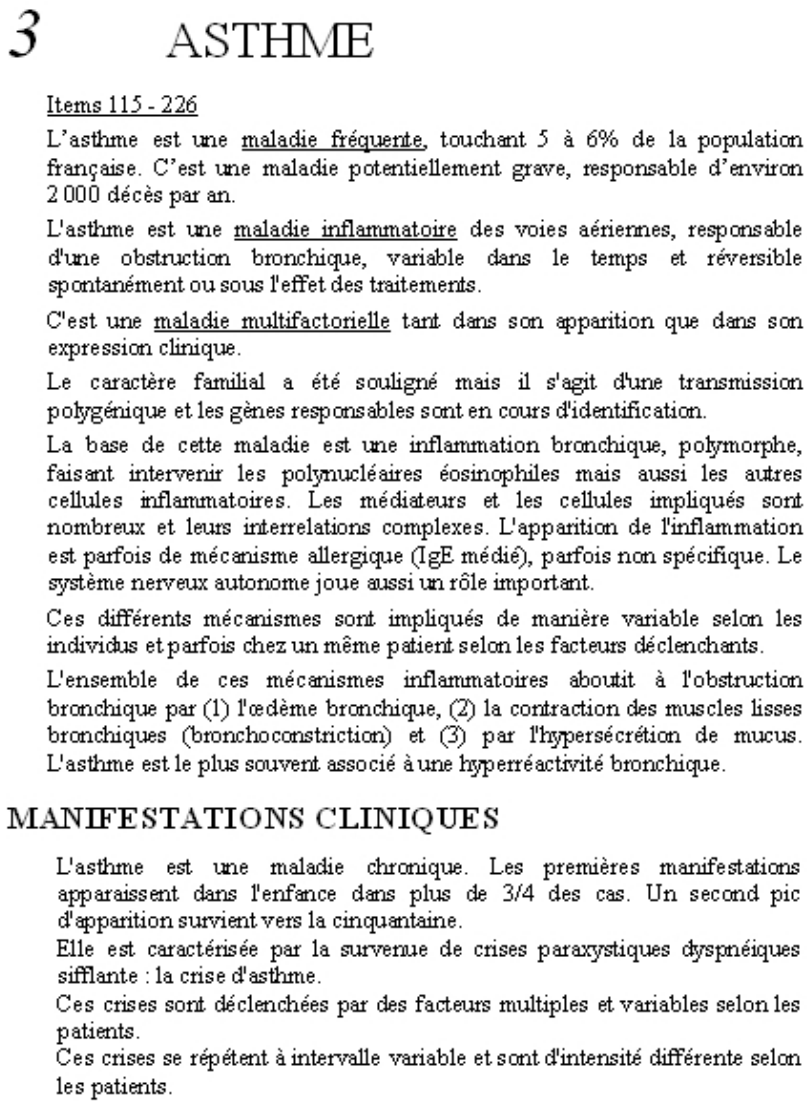


FIG. 4.4 – Extrait d'un CRH transformé dans un format compatible avec SYNTEX-UPERY.

Au contraire, les données du corpus [CRH] sont exprimées par des pneumologues dans leur vocabulaire métier et sont donc plus représentatives. C'est pourquoi nous choisissons de traiter séparément ces deux ressources qui ne paraissent pas avoir la même pertinence pour nos recherches.

3 Traitement des corpus

Le traitement des corpus revient à systématiser et à rendre plus efficace la recherche des connaissances pertinentes dans les textes en utilisant des outils de Traitement automatique de la langue. Nous traitons nos deux corpus par l'analyse syntaxique puis par l'analyse distributionnelle et par l'application de patrons lexico-syntaxiques. Nous comparons ensuite les résultats de ces deux méthodes en pensant que la compatibilité, comme la divergence, des résultats obtenus sont une aide précieuse pour l'ingénieur des connaissances, pour construire l'ontologie et particulièrement concernant l'étape de mise en œuvre des principes différentiels.



3 ASTHME

Items 115 - 226

L'asthme est une maladie fréquente, touchant 5 à 6% de la population française. C'est une maladie potentiellement grave, responsable d'environ 2 000 décès par an.

L'asthme est une maladie inflammatoire des voies aériennes, responsable d'une obstruction bronchique, variable dans le temps et réversible spontanément ou sous l'effet des traitements.

C'est une maladie multifactorielle tant dans son apparition que dans son expression clinique.

Le caractère familial a été souligné mais il s'agit d'une transmission polygénique et les gènes responsables sont en cours d'identification.

La base de cette maladie est une inflammation bronchique, polymorphe, faisant intervenir les polynucléaires éosinophiles mais aussi les autres cellules inflammatoires. Les médiateurs et les cellules impliqués sont nombreux et leurs interrelations complexes. L'apparition de l'inflammation est parfois de mécanisme allergique (IgE médié), parfois non spécifique. Le système nerveux autonome joue aussi un rôle important.

Ces différents mécanismes sont impliqués de manière variable selon les individus et parfois chez un même patient selon les facteurs déclenchants.

L'ensemble de ces mécanismes inflammatoires aboutit à l'obstruction bronchique par (1) l'œdème bronchique, (2) la contraction des muscles lisses bronchiques (bronchoconstriction) et (3) par l'hypersécrétion de mucus. L'asthme est le plus souvent associé à une hyperréactivité bronchique.

MANIFESTATIONS CLINIQUES

L'asthme est une maladie chronique. Les premières manifestations apparaissent dans l'enfance dans plus de 3/4 des cas. Un second pic d'apparition survient vers la cinquantaine.
Elle est caractérisée par la survenue de crises paraxystiques dyspnéiques sifflante : la crise d'asthme.
Ces crises sont déclenchées par des facteurs multiples et variables selon les patients.
Ces crises se répètent à intervalle variable et sont d'intensité différente selon les patients.

FIG. 4.5 – Extrait du livre de cours au format Microsoft Word, (Housset, 1999).

3.1 Approche syntaxique et distributionnelle : SYNTEX-UPERY

3.1.1 SYNTEX

Dans un premier temps, le module SYNTEX [3] extrait des corpus des syntagmes nominaux maximaux en s'arrêtant aux unités linguistiques identifiées comme étant des frontières des syntagmes. Ces frontières peuvent être une ponctuation forte, un pronom, un verbe conjugué ... Ensuite, le module acquiert par lui-même, à l'aide d'une méthode d'apprentissage endogène sur

[3] Pour une présentation complète de l'analyseur de corpus SYNTEX, le lecteur intéressé peut se reporter aux travaux de D. Bourigault *et al.* (2005).

```
<#masson3-Chap1-titre>
Asthme
<#masson3-Chap1-par>
L'asthme est une maladie fréquente, touchant 5 à 6% de la population française. C'est ur
<#masson3-Chap1-par>
L'asthme est une maladie inflammatoire des voies aériennes, responsable d'une obstructic
<#masson3-Chap1-par>
C'est une maladie multifactorielle tant dans son apparition que dans son expression clir
<#masson3-Chap1-par>
Le caractère familial a été souligné mais il s'agit d'une transmission polygénique et le
<#masson3-Chap1-par>
La base de cette maladie est une inflammation bronchique, polymorphe, faisant intervenir
<#masson3-Chap1-par>
Ces différents mécanismes sont impliqués de manière variable selon les individus et parf
<#masson3-Chap1-par>
L'ensemble de ces mécanismes inflammatoires aboutit à l'obstruction bronchique par (1) 
<#masson3-Chap1-Sec1-titre>
Manifestations cliniques
<#masson3-Chap1-Sec1-par>
L'asthme est une maladie chronique. Les premières manifestations apparaissent dans l'enf
<#masson3-Chap1-Sec1-par>
Elle est caractérisée par la survenue de crises paraxystiques dyspnéiques sifflante : la
<#masson3-Chap1-Sec1-par>
Ces crises sont déclenchées par des facteurs multiples et variables selon les patients.
<#masson3-Chap1-Sec1-par>
Ces crises se répètent à intervalle variable et sont d'intensité différente selon les pa
<#masson3-Chap1-Sec1-par>
L'asthme peut rester symptomatique toute la vie ou devenir inactif, le plus souvent à l'
<#masson3-Chap1-Sec1-par>
Les 2 principaux risques évolutifs sont l'évolution vers un asthme à dyspnée continue et
<#masson3-Chap1-Sec1-sSec1-titre>
Crise d'asthme
<#masson3-Chap1-Sec1-sSec1-par>
La crise d'asthme, volontiers vespérale ou nocturne, peut être précédée de divers sympto
<#masson3-Chap1-Sec1-sSec1-par>
une phase sèche caractérisée par une polypnée avec allongement du temps expiratoire. Cet
<#masson3-Chap1-Sec1-sSec1-par>
l'évolution se fait vers la phase humide ou catarrhale avec l'apparition d'une hypersécr
<#masson3-Chap1-Sec1-sSec2-titre>
État de mal asthmatique ou asthme aigu grave (AAG)
<#masson3-Chap1-Sec1-sSec2-par>
C'est une crise qui met en jeu le pronostic vital et constitue donc une urgence diagnost
```

FIG. 4.6 – Extrait du livre de cours transformé dans un format compatible avec SYNTEX-UPERY.

corpus, les informations de sous-catégorisation des noms et des adjectifs, propres aux corpus, nécessaires pour résoudre les cas d'ambiguïté de rattachement prépositionnel (Bourigault & Frérot, 2004). Ainsi, sur le corpus [CRH] par exemple, le module SYNTEX a pu apprendre l'information selon laquelle, le nom « exposition » est associé à la préposition « à » et peut ainsi extraire les syntagmes « exposition à l'amiante », « exposition à des polluants », « exposition à la silice cristalline », « exposition à des allergènes » ... (Bourigault & Jacquemin, 2000). Enfin, le module calcule les dépendances syntaxiques au sein des syntagmes nominaux maximaux identifiés précedemment et construit un réseau de têtes et d'expansions des candidats termes[4].

3.1.2 UPERY

UPERY [5] est un module qui met en œuvre une méthode d'analyse distributionnelle dite « étendue ». L'analyse distributionnelle est un type d'exploration de corpus fondé sur les principes

[4]Un candidat terme est un syntagme nominal composé d'une tête et d'une expansion. Par exemple, dans le syntagme nominal *Opacité dans le poumon gauche*, le terme *Opacité* est la tête du syntagme et *dans le poumon gauche* est son expansion.

[5]Pour une présentation complète de l'analyseur de corpus UPERY, le lecteur intéressé peut se reporter à (Bourigault, 2002).

de Harris (1968) et mis en œuvre notamment dans le logiciel LEXICLASS (Assadi & Bourigault, 2000) d'après les résultats fournis par le logiciel LEXTER (Bourigault, 1994a), puis dans SYNTEX-UPERY (Bourigault, 2002). Étant donné les relations de dépendances syntaxiques entre mots (ici noms, verbes et adjectifs) dans chaque énoncé d'un corpus, cette analyse pose que les mots qui ont un sens proche se caractérisent par des dépendances similaires. La mise en œuvre de cette méthode se découpe donc classiquement en deux parties : collecte des dépendances syntaxiques entre mots ou syntagmes, puis analyse de leur distribution.

L'analyseur syntaxique de corpus SYNTEX effectue l'analyse en dépendance de chacune des phrases des corpus, puis construit un réseau de mots et syntagmes, dans lequel chaque syntagme est relié à sa tête et à ses expansions. À partir de ce réseau, le module d'analyse distributionnelle UPERY construit pour chaque terme du réseau l'ensemble de ses contextes syntaxiques. Les termes et les contextes syntaxiques peuvent être simples ou complexes. Le module rapproche ensuite les termes, ainsi que les contextes syntaxiques, sur la base de mesures de proximité distributionnelle. L'analyse distributionnelle rapproche deux à deux des termes qui partagent les mêmes contextes. Elle est symétrique, en ce sens qu'elle peut rapprocher aussi les contextes, en fonction des termes qu'ils partagent. Ainsi, on dispose pour un contexte donné de l'ensemble des termes (termes ou syntagmes) qui apparaissent dans ce contexte, et pour un terme donné de l'ensemble des contextes (simples ou complexes) dans lesquels il apparaît. La productivité d'un contexte et la productivité d'un terme sont définies ainsi :

- la productivité d'un contexte est égale au nombre de termes qui apparaissent dans ce contexte ;
- la productivité d'un terme est égale au nombre de contextes dans lesquels ce terme apparaît.

UPERY utilise trois mesures différentes de la proximité :

le coefficient a
: Si l'on considère deux termes t1 et t2, alors le coefficient a est égal au nombre de contextes syntaxiques partagés par les deux termes.

le coefficient prox
: Ce coefficient sert à formaliser l'idée que si un contexte partagé par deux termes est très productif alors sa contribution au rapprochement des deux termes est a priori plus faible que celle d'un contexte peu productif

les coefficients j1 **et** j2
: La proximité entre deux termes est également caractérisée à l'aide de ces deux indices. Il s'agit du rapport entre le nombre de contextes partagés et le nombre total de contextes, soit $j1 = a/prod(t1)$ et $j2 = a/prod(t2)$.

Ce module calcule pour chaque couple de termes l'ensemble de ces coefficients. L'interface TERMONTO (*cf.* figure 4.7) ne présente à l'utilisateur que les couples dont les coefficients dépassent certains seuils. Ces seuils sont définis de façon empirique et varient en fonction d'une part de l'homogénéité et de la redondance du corpus et d'autre part du contexte dans lequel doivent être exploités les résultats de l'analyse distributionnelle.

productivité T	productivité E	nb voisins T	nb voisins E	nb var	nbdoc	freq	fsc1	fsc2	cat	terme	validité
22	30	9	27		538	**746**	741	5	SNom	voies aériennes	
83	57	11	38		585	**661**	168	493	SNom	Gaz du sang	
44	54	7	5		538	**597**	40	557	SNom	air ambiant	
41	61	25	65		460	**561**	397	164	SNom	embolie pulmonaire	
128	31	30	9		267	**554**	0	554	SNom	1 cp	
44	68	7	37		476	**515**	184	331	SNom	examen clinique	
53	66	39	88		467	**503**	345	158	SNom	insuffisance respiratoire	
98	59	67	107		426	**482**	249	233	SNom	épanchement pleural	
30	38	10	1		465	**475**	87	388	SNom	état général	
79	161	17	71		426	**473**	92	381	SNom	scanner thoracique	
39	97	8	53	1	400	**462**	200	262	SNom	radiographie de thorax	
37	88	8	43		439	**461**	131	330	SNom	radiographie thoracique	
260	37	52	12		375	**454**	0	454	SNom	cure de chimiothérapie	
66	42	20	10		373	**403**	83	320	SNom	murmure vésiculaire	
29	73	7	9		339	**403**	84	319	SNom	membres inférieurs	
46	8	18	25		332	**369**	269	100	SNom	hypertension artérielle	
21	11	3	8		350	**361**	140	221	SNom	insuffisance cardiaque	
10	4				353	**353**	5	348	SNom	lonogramme sanguin	
82	37	13	55		307	**337**	141	196	SNom	douleur thoracique	
19	19				320	**331**	23	308	SNom	Bilan hépatique	
23	29	2	36		255	**315**	298	17	SNom	cancer broncho-pulmonaire	

FIG. 4.7 – TERMONTO, l'interface d'accès aux données du logiciel SYNTEX-UPERY.

3.1.3 Résultats

Les deux corpus [CRH] et [LIVRE] ont été traités par SYNTEX-UPERY. Le corpus [CRH] a produit 36 881 syntagmes nominaux et le corpus [LIVRE] en a produit 17 666. Après étude, l'analyse distributionnelle ne donne pas de résultats satisfaisants sur le corpus [LIVRE] :

1. Les termes les plus fréquents extraits par SYNTEX-UPERY ne sont pas souvent pertinents pour construire la hiérarchie des concepts primitifs, essentiels à la représentation du domaine. Par exemple, le candidat terme *rapport de vraisemblance* a la plus forte fréquence d'apparition (177) dans le corpus. Or, il est sémantiquement pauvre pour le domaine de la pneumologie, il n'est donc pas caractéristique et ne sera pas retenu.

2. Par ailleurs, le nombre de voisins en tête et en expansion est faible : les candidats termes sont souvent sémantiquement éloignés car le corpus est faiblement redondant. Il est donc difficile pour l'ingénieur des connaissances de savoir où les placer dans la hiérarchie ontologique sur la base de leur distribution.

3.2 Repérage d'énoncés définitoires par patrons lexico-syntaxiques

Le travail présenté ci-dessous a été fait en collaboration avec V. Malaisé[6]. Nous avons essayé d'identifier dans les corpus [CRH] et [LIVRE] des définitions à l'aide de patrons lexico-syntaxiques. Ces patrons, mis au point par V. Malaisé, sont disponibles en annexe (*cf.* annexe C, figures C.1 et C.2). La recherche par patrons lexico-syntaxiques est fondée sur la définition *a priori* d'une relation sémantique, par exemple l'hyperonymie, puis sur l'observation de séquences en corpus qui véhiculent la relation souhaitée. Cette observation permet de schématiser le contexte lexical et syntaxique des unités lexicales en relation et de construire une synthèse de ce contexte sous la forme d'un patron lexico-syntaxique. Le patron est ensuite comparé aux occurrences en corpus et permet d'en extraire d'autres couples d'unités lexicales correspondant au motif spécifié. L'hypothèse est alors que ces nouvelles unités lexicales sont liées par la relation sémantique souhaitée. Les patrons lexico-syntaxiques s'appuient sur un marqueur ou pivot (une unité linguistique qui peut être un indice d'une relation lexicale, comme *entre autres* pour la relation d'hyperonymie) et sur un ensemble de contraintes que le contexte lexical ou syntaxique de ce pivot doit remplir. La méthode utilisée tient compte du contexte structurel dans lequel apparaît le terme. Par exemple, dans le cas de l'hyperonymie et du marqueur *entre autres*, il faut que la forme syntaxique corresponde au patron *DET SN*[7]*, entre autres SN*. Ce patron permet d'extraire une phrase contenant *Les méningites, entre autres pathologies ...* et de mettre en relation *méningites* et *pathologies*. Cette méthodologie a été présentée dans (Hearst, 1992) et mise en œuvre notamment dans (Morin, 1999) et (Séguéla, 2001). Les patrons lexico-syntaxiques liés à l'hyperonymie mettent en relation des couples père-fils potentiels qui sont intéressants pour la structuration hiérarchique d'une ontologie. Dans le cadre spécifique de la construction d'ontologies différentielles, nous appliquons cette technique à la recherche d'énoncés définitoires en corpus. Nous nous appuyons pour cela sur les travaux de (Rebeyrolle, 2000). Ces énoncés définitoires sont ensuite mobilisés dans des traitements visant à donner des pistes terminologiques pour la construction des principes différentiels.

3.2.1 Résultats

Nous appliquons sur le corpus [CRH] les patrons lexico-syntaxiques de recherche d'énoncés définitoires développés dans (Malaisé *et al.*, 2004). Les procédés techniques de cette opération de projection sont décrits dans (Malaisé, 2005). Le genre textuel n'étant pas adapté à la reformulation ou à l'explicitation du sens des unités lexicales[8], les programmes développés n'ont extrait que 31 phrases, ou ensembles de phrases, correspondant effectivement à des énoncés définitoires[9], sur un total de 199 extractions[10]. Il s'agit d'un résultat trop limité pour que cette méthode

[6] `http://www.few.vu.nl/~vmalaise/`

[7] Syntagme nominal (SN)

[8] Les textes sont destinés à des personnes de même degré de compétence, et traitent de leur domaine de connaissance, pouvant donc se fonder sur tout leur « aquis terminologique commun ».

[9] Parmi ces énoncés, 5 correspondent également à des paradigmes, relation que nous avons jugée intéressante dans la mesure où elle permet de proposer des « candidats co-hyponymes ».

[10] Pour un aperçu plus détaillé de la dépendance entre genre textuel et patrons lexico-syntaxiques, voir les travaux de (Condamines, 2003).

présente un réel intérêt sur ce corpus précis. Les principales erreurs sont les suivantes :

- Concernant les énoncés extraits autour du marqueur de la parenthèse, le patron N(N) supposé renvoyer l'hyperonyme du nom précédent la parenthèse est à l'origine de beaucoup de bruit. En effet, suite à un étiquetage par défaut de l'outil développé par V. Malaisé, des énoncés correspondant aux schémas suivants ont été renvoyés : N(HÔPITAL), Dr X (SPÉCIALITÉ), N(DATE), EXAMEN(REF. du dossier), MÉDICAMENT(DOSE) ...
- Concernant ceux extraits sur la base de marqueurs métalinguistiques (comme *expression* ou le verbe *définir*), les erreurs sont liées au genre des CRH, comprenant des passages comme *l'expression de mes salutations distinguées*, ou au domaine médical *définir les modalités d'une opération* ...
- Enfin, concernant les énoncés extraits à partir de marqueurs linguistiques plus génériques (*il s'agit de*, *indiquer* ...), nous remarquons trois grands types d'erreurs :
  1. Certains patrons associent un diagnostic à une pathologie, association qui est intéressante au niveau de la modélisation du domaine mais qui n'est pas directement définitoire.
  2. La structure même des CRH donne lieu à des extractions erronées car ils associent à un titre de paragraphe (comme *Evolution*) la description d'un patient, en commençant la phrase par *Il s'agit de* ... Il s'agit d'un problème de rattachement sémantique, la mention *Il s'agit de* ne se rapportant pas à l'*évolution*. En revanche, dans le corpus [LIVRE], ce patron est intéressant car il permet d'associer aux titres de section leurs descriptifs commençant par ce même marqueur.
  3. Le troisième type d'erreurs rencontré rejoint le comportement observé dans (Malaisé *et al.*, 2004) sur un corpus de diététique : il semblerait que le marqueur *indiquer* ne soit pas pertinent ou demande des contraintes spécifiques dans le domaine médical.

L'analyse des résultats soulève qu'il est, d'une part, toujours problématique de contraindre des patrons lexico-syntaxiques de peur d'induire du silence informationnel, et, d'autre part, que le fonctionnement de certains patrons est fortement lié aux différences des genres textuels.

Nous appliquons ensuite ces patrons au corpus [LIVRE]. Il s'agit d'un corpus d'enseignement, un genre textuel particulièrement propice à la découverte d'énoncés définitoires. Nos programmes ont extrait 799 phrases ou groupes de phrases, nous en avons validé 119, ce qui représente une précision de 15 %.

V. Malaisé a suivi la méthode présentée dans (Malaisé *et al.*, 2004) pour exploiter ces énoncés définitoires. Nous avons ensuite validé les groupes extraits dans une interface HTML prévue à cet effet par V. Malaisé (*cf.* figure 4.8). Il s'agit d'un formulaire permettant de modifier les groupes extraits, de valider les relations sémantiques et les énoncés jugés pertinents pour la construction d'ontologie.

Les données sont ensuite insérées automatiquement dans une base de données MYSQL avec un export au format d'ontologie OWL. Les hiérarchies ainsi créées sont visualisées dans l'éditeur d'ontologie DOE.

Les patrons utilisés dans cette expérimentation peuvent être améliorés. L'intérêt du travail présenté ci-dessus était véritablement de tenter « l'expérience » avec les ressources que nous avions alors à disposition. Les résultats étant intéressants, il faudrait compléter cette recherche.

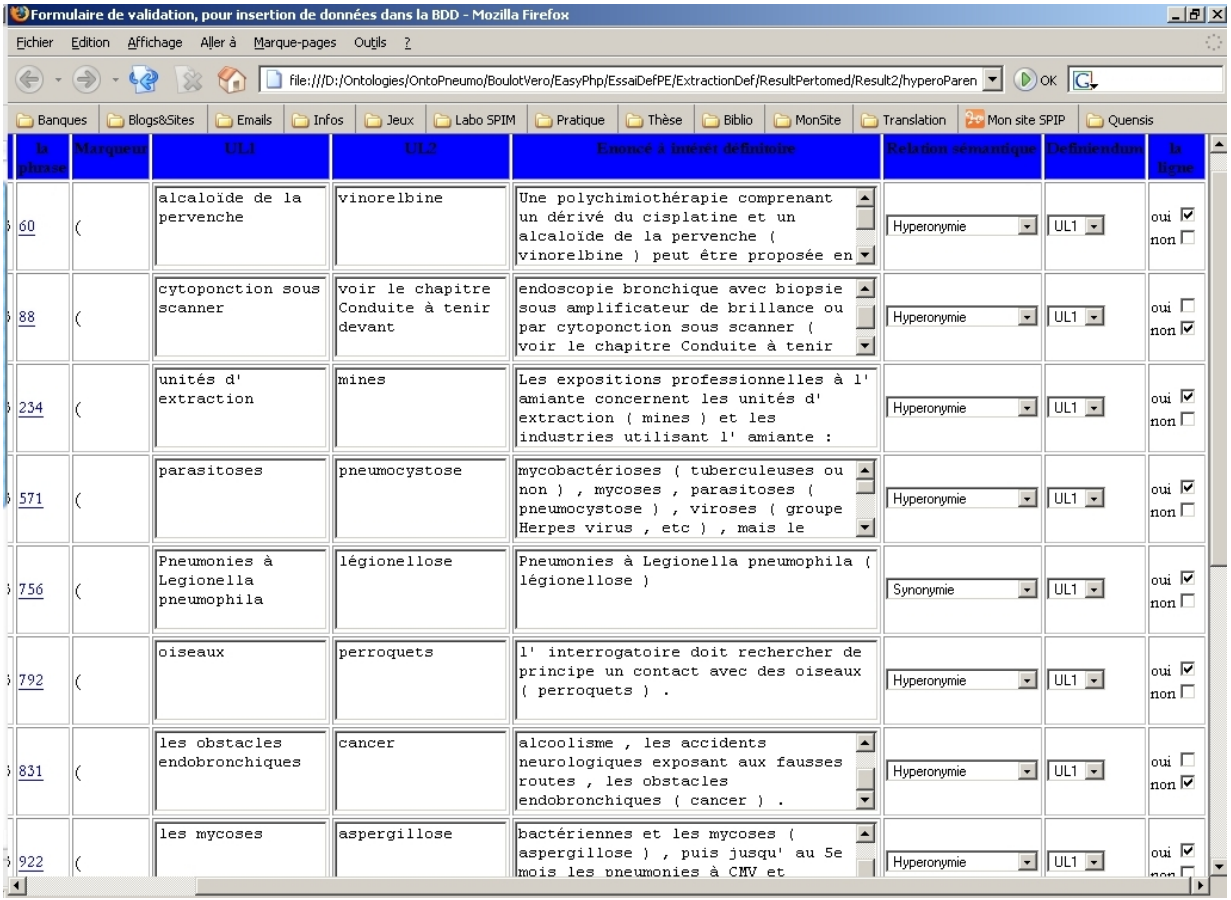


FIG. 4.8 – Interface de visualisation et de validation des extractions.

Cela fera peut-être l'objet d'une nouvelle collaboration avec V. Malaisé ou avec quelqu'un ayant des compétences semblables.

4 Sélection des candidats termes du domaine

Pour construire la hiérarchie de l'ontologie, nous avons besoin de sélectionner les briques de connaissances, à partir des candidats termes, spécifiques au domaine de la pneumologie.

4.1 Définir les termes du domaine

Il faut veiller à filtrer ou identifier les termes spécifiques du domaine de la pneumologie de ceux qui ne le sont pas car nous utilisons une technique automatique d'extraction des candidats termes. Nous nommons les termes qui nous intéressent *termes du domaine*. Ainsi, les résultats fournis par SYNTEX-UPERY sur la base du corpus [CRH] servent de support dans le choix de

candidats termes représentatifs de la pneumologie en tant qu'activité médicale. Les termes du domaine sont de plusieurs sortes :

- il peut s'agir de termes spécifiques au domaine de la pneumologie : *Pneumothorax*[11], *Chylothorax*[12], *Dyspnée*[13], *Spirométrie*[14] ...
- les termes du domaine peuvent également comporter un sens médical et être d'utilisation commune : *Fièvre*, *Palpation*, *Prélèvement*, *Diagnostic* ...
- enfin, les termes du domaine peuvent être polysèmes ayant un sens médical et un sens autre dans la langue générale : *Plaquette*, *Interne*, *Opération* ...

Les termes du domaine sont ainsi définis comme un ensemble de termes ayant au minimum un sens médical.

Le corpus de référence [LIVRE] a une structure interne telle qu'elle permet d'identifier des *termes fondamentaux du domaine*. En effet, l'importance de certains termes peut être présumée au vu de la place de choix qu'ils occupent dans la structure des douments. Ainsi, un terme, comme *Asthme* par exemple (*cf.* figure 4.5), présent dans le titre, le résumé ou le premier paragraphe d'un texte peut être identifié comme plus important que les autres car son « enjeu informationnel » n'est vraissemblablement pas le même. Nous verrons que la méthode de répérage par patrons lexico-syntaxiques utilisée sur ce corpus donne des résultats intéressants, notamment parce qu'elle tient compte du contexte structurel dans lequel apparaît le terme.

4.2 Extraction, filtrage et sélection

Comme le montre le tableau 4.1, les syntagmes extraits des corpus constituent un ensemble très hétérogène. Par conséquent, nous avons choisi de n'examiner que les noms et les syntagmes nominaux car nous pensons qu'ils véhiculent les connaissances nécessaires à l'identification des objets du domaine. Nous choisissons donc d'ignorer, pour l'instant, la sémantique portée par les autres catégories grammaticales.

SYNTEX extrait 54 547 syntagmes nominaux sur l'ensemble de nos deux corpus de référence. La quantité, sans être énorme, est suffisamment conséquente pour qu'un premier tri s'impose. Ainsi, nous allons éliminer un certain nombre de syntagmes inappropriés :

1. les syntagmes contenant des chiffres : *1 cp x3*, *Température à 37 °*, *4ème cure* ...
2. les syntagmes qui contiennent des caractères qui ne font pas partie de l'alphabet : *air à %*, *Saturation à 97 %*, *météorisme +++* ...

[11] Il s'agit de la présence d'air ou de gaz dans la cavité pleurale qui peut être spontanée ou provoquée artificiellement.

[12] Il s'agit d'un épanchement de chyle dans la cavité pleurale, à la suite de la perforation traumatique ou tumorale du canal thoracique.

[13] Il s'agit de la difficulté de respirer accompagnée d'une sensation d'oppression ou de gêne.

[14] La spirométrie est une méthode servant à mesurer la fonction ou la capacité pulmonaire et à la comparer à la fonction pulmonaire moyenne d'une personne de race, de taille, de poids et d'âge identiques. Sur base de cette comparaison, on déterminera si le patient présente une affection pulmonaire et de quel type d'affection il est question.

Abréviation	Catégorie	Exemple
Adj	Adjectif	*veineux*
Adv	Adverbe	*totalement*
Ppa	Participe passé	*diagnostiqué*
NomPr	Nom propre	*TAXOTERE*
Nom	Nom	*dilatation*
V	Verbe	*hospitaliser*
SAdj	Syntagme adjectival	*favorable sous antibiothérapie*
SNom	Syntagme nominal	*auscultation pulmonaire*
SPpa	Syntagme participial	*observé à l'échographie*
SV	Syntagme verbal	*débuter un traitement*

TAB. 4.1 – Exemples de syntagmes extraits par SYNTEX sur le corpus [CRH].

3. les termes vides : *point de vue*, *date*, *probabilité* ... Ce sont des termes qui apparaissent dans les corpus de référence avec une certaine fréquence et qui ne sont pas des termes du domaine. Ils n'apportent pas d'informations intéressantes pour la modélisation.
4. les syntagmes contenant plusieurs lettres en majuscules sont examinés manuellement avec soin : les syntagmes comme par exemple *Unité du Service de Médecine Interne* ne nous intéressent pas, par contre, nous souhaitons garder les noms de médicaments comme dans *cure de TAXOTERE* et les abréviations comme dans *mesure de la CPT*.

Nous obtenons une liste réduite de syntagmes nominaux. Des expérimentations ont été menées dans (Lame, 2002) pour différencier les syntagmes nominaux susceptibles d'être des termes du domaine de ceux qui ne le seraient pas. Ces expériences reposent sur la mise en œuvre de techniques statistiques fondées sur les indices classiques de pondération de termes. Comme le dit G. Lame, ces expérimentations se sont soldées par des échecs. On peut ainsi en conclure que l'identification des termes du domaine doit se faire manuellement.

En pratique, nous distinguons deux étapes dans la sélection des candidats termes du domaine. Nous étudions les sorties de SYNTEX-UPERY sur le corpus [CRH]. Le travail est manuel et s'appuie sur les fonctionnalités de TERMONTO.

1. Nous parcourons l'ensemble des résultats fournis par l'analyse syntaxique et choisissons d'étudier, en premier lieu, les syntagmes nominaux dont la fréquence d'apparition en corpus est supérieure à 12 (2 % du corpus). Nous repérons les grands axes conceptuels typiques du corpus et donc du domaine représenté. À chaque candidat terme, nous associons un critère de validité qui agit comme un filtre dans TERMONTO. Ce critère, comme le montre la figure 4.7, est compris dans un intervalle allant de 1 à 6, correspondant à l'un des axes conceptuels :
   – 1 : permet de regrouper les candidats termes non pertinents appartenant à l'axe *Autres*,
   – 2 : est réservé aux candidats termes déjà modelisés dans l'ontologie,
   – 3 : permet de regrouper les candidats termes appartenant à l'axe *Signes/Symptômes*,
   – 4 : permet de regrouper les candidats termes appartenant à l'axe *Pathologies*,
   – 5 : permet de regrouper les candidats termes appartenant à l'axe *Qualificatif*,

Descendants en tête	Descendants en expansion	Voisins en tête	Voisins en expansion
Épanchement pleural droit	Lame d'épanchement pleural	Lésions	Liquide
Épanchement pleural liquidien	Récidive d'épanchement pleural	Infection	Infiltrats
Épanchement pleural de la grande cavité	Lier la dyspnée à l'épanchement pleural	Signes	Décompensation

TAB. 4.2 – Exemple de résultats du rapprochement contextuel pour le syntagme nominal *Epanchement pleural*.

– 6 : permet de regrouper les candidats termes appartenant à l'axe *Traitements/Examens*. Par exemple, nous fixons à 6 le critère de validité pour tous les candidats termes de ce dernier axe - *e.g. examen*, *doppler*, *radiographie*, *etc*. Au début du processus, tous les candidats termes ont un critère de validité égale à 1 et à la fin égale à 2 car ils sont, en principe, tous définis dans l'ontologie. Les critères de validité 3, 4, 5 et 6 sont utilisés temporairement durant la phase de construction. Ces regroupements permettent une première phase de travail sur les rapprochements des candidats termes par contexte. La sélection par critère de validité laisse 35 % des candidats termes sur lesquels élaborer le cœur de notre ontologie.

2. Rappelons que l'analyse distributionnelle rapproche deux à deux les termes partageant les mêmes contextes (descendants en tête et en expansion). Comme cette analyse est symétrique, elle rapproche également les contextes en fonction des termes qu'ils partagent (voisins en tête et en expansion). Dans le tableau 4.2 *épanchement* est la tête du syntagme nominal *épanchement pleural* et *pleural* est son expansion. Les descendants en tête donnent des informations sur ce qui pourraient être des concepts fils ou des concepts définis. Les descendants en expansion donnent des informations sur la place du concept dans la hiérarchie, sur le concept père. Les voisins en tête et en expansion nous permettent de constituer des regroupements de candidats termes sémantiquement proches du candidat terme étudié. Ces regroupements sont d'une grande aide pour élaborer la structure hiérarchique de l'ontologie, aussi bien l'axe horizontal que vertical. L'exemple suivant montre un premier rapprochement possible : nous pouvons mettre en rapport le groupe A *{épanchement, lésion, infection, décompensation}* avec *{signes}*. Les candidats termes du groupe A partagent un même contexte sémantique, la première hypothèse est donc qu'il peut s'agir de concepts frères dont *signes* est possiblement le concept père.

5 Mise en œuvre des principes différentiels

La méthodologie de construction d'ontologies différentielles utilisée est constructive, elle permet de placer de manière précise chaque concept dans la structure hiérarchique. Donc, pour

élaborer cette hiérarchie, il convient d'articuler les candidats termes choisis dans la précédente étape en précisant les principes différentiels qui les définissent. Par exemple, le concept *Ultrasonographie* et le concept *ExamenIsotopique* sont des concepts frères dont le concept père est *ImagerieParRayonnement* (*cf.* figure 4.9). Le principe de communauté avec le concept père est la projection ou l'injection d'une substance artificielle dans le but d'effectuer des mesures. Le principe de communauté entre les concepts frères est lié au média d'injection. Le principe différentiel entre les concepts frères est relatif au type de média artificiel mis en œuvre : un isotope dans le cas de l'*examen isotopique* (la scintigraphie est un exemple d'examen isotopique) et les ultrasons pour l'*ultrasonographie*. L'ensemble des candidats termes regroupés sous les axes conceptuels présentés à la section 4.2, est défini selon ces principes. Nous pensons que la com-

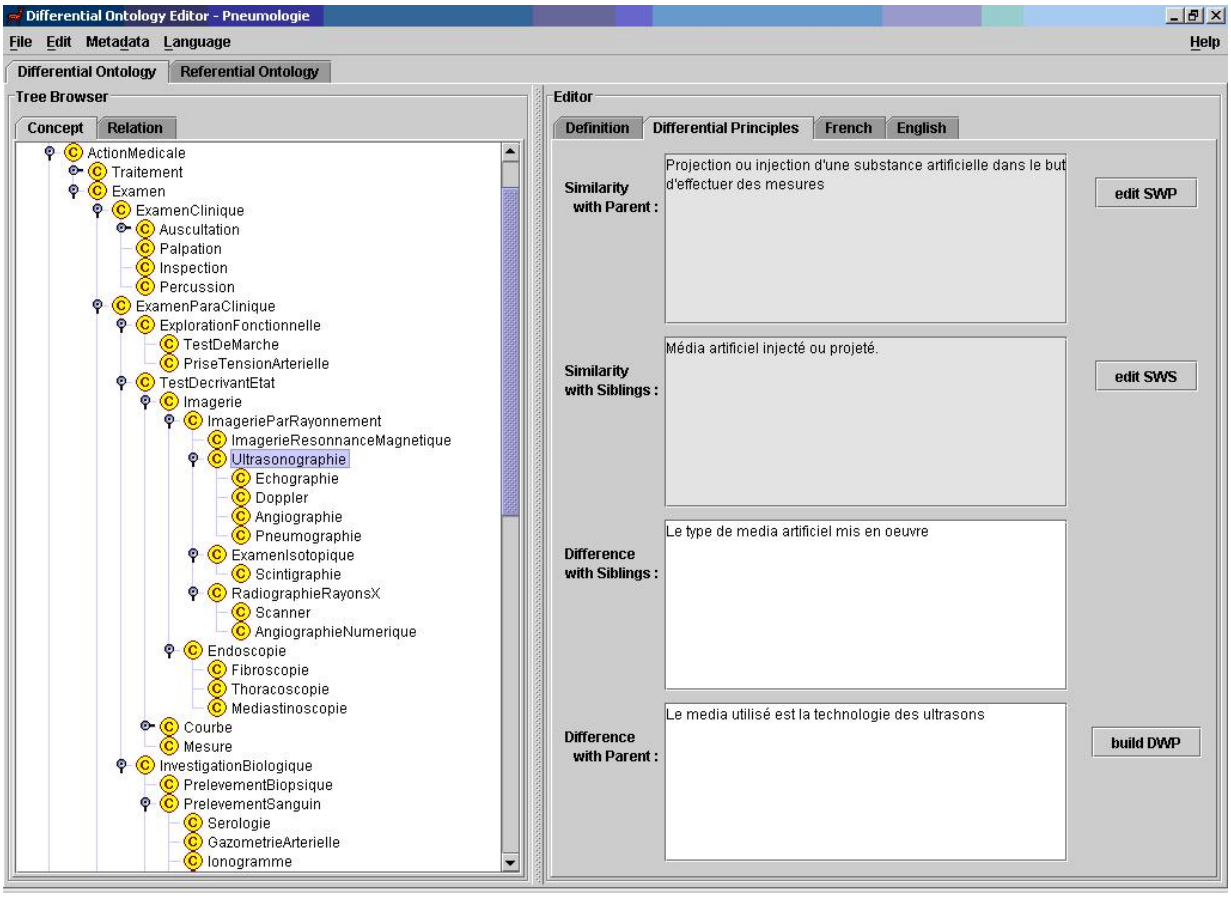


FIG. 4.9 – Extrait d'OntoPneumo sous DOE.

paraison des résultats obtenus (1) par l'analyse distributionnelle sur le corpus [CRH] et (2) par le repérage d'énoncés définitoires à l'aide de patrons lexico-syntaxiques sur le corpus [LIVRE], va enrichir la démarche cognitive de l'ingénieur des connaissances dans la phase de mise en œuvre des principes différentiels.

Type	Nb	Exemple
Termes identiques, à la normalisation terminologique près	3	*Asthme*[CRH] ***vs*** *asthme*[LIVRE], *Emphyseme*[CRH] ***vs*** *emphysème*[LIVRE]
Variantes lexicales comparables	3	*SaturationEnAir*[CRH] ***vs*** *saturation en oxygène*[LIVRE]
Niveaux de granularité différents	18	*Adenopathie*[CRH] ***vs*** *Adénopathies médiastinales* [LIVRE]

TAB. 4.3 – Comparaison des termes issus des deux hiérarchies terminologiques.

5.1 Procédure de comparaison des hiérarchies obtenues

Nous avons comparé les terminologies structurées suivant les deux méthodes précédentes pour voir dans quelle mesure elles sont compatibles ou complémentaires. Pour cela, nous comparons tout d'abord manuellement les 370 termes de la future ontologie (arborescence [CRH]) avec la structuration à un niveau de profondeur issue de la validation des extractions d'énoncés définitoires (arborescence [LIVRE]), comprenant 119 candidats termes ou groupes syntaxiques plus larges. Nous trouvons 24 ensembles de termes, ou termes, comparables à différents titres. Par exemple, au concept de *Tumeur* issu de la terminologie [CRH] correspond un ensemble de différentes tumeurs dans la terminologie [LIVRE]. Ces variantes sont détaillées dans le tableau 4.3.

Nous comparons ensuite les structures autour de ces termes communs ou similaires. Là encore, il y a plusieurs cas de figure. En effet, lorsque la structuration terminologique liée à la validation du formulaire correspond à une hiérarchie (c'est-à-dire lorsque l'énoncé définitoire ne met pas en relation deux « synonymes » au sens large), nous observons que les deux structures peuvent être identiques, complémentaires ou divergentes (*cf.* tableau 4.4).

Les derniers exemples du tableau 4.4, concernant les hiérarchies divergentes, montrent l'intérêt d'avoir deux hiérarchies à confronter pour pré-valider les données issues d'extraction à partir de corpus avant de les proposer aux experts. Dans la majorité des cas, les hiérarchies sont soit équivalentes, soit complémentaires. En effet, nous n'avons identifié que quatre contre-exemples sur 24 termes, incluant deux cas où la hiérarchie [LIVRE] n'est pas juste puisqu'elle associe un terme à ses caractéristiques et non pas à son hyperonyme. Nous constatons que les candidats termes hiérarchisés à partir du corpus [LIVRE] sont souvent plus spécifiques que les candidats termes issus du corpus [CRH]. Cette différence peut être due au fait que les termes de base ne sont pas définis dans le livre de cours à l'origine du corpus [LIVRE] car ils correspondent à des notions supposées acquises. Elle peut également être due au mode d'extraction de ces termes et marquer une des spécificités de l'extraction par patrons lexico-syntaxiques.

Enfin, nos deux dernières expérimentations concernent les « termes spécifiques » des deux analyses terminologiques : nous avons cherché à comprendre pourquoi certains termes ne se retrouvent pas dans les deux arborescences.

Type	Exemple	Commentaire
Identique ou comparable	*Broncho pneumopathie* / *Asthme* [CRH] ***vs*** *Bronchopathie* / *Asthme à dyspnée continue* [LIVRE]	Pour comparer ces deux hiérarchies, nous avons regardé comment ces trois notions étaient organisées dans le MeSH. Les deux premières sont classifiées sous *Poumon, maladie*, alors qu'*Asthme* est une notion plus spécifique dans la même branche hiérarchique : ce qui tend à valider la cohérence et la compatibilité des deux hiérarchies terminologiques trouvées.
Complémentaire	*Signe* / [...] / *Signe-Respiratoire* / *Insuffisance-Ventriculaire* [CRH] ***vs*** *Signe* / *Insuffisance ventriculaire droite* [LIVRE]	La deuxième arborescence vient confirmer la première et permet de la compléter d'un niveau, celui de *Insuffisance ventriculaire droite*.
Divergente	*EtatPathologique* / *Maladie-Respiratoire* / *Bronchite* ET *Signe* / *Toux* [CRH] ***vs*** *Toux avec expectoration* / *Bronchite chronique* [LIVRE]	Dans le MeSH, la toux est classifiée à la fois comme *signe-symptôme* et comme *pathologie* : les deux sources textuelles illustrent chacune un de ces aspects. Nous avons choisi de privilégier le point de vue abordé dans le corpus [CRH].
	EtatMorphologique / *AnomalieMorphologique* / *Lesion* / *Atelectasie – Adenopathie* [CRH] ***vs*** *Opacité médiastinale* / *Adénopathies médiastinales* ET *Opacité dense arrondie* / *Atéléctasie par enroulement* ... [LIVRE]	Dans les deux cas, *Atélectasie* et *Adénopathie* sont cohyponymes, mais leurs hyperonymes sont contradictoires : dans l'arborescence [CRH], les *opacités* sont classifiées sous *Signes*. Il s'agit d'une erreur d'interprétation de la relation sémantique dans les extractions à partir du corpus [LIVRE]. Une *opacité* est bien un élément d'une image médicale qui est interprétée comme le *signe* d'une *adénopathie*. Cependant, l'extrait du corpus était ambigu : *« Adénopathies médiastinales. Il s'agit des opacités médiastinales les plus fréquentes ... »*

TAB. 4.4 – Comparaison des résultats des deux analyses terminologiques.

5.2 Comparaison des termes issus du corpus [LIVRE]

L'automatisation du processus de comparaison des deux hiérarchies a permis d'isoler les termes qui n'apparaissent que dans la structure terminologique issue du corpus [LIVRE]. Nous avons cherché à comprendre pourquoi ces termes ne se retrouvent pas dans la structure issue du corpus [CRH]. Dans cette optique, nous avons étudié la présence d'une ou plusieurs occurrences dans le corpus [CRH]. L'analyse distributionnelle aurait dû les isoler. Il y a 83 termes « spécifiques »[15]), à la terminologie issue du corpus [LIVRE]. Les résultats de l'observation de leurs occurrences dans le corpus [CRH] sont présentés dans le tableau 4.5.

[15]Il y a 83 termes « spécifiques » et non 95 (119 termes - 24 termes communs). Le décompte des 24 termes communs regroupe des termes composés comparables, comme nous l'avons vu plus haut.

Type	Nb
Termes non présents dans le corpus [CRH], ou sous une forme non identifiée	20
Termes présents sous la même forme dans [CRH] mais à faible nombre d'occurrences	15
Termes correspondant à une même racine dans [CRH] (*spondylarthrite* [LIVRE] *vs spondylarthropathie* [CRH]), ou présence de la tête du syntagme complexe (*mésothéliome malin diffus* [LIVRE] *vs mésothéliome ou mésothéliome pleural* [CRH])	42
Termes correspondant à des énoncés définitoires, validés comme paraphrases synonymiques ou par erreur en tant qu'hyperonymes, et n'ayant donc aucun équivalent terminologique dans [CRH]	6

TAB. 4.5 – Comparaison des termes spécifiques à la terminologie construite à partir du corpus [LIVRE] avec le corpus [CRH].

Les termes qui ne figurent pas dans le corpus [CRH] ne peuvent pas être proposés comme candidats termes par l'analyse distributionnelle. Ceux qui sont présents sous la même forme en corpus n'ont pas un nombre d'occurrences supérieur ou égal à 12 et ne sont donc pas encore pris en compte dans notre analyse des résultats de SYNTEX. Enfin, la moitié des termes extraits par les patrons lexico-syntaxiques ont des « homologues » ou termes proches dans le corpus [CRH]. Pour les rapprocher, il faut disposer de connaissances morphologiques sur la dérivation et la composition que nous n'avons pas mises en œuvre dans cette expérimentation (Namer & Zweigenbaum, 2004).

5.3 Comparaison des termes issus du corpus [CRH]

Nous avons ensuite extrait les termes qui ne sont présents que dans l'arborescence [CRH] et nous avons regardé s'ils étaient définis ou caractérisés dans le corpus [LIVRE]. Cette comparaison étant manuelle, nous sommes passés outre la normalisation terminologique des termes[16] [CRH] et vérifions également la présence d'éventuelles variantes terminologiques. Le résultat de la comparaison est détaillé dans le tableau 4.6.

68 termes correspondent à des énoncés pouvant être interprétés comme définitoires ; nous avons voulu comprendre pourquoi les patrons lexico-syntaxiques ne les ont pas renvoyés. L'analyse de ces énoncés nous a donné plusieurs explications présentées dans le tableau 4.7.

Nous détaillons ensuite l'analyse des énoncés à intérêt définitoire semblant pouvoir être extraits au moyen de patrons lexico-syntaxiques (sachant que les deux autres types de contextes ne peuvent pas être trouvés par ce genre de méthode, mais demandent plutôt des solutions apparentées à la résolution d'anaphore), afin de savoir s'il faut augmenter le système en développant de nouveaux patrons, en adapter certains ou procéder par relâchement de contraintes (*cf.* tableau 4.8). 23 des 47 énoncés analysés sont extraits par le système, mais ne sont pas validés

[16]Nous rappelons que, conformément aux principes d'ARCHONTE, nous distinguons les labels des concepts des concepts eux-mêmes. Les labels sont les termes qui servent à désigner les concepts dans le langage. Dans notre ontologie, pour des raisons informatiques ensuite parce que ça les différentie des termes préférés, ils sont désaccentués, mis au singulier et chacun des termes commence par une majuscule. Ainsi, par exemple, le label *ElementAnatomique*

Type	Nb
Termes définis, classifiés ou caractérisés, plus ou moins précisément dans le corpus [LIVRE]	68
Termes sans occurrence dans le corpus [LIVRE]	60
Termes non définis ou caractérisés dans le corpus [LIVRE]	49
Termes de haut niveau (comme *Inanime*, *SigneFonctionnel*, ...), ne correspondant pas forcément à une occurrence dans le corpus [CRH]	48
Termes exprimant une caractéristique : des qualificatifs (*Gauche*, *Positif*, ...)	21

TAB. 4.6 – Comparaison des termes spécifiques à l'ontologie basée sur les corpus [CRH] et [LIVRE].

Type	Nb d'énoncés
Contextes sur plusieurs phrases pouvant être interprétés comme donnant des éléments de définition, mais étant relativement vagues	11
Contextes étant des définitions plus claires, mais ne pouvant pas être retrouvés au moyen de patrons lexico-syntaxiques	9
Contextes étant des définitions et pouvant, ou semblant pouvoir, être extraits au moyen de patrons lexico-syntaxiques	47

TAB. 4.7 – Évaluation des énoncés à intérêt définitoire ou assimilés du corpus [LIVRE] non renvoyés par nos patrons.

Type de patron	Exemple
Des patrons envisagés mais pas encore implémentés	Liste, virgule, double points
Des patrons classiques pas encore implémentés	de NOM tel que le NOM, des NOM et d'autres NOM
Des patrons implémentés, mais pour lesquels il faudrait pousser l'analyse, ou voir si des modifications sont envisageables	parenthèse, ou/et
Des patrons correspondant à la méronymie	comportent, consistent en
Des patrons correspondant à des relations spécifiques à la médecine	Le traitement des formes cryptogéniques *repose sur* la corticothérapie ...

TAB. 4.8 – Évaluation des énoncés à intérêt définitoire pouvant être renvoyés par des patrons lexico-syntaxiques.

lors de l'analyse des réponses. La principale raison est qu'ils donnent une définition partielle ou associent un terme à un hyperonyme inattendu, jugé non pertinent lors de la validation des formulaires. Les autres cas de figure nous donnent des pistes pour compléter les patrons existants.

6 Ontologie de haut niveau

Le travail décrit jusque ici nous permet de construire une ontologie de domaine, c'est-à-dire une ontologie des concepts de la pneumologie tels qu'ils sont manipulés par les médecins, dans le cadre de l'enseignement – Abrégé de pneumologie (Housset, 1999) – ou de leur activité de soin – CRH.

À ce stade intermédiaire, l'ontologie est bien ce qu'on appelle une ontologie de domaine et les concepts supérieurs sont soit des concepts manifestement du domaine – *e.g. Signe, RoleEnMilieuHospitalier, ModeAdministration, ReflexionMedicale, ConditionExamen, …* – soit des concepts plus généraux, ayant une signification ou des instantiations particulières dans le domaine médical – *e.g. ObjetIndenombrable, EtatObjetPhysique, ObjetNaturel, ….* Toujours à ce stade, les relations, compléments des concepts, ont été définies, sans que leur champ d'application – l'arité en logique de description – n'ait été précisée.

Une ontologie de haut niveau, appelée *top-Ontologie*, vocable que nous conserverons par la suite, est alors nécessaire pour organiser les concepts médicaux les uns par rapport aux autres. Une conséquence indispensable et intéressante de cette démarche est de définir les relations entre les concepts à ce niveau-là (par exemple [abstract-object]-(has-view-point)->[meta-abstract-object]), pour que leur applicabilité – leur *range* – se propage aux niveaux inférieurs.

La top-ontologie de MENELAS a 800 concepts et 300 relations. Que ce soit pour les concepts ou les relations, tout n'est pas utile à notre modélisation de la pneumologie car le système MENELAS (*cf.* section 2.10) tentait de prendre en compte une modélisation plus complexe que notre travail au sein de la pneumologie. Mais l'hypothèse que nous voulons tester ici est qu'il y a possibilité de définir une top-ontologie de la médecine. En effet, si nous pensons, que les travaux visant à définir une top-ontologie générale à l'ensemble des domaines du monde est vouée à l'échec, nous voudrions vérifier si cela est possible en réduisant la généricité à la seule médecine alors que chaque spécialité médicale a un point de vue spécifique sur l'ensemble du corps humain. Ces problèmes sont discutés en perspectives (*cf.* chapitre 7).

7 Formalisation et opérationalisation : PROTÉGÉ 3.2

L'ontologie que nous avons obtenu à ce stade est une ontologie référentielle rassemblant des concepts et des termes du domaine (le « terme préféré » en français et en anglais, les synonymes), des connaissances de type encyclopédique sur le domaine, ainsi que des relations sémantiques bien connues comme *est_un*, *est_une_partie_de*, *est_caracterise_par*, *est_pratique_par* … Cette ontologie a été traduite en HTML et peut être consultée à l'adresse web suivante `http://baneyx.net/OntoPneumoHTML/index.html`.

désigne le concept d'*élément anatomique*.

Les étapes de formalisation et d'opérationalisation se sont faites à l'aide du logiciel PROTÉGÉ qui offre notamment la possibilité de représenter graphiquement l'ontologie (*cf.* figure 4.12). Concrètement, nous avons utilisé l'export OWL de DOE , compatible avec le format d'import OWL de PROTÉGÉ (Troncy *et al.*, 2003). Cependant, cette compatibilité est pour l'instant limitée. Les seules informations conservées d'un logiciel à l'autre sont la définition en langage naturel des principes différentiels et le label du concept. Aussi, nous avons développé un certain nombre de programmes annexes qui permettent de récupérer et de réinjecter les informations manquantes.

L'étape de formalisation permet d'introduire des axiomes logiques qui définissent le comportement des individus qui constituent les extensions des concepts formels. Comme le montre la figure 4.11, nous rajoutons des définitions formelles aux concepts (les concepts sont en noir dans la figure) en précisant les individus des concepts telle que la liste des pneumologues pour le concept *Pneumologue*, la liste des êtres humains pour le concept *EtreHumain* ... La liste des pneumologues (les individus sont en bleu dans la figure) recoupe celle des êtres humains. Ainsi, la formalisation permet de créer un nouveau concept formel *PersonnePneumologue* (les concepts formels définis sont en rouge dans la figure) défini à l'intersection de ses deux concepts formels pères *EtreHumain* et *Pneumologue*. Comme l'explique B. Bachimont (2000), les concepts formels primitifs ont une sémantique référentielle déterminée par les engagements sémantiques et ontologique tandis que les concepts formels définis sont uniquement déterminés en fonction d'un engagement ontologique. L'ajout de nouveaux concepts à ce stade de développement modifie la structure hiérarchique car on passe d'une arborescence fondée sur des relations de similarités et de différences à une arborescence fondée sur une logique d'inclusion ensembliste. Dans l'ontologie formelle, les concepts qui n'entretiennent pas de relation de subsomption (couple père-fils) s'excluent mutuellement (*cf.* figure 4.10). Dans l'ontologie référentielle, les concepts peuvent admettre des extensions qui ont un sous-ensemble commun. L'héritage multiple est donc possible et la structure construite n'est plus un arbre mais un treillis (*cf.* figure 4.11).

Il reste à opérationaliser OntoPneumo. Il s'agit de traduire l'ontologie obtenue à l'étape précédente en une ontologie destinée à servir dans un système informatique. Pour cela, elle doit être spécifiée en un langage de représentation des connaissances doté de capacité d'inférence. Nous avons choisi d'utiliser le langage OWL qui répond parfaitement à nos besoins en termes d'expressivité et de maniabilité.

8 Synthèse sur la construction de la hiérarchie

Notre méthodologie est fondée sur l'utilisation de corpus textuels comme source des connaissances du domaine. Nous disposons de deux corpus de types différents sur lesquels nous appliquons deux techniques relevant du Traitement automatique des langues : l'analyse syntaxique suivie de l'analyse distributionnelle sur le corpus de comptes rendus d'hospitalisation et la recherche d'énoncés définitoires par projection de patrons lexico-syntaxiques sur le second corpus constitué d'un livre de cours de pneumologie. Nous utilisons les candidats termes issus de l'analyse distributionnelle sur le corpus [CRH] comme base pour construire la hiérarchie de l'ontologie. Les regroupements de ces candidats termes sous de grands axes conceptuels nous permettent très rapidement de développer des micro-structures hiérarchiques. Ainsi, nous n'adoptons pas vé-

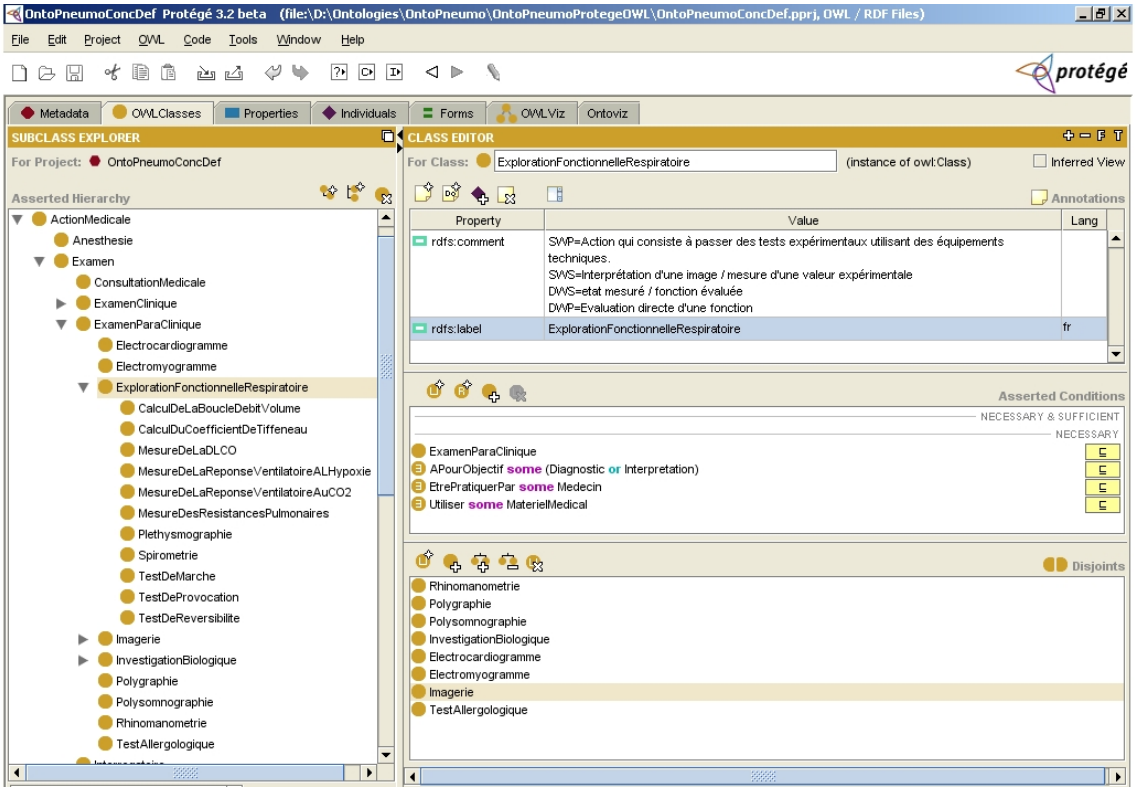


FIG. 4.10 – Extrait d'OntoPneumo sous PROTÉGÉ.

ritablement de démarches ascendantes ou descendantes dans notre première étape de construction puisque les axes que nous avons identifiés relèvent plutôt de la *« middle »* ontologie. À ce stade, la définition des principes différentiels est entièrement manuelle. Nous essayons, autant que faire ce peut, de réfléchir à la place de chaque concept dans la hiérarchie et de respecter les choix conceptuels qui sont obligatoirement faits à ce moment là de la construction. Nous utilisons ensuite les résultats de l'analyse par patrons lexico-syntaxiques sur le corpus [LIVRE] pour faire évoluer l'ontologie :

1. ces résultats nous aident à renseigner les principes différentiels et donc à vérifier la cohérence de notre modélisation,
2. ils nous permettent d'enrichir la hiérarchie de l'ontologie en prenant en compte les informations apportées par la comparaison des analyses terminologiques – complémentaires ou divergentes – issues des deux corpus,
3. ils apportent des renseignements intéressants à intégrer à l'ontologie : synonymes, acronymes ...

La *« middle »* ontologie de la pneumologie compte à ce stade 1 460 concepts primitifs auxquels viennent s'ajouter les 800 concepts primitifs de la top-ontologie du projet MENELAS et les concepts définis à l'aide du logiciel PROTÉGÉ. Cette ontologie complète de la pneumologie compte au total 2 260 concepts primitifs.

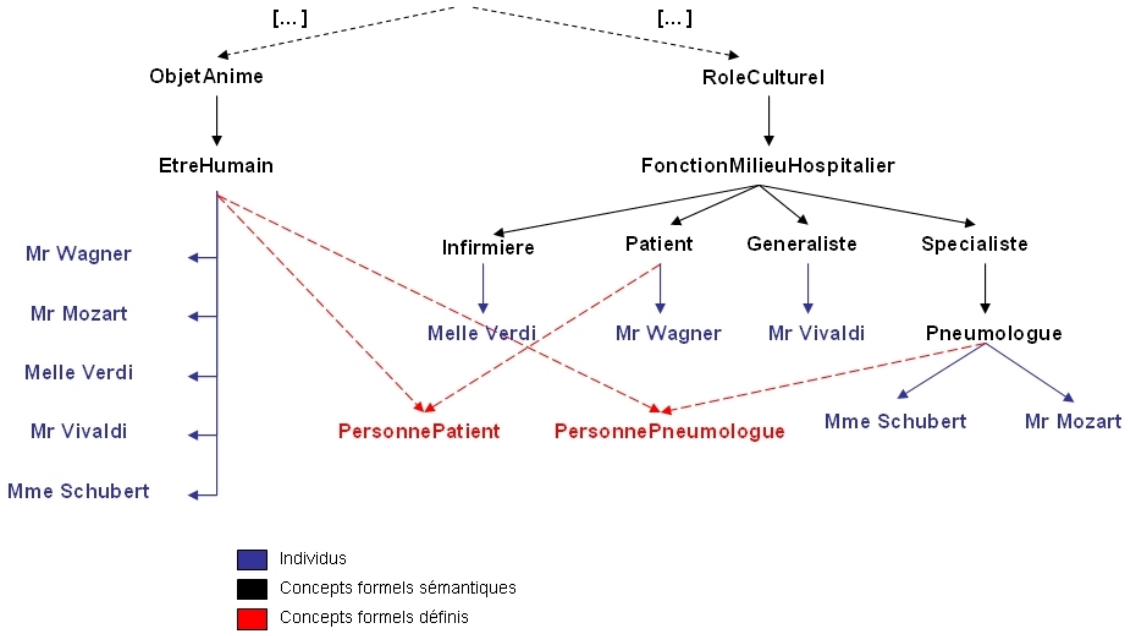


FIG. 4.11 – Illustration d'une structure de treillis à l'étape de formalisation.

Notre travail s'est centré sur la création des concepts et nous n'avons que très peu réfléchi aux relations en général. Nous avons choisi de reprendre les relations présentes dans MENELAS en adaptant leur dénomination à nos besoins. Leur validation devra obligatoirement être entreprise dans le cadre de notre réflexion sur la réutilisation de la « top-ontologie » de MENELAS (*cf.* chapitre 7, section 1). Dans le cadre de ce travail, il sera probablement nécessaire de valider les choix faits dans le projet MENELAS en les comparant à des recherches sur les relations à partir de nos corpus de travail (Aussenac-Gilles *et al.*, 2005; Aussenac & Seguela, 2000).

9 Discussion et conclusion

Nous avons présenté un ensemble de traitements pour la construction de hiérarchies terminologiques fondées sur deux méthodologies de Traitement automatique de la langue, adaptées chacune à un type et genre de corpus : l'analyse distributionnelle sur un corpus redondant et riche en termes spécialisés, et l'extraction par patrons lexico-syntaxiques sur un corpus didactique à la structure régulière. Nos résultats ont montré :

1. l'intérêt de chacune des méthodes en fonction du type et du genre de corpus,
2. la nature des connaissances pouvant être identifiées par chaque approche,
3. la complémentarité et la différence des résultats de chacune.

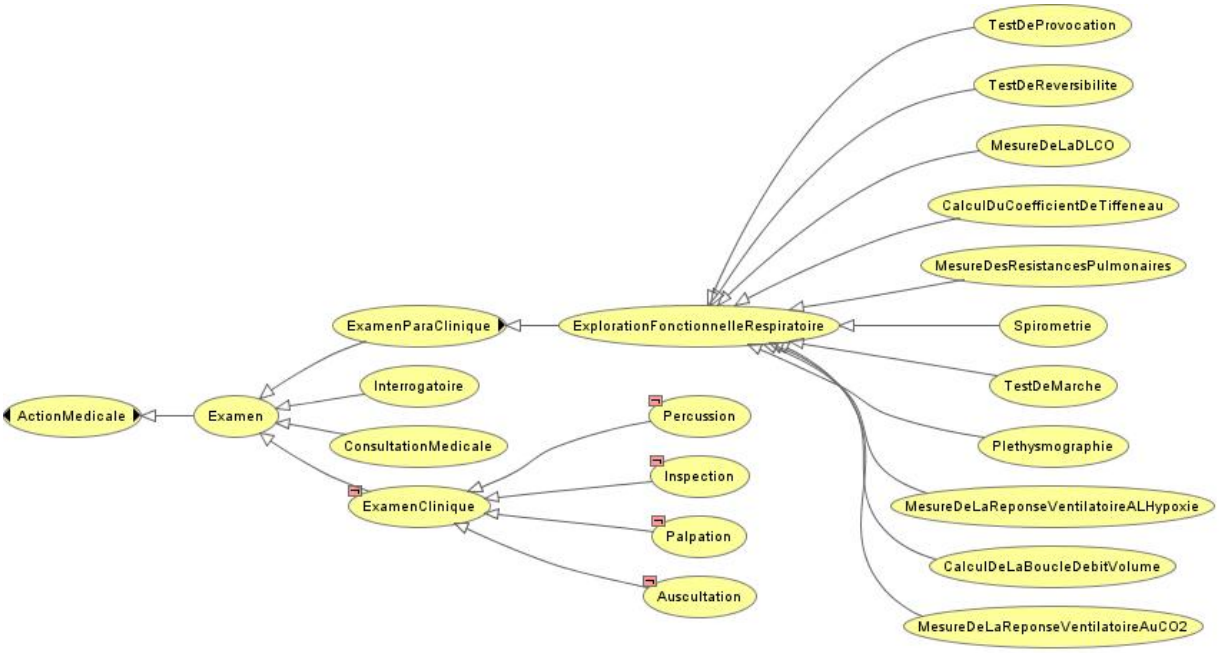


FIG. 4.12 – Extrait d'OntoPneumo vu avec le pluggin OWLViz de PROTÉGÉ.

Bien que les traitements aient porté sur des corpus différents, il est intéressant d'observer la relative compatibilité des deux ensembles terminologiques extraits. Comme nous l'avons dit dans ce chapitre, la question des genres textuels n'est pas nouvelle. Cependant, l'approche comparative de ce travail est plus rarement exploitée.

La divergence des structures terminologiques obtenues est également un point intéressant. Elle tend à prouver l'existence d'organisations conceptuelles différentes au sein d'un même domaine de connaissances. Ce point va à l'encontre d'un certain nombre de travaux sur la modélisation d'ontologies universelles. En effet, une ontologie est une modélisation conceptuelle, non contextuelle et non ambiguë. Elle n'a donc qu'un seul contexte d'interprétation possible. Nous sommes ainsi convaincus qu'il existe de nombreuses modélisations possibles pour un domaine donné, en fonction de la tâche à réaliser, c'est-à-dire en fonction du contexte. Il en va de même des raisons qui gouvernent la création des corpus et des ressources qui les constituent. La figure 4.13 extraite de MENELAS et la figure 4.14 extraite de ROME, une *« middle »* ontologie construite par D. Pisanelli (LOA, Rome, Italie), illustrent deux manières différentes de considérer les médicaments (*« drug »*). Ainsi, le travail de structuration ontologique vise à expliciter les choix faits parmi l'ensemble des modélisations potentielles. Il s'agit bien de construire des ontologies régionales car toute modélisation n'est jamais qu'un point de vue sur le monde. Ces conclusions sont corroborées par les travaux du groupe Terminologie et Intelligence Artificielle : proposer à un ingénieur des connaissances, non spécialiste du domaine, des vues complémentaires ou divergentes sur le domaine lui donne des arguments critiques pour faire ses choix et les valider. L'idée est alors de développer un outil à même de proposer ces vues complémentaires en

gardant à l'esprit les necessités ergonomiques qu'entraine une telle démarche.

FIG. 4.13 – Concept *Medicament* (*commercial-drug*) dans MENELAS.

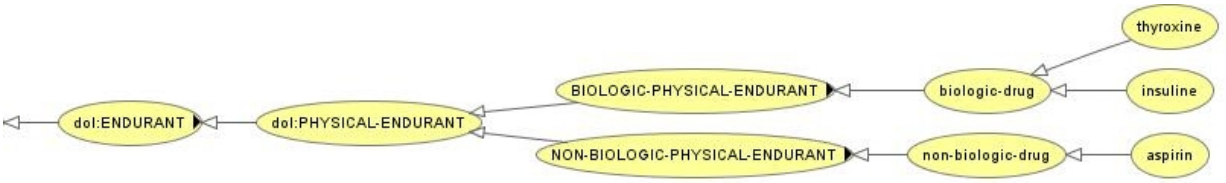


FIG. 4.14 – Concept *Medicament* (*biologic-drug* et *nonbiologic-drug*) dans ROME.

Nous avons toutefois cerné certaines des limites liées à cette comparaison : un rapprochement plus automatique de ces deux ensembles terminologiques nécessiterait, d'une part, de mettre en œuvre des techniques plus sophistiquées d'appariement, et, d'autre part, d'améliorer la précision (et probablement le rappel) des patrons lexico-syntaxiques et des marqueurs de relation définitoire en les adaptant spécifiquement au domaine médical. Ainsi, des marqueurs comme *indiquer* et *définir* sont à spécifier plus finement : *indiqué* est souvent utilisé dans le cadre de « traitements *indiqués* pour soigner une pathologie », *définir* intervient surtout dans des phrases telle que « Ces résultats ont permis d'acquérir certaines connaissances et de *définir* les meilleurs traitements pour soigner les 30 patientes ». Toutefois, ces mêmes patrons permettent déjà de repérer d'autres relations propres au domaine médical, qu'il serait alors intéressant d'isoler et de caractériser plus précisément. Il serait également pertinent de comparer les hiérarchies non redondantes entre les deux structures modélisées avec une terminologie ou un thésaurus de référence comme le MESH , pour vérifier leur cohérence et validité propres ou, le cas échéant, proposer des suggestions de compléments au MESH. Il serait également intéressant de comparer notre ontologie avec la SNOMED-CT et peut-être y trouver des concepts utiles à inclure dans notre modélisation. Même si des travaux menés dans notre équipe montrent les manques de la SNOMED par rapport à un problème précis (Steichen *et al.*, 2007).

Ce travail de modélisation des connaissances à partir de textes nous a également permis de mesurer la nécessité d'utiliser conjointement des outils de Traitement automatique du langage (SYNTEX-UPERY) et de modélisation (DOE, PROTÉGÉ). Il semblerait intéressant d'intégrer SYNTEX-UPERY et DOE pour faciliter le passage des candidats termes à la représentation des concepts en suivant la méthode ARCHONTE , tout en assurant de pouvoir revenir aux textes. C'est ce qui est fait dans TERMINAE (Szulman & Biébow, 2004) et sur lequel nous reviendrons dans le projet DAFOE4APP .

Enfin, nous précisons que la validation de l'ontologie s'est faite selon un processus itératif encadré par un expert du domaine, le docteur F.-X. Blanc de l'hôpital du Kremlin-Bicêtre. La

dernière phase de validation se fera par l'usage, un déploiement de l'application d'aide aux codages utilisant l'ontologie (*cf.* chapitre 5) est prévu dans deux hôpitaux de l'Assistance-Publique Hôpitaux de Paris dans le courant de l'année 2008. L'ontologie doit être mise à disposition du corps médical dans le cadre de la plate-forme terminologique du projet DaFOE4App. Nous tenterons alors de quantifier et de qualifier l'aide que notre travail apporte aux pneumologues selon des scénarios d'usage précis.

Pour conclure, nos contributions au domaine de l'Ingénierie des connaissances sont au nombre de quatre :

1. Nous mettons au point une méthodologie d'Ingénierie ontologique unifiée (tenant compte des principes et des méthodes de l'Ingénierie ontologique, de l'Ingénierie des connaissances, du Traitement automatique des langues, de la logique et de la sémantique différentielle) pour la construction d'ontologies, à partir de textes. À cet effet, nous avons complété et précisé la méthodologie ARCHONTE mise au point pas B. Bachimont, en montrant comment s'enchaînent les différentes étapes : 1) élaboration et traitement des corpus textuels, 2) identification, sélection et extraction des termes pertinents, 3) normalisation, 4) formalisation et 5) opérationalisation.
2. Nous expérimentons la complémentarité de deux modes d'analyse de la langue en usage dans les textes, l'analyse distributionnelle et l'approche par patrons lexico-syntaxiques et montrons que l'utilisation conjointe de ces méthodes facilite la structration hiérarchique des concepts de l'ontologie en aidant la définition des axes différentiels.
3. Nous expliquons comment la réutilisation d'une ontologie de haut niveau de la médecine vient guider la réorganisation d'une première structuration des concepts.
4. Enfin, nous essayons d'abstraire la méthode du seul domaine médical et proposons en annexe un guide de bonne pratique méthodologique à l'usage de l'ingénieur des connaissances. Nous espérons ainsi qu'elle pourra être réutilisée dans d'autres travaux de construction d'ontologies.

CHAPITRE 5

MedCKARe, un outil pour le codage des CRH

« La théorie, c'est quand on sait tout et que rien ne fonctionne.
La pratique, c'est quand tout fonctionne et que personne ne sait pourquoi.
Ici, nous avons réuni théorie et pratique : Rien ne fonctionne ... et personne ne sait pourquoi ! »
Albert Einstein

*Ce chapitre est consacré à Med*CKARe *(Medical Coding by Knowledge Acquisition and Representation), l'outil d'aide au codage que nous avons développé dans le cadre du projet* PERTOM*ed. Cet outil utilise comme principale ressource l'ontologie de la pneumologie dont nous avons décrit le processus de construction dans le chapitre précédent. Nous présentons, en section 1, les objectifs que nous nous sommes fixés et les hypothèses sur lesquelles nous avons fondé l'élaboration de Med*CKARe*. La section 2 s'intéresse aux outils industriels de codage et en propose une modeste revue. Nous détaillons ensuite, dans la section 3, les ressources dont nous disposons pour construire le cœur du système. La section 4 est consacrée à l'outil dont nous nous sommes servi. La section 5 explique le mode de fonctionnement de Med*CKARe *en détaillant chacune des étapes de traitement. La section 6 présente les résultats que nous avons obtenus sur nos corpus d'évaluation. Enfin, la section 7 clôt ce chapitre et propose un certain nombre de pistes pour les développements futurs.*

1 Objectifs et hypothèses

La réduction des inégalités de ressources entre les établissements de santé figure dans la réforme de l'hospitalisation (ordonnance du 24/04/96). Afin de mesurer l'activité et les ressources des établissements, le gouvernement souhaite disposer d'informations quantifiées et standardisées. Telle est la vocation du Programme de Médicalisation des Systèmes d'Information (*cf.* cha-

pitre 3 section 2.3). Ainsi, le codage des diagnostics et des actes médicaux est devenu une obligation légale des structures de soins. Nous avons vu dans le chapitre 2 que cette obligation touche toutes les spécialités médicales et que, loin d'être un processus anodin, elle demande une vraie disponibilité du médecin responsable ainsi que l'acquisition de nouvelles compétences. Par rapport aux systèmes classiques d'aide au codage (Friedman *et al.*, 2004), nous recherchons d'abord une représentation conceptuelle, là où les travaux habituels recherchent directement une représentation dans un thésaurus, le MeSH , l'UMLS (*via* son métathésaurus) ou la SNOMED (Dolin *et al.*, 2001). Notre hypothèse est que, pour proposer un environnement d'aide au codage performant, il faut : (1) permettre aux médecins de se réapproprier le codage des patients en mettant en place un codage médical de ceux-ci, (2) proposer un système semi-automatique pour assister, et non pas remplacer, le médecin dans sa tâche de codage médico-économique. Par codage « médical », nous entendons un codage où le médecin exprime les constats médicaux qu'il a fait sur le patient. À terme, nous souhaitons utiliser ce codage pour indexer, de manière évoluée, les informations pertinentes contenues dans un CRH et permettre au personnel médical de rechercher, par exemple, tous les cas de patients atteints de *sténose serrée de la trachée à la fois par compression extrinsèque et par envahissement de la muqueuse* diagnostiquée par *une endoscopie bronchique*. Pour l'instant, aucun outil ne permet de faire de telles recherches mais nous ne sommes pas la seule équipe en France à chercher une solution de ce type. Nous pensons en particulier à un travail proche du notre, celui de S. Pereira *et al.* (2006; 2007).Ce codage médical est évidemment une représentation réductrice de ce que contient le dossier médical du patient mais c'est parce qu'elle est réductrice que cette représentation permet au médecin un rappel rapide, une représentation résumée de ce qu'il sait sur le patient. Comme nous le verrons dans le chapitre 7, cette fonctionnalité sera développée dans le cadre d'un projet post-doctoral baptisé MedOC . Enfin, la représentation n'étant pas, de façon figée, dédiée au codage médico-économique, c'est bien un « environnement » de codage que nous proposons ici plutôt qu'un outil « d'aide au codage ».

Contrairement aux outils commerciaux existants (*cf.* section suivante) dont les médecins ne sont pas entièrement satisfaits, nous ne souhaitons pas automatiser complètement la procédure de codage. Notre idée est plutôt de proposer un système de représentation des connaissances avec lequel le médecin puisse interagir pour construire la représentation du patient qu'il désire, en tenant compte de ses propres capacités de choix et d'interprétation mais en l'aidant dans sa tâche. Le problème du codage se situe au niveau du passage au formalisme, toujours difficile. L'expression des connaissances, qui se présente le plus souvent et le plus naturellement sous forme textuelle, doit être transformée en un codage qui, lui, est toujours réducteur en termes de représentation du sens. Nous devons donc proposer un outil qui permette le codage en même temps qu'il permette aux médecins d'assumer le caractère réducteur de ce processus. Il faut également qu'ils puissent interagir avec le système formel dans les termes de leur domaine de spécialité, répertoriés dans des thésaurus (CCAM, CIM-10 ...).

2 Outils existants

De nombreux outils industriels d'aide au codage des actes sont actuellement disponibles. Bien que nous ayons privilégié le codage des diagnostics dans MedCKARe, nous pensons que

cette modeste revue des outils (nous n'avons pu en tester qu'un très petit nombre) dédiés aux actes peut nous donner des renseignements intéressants pour notre outil : fonctionnalités existantes, types de recherches, temps de réponse d'une recherche, communautés d'utilisateurs ... Nous les avons classés selon deux critères : leur organisation interne et les services qu'ils se proposent de rendre aux médecins.

Outils proposant une recherche hiérarchisée

L'utilisateur choisit un domaine puis navigue dans l'arborescence pour choisir l'expression la plus adaptée à la pathologie ou à l'acte médical et trouver le code correspondant. Une recherche dans l'index analytique de la CCAM ou de la CIM-10 suit les mêmes principes (Organisation mondiale de la santé, 1995).

- L'outil d'aide au codage des actes C.O.R.I.M[1] (*cf.* figure 5.1) propose une navigation hiérarchique, par mots-clés, par code, ou une recherche directe dans le thésaurus. Cet outil a été développé par le Collège Régional de l'Information Médicale – association « Loi 1901 » créée en 1992 – à l'initiative des médecins responsables des Départements d'Information Médicale de la Région Poitou-Charentes. L'outil n'est pas très ergonomique : le seul moyen de naviguer dans la hiérarchie proposée est le double-clic et les résultats des recherches qui peuvent être très nombreux sont présentés à la suite les uns des autres. Par contre, il est intéressant de pouvoir visualiser les temps de réponse des recherches, de pouvoir effectuer une nouvelle recherche à partir de n'importe quelle page du site. De plus, le site propose une annexe qui explique comment sont construits les codes de la CCAM.
- WEBCCAM[2], CODAM.S[3] et EDOCOD[4] sont des outils de recherche de codes (par mots-clés et hiérarchique) classique mais proposent un système de hiérarchie personnalisée : un utilisateur peut ne conserver qu'une partie de la hiérarchie, utile à la description de son activité. Ils gèrent également un système de favoris : l'utilisateur se voit proposer les codes qu'il utilise régulièrement. L'outil EdoCod a pour originalité de proposer à l'utilisateur de créer une hiérarchie personnalisée et de conserver uniquement les parties utiles à la description de son domaine. Cela dit, l'ensemble de ces outils nécessitent une bonne connaissance des principes d'organisation de la classification utilisée.

Outils proposant une recherche « lexicale »

Les outils fondés sur une recherche lexicale permettent de chercher un mot ou un texte pour trouver le code correspondant. Ils permettent à l'utilisateur de formuler des requêtes en langage naturel. Ces outils sélectionnent les termes saisis par l'utilisateur, jugés pertinents par le système, et proposent des expressions qui représentent au mieux ces termes. Cependant, l'utilisateur rencontre rapidement des problèmes car les variantes lexicales et les synonymes semblent très mal gérés. Le principal intérêt de ces outils est de permettre aux utilisateurs, peu familiers des termes employés dans la CIM , de pouvoir coder leurs diagnostics.

[1] http://corim.pc.free.fr/toutpublic/corim_tout_public.htm
[2] http://www.webccam.net/
[3] http://www.symphonieonline.com/files/simccam.pdf
[4] http://www.micro6.fr/CCAM.htm

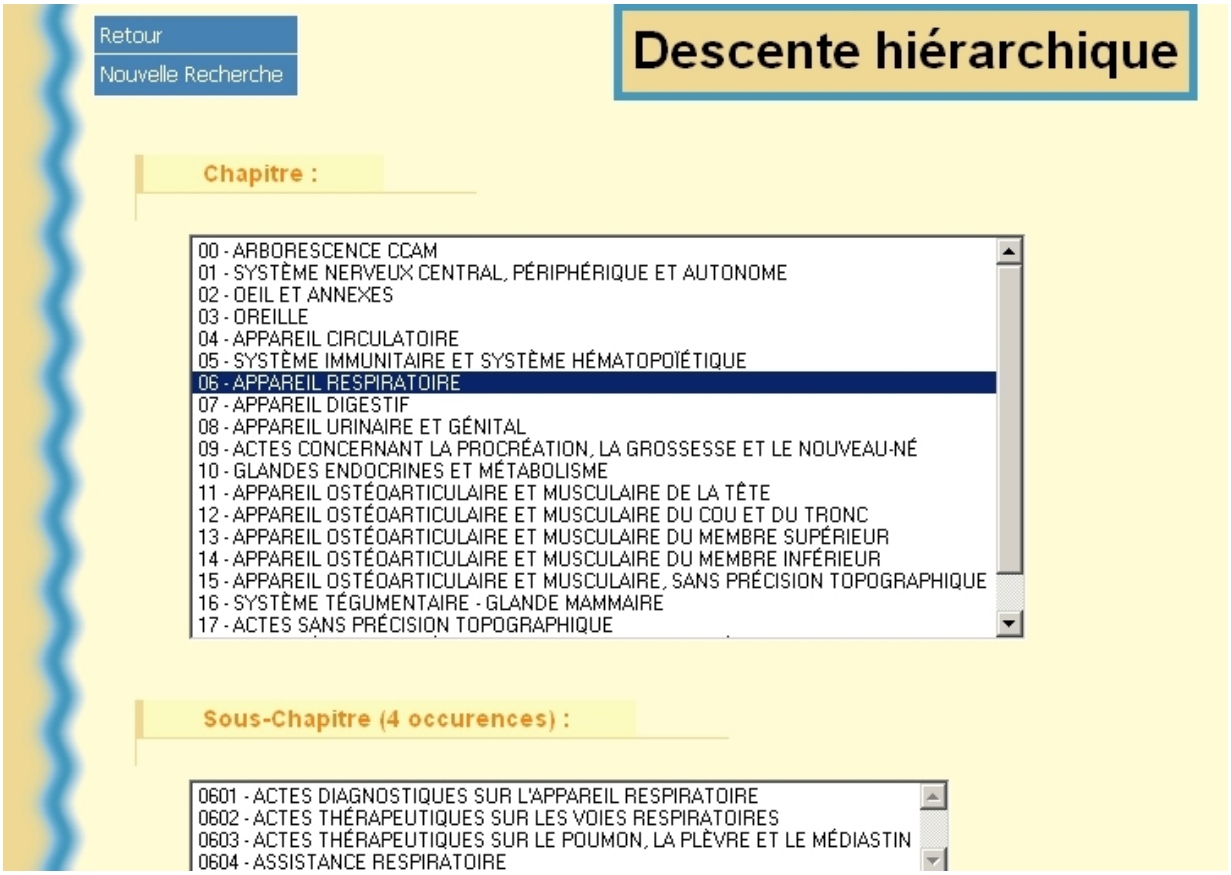


FIG. 5.1 – Copie d'écran de l'interface de recherche hiérarchique de l'outil C.O.R.I.M.

– LUCID[5] est un outil qui permet la recherche, en utilisant le langage naturel, de codes diagnostics et d'opérations sur la base des classifications CIM-10 et CIM-9. Les résultats des recherches peuvent être gardés dans une base centrale d'archives à des fins statistiques et d'aide à la décision dans le pilotage de l'établissement.

Outils fondés sur un système documentaire

Les systèmes documentaires, plus complexes, indexent les codes à l'aide des mots d'un thésaurus qui est un répertoire alphabétique de termes. La présence d'un thésaurus permet de définir des synonymes. Ainsi une requête avec le mot *« angor »* permettra de retrouver le code *« I209, angine de poitrine sans précision »*, bien que le terme *angor* ne soit pas contenu dans le libellé. Ces systèmes reposent sur une représentation des données plus sophistiquée étant donné que les différents termes représentant un concept sont reliés par une relation « is-a » (Kohler *et al.*, 1990; Cimino *et al.*, 1994).

– Le logiciel d'aide au codage CIM et CCAM du Dr J. Ruiz proposent deux types de recherche : une recherche classique par parcours des hiérarchies des thésaurus et une autre pour des utilisateurs expérimentés faisant intervenir des dictionnaires correspon-

[5] http://www.nicecomputing.ch/lucid/frameset.html

dant au découpage des codes[6] CCAM. La recherche pour utilisateurs expérimentés est très difficile à utiliser sans manuel, même pour des spécialistes.

Outils fondés sur un système d'ingénierie des connaissances

Il existe également des outils fondés sur des systèmes utilisant des techniques très proches de l'ingénierie des connaissances dont la spécificité est de prendre en compte le contexte et de guider l'utilisateur vers certains codes en lui proposant de choisir des « concepts » ou « notions » clés.

– NEPAL[7] est un outil d'aide à la codification CCAM sous licence CeCiLL. Il intègre la base CCAM version 1. Cet outil propose un mode de recherche avancée qui guide l'utilisateur à l'aide de notions prédéfinies. Les recherches peuvent être élargies aux mots proches ou aux synonymes (*cf.* figure 5.2). Les résultats proposent un lien vers les codifications de référence (*cf.* figure 5.3).

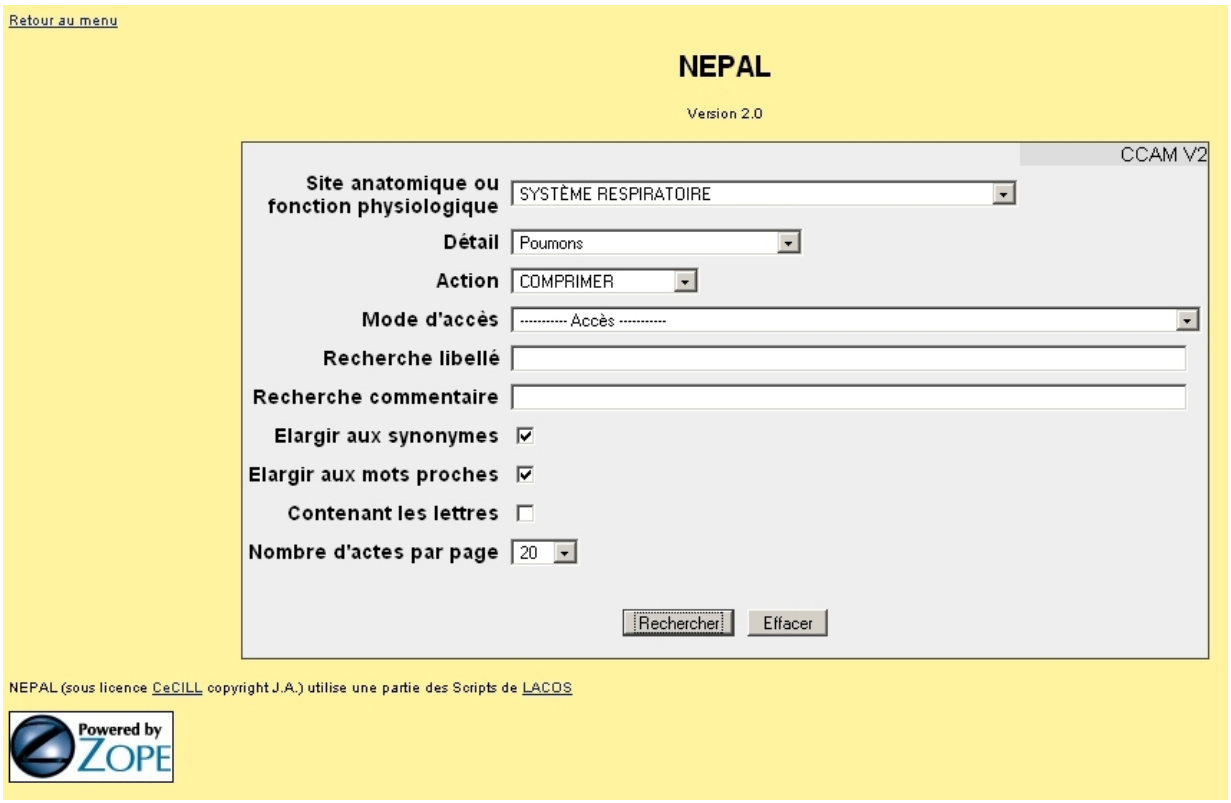


FIG. 5.2 – Copie d'écran de l'interface de recherche de l'outil NEPAL.

– CODAZ[8] propose plusieurs types de recherche (mots-clés, parcours des hiérarchies, « concepts » …) des codes CIM-10 et CCAM. Concernant la recherche par mots-

[6] Le code de chaque acte CCAM comprend 7 caractères. Les 4 premiers caractères à gauche sont des lettres dont le choix est précisé dans des dictionnaires (*cf.* Manuel d'utilisation de la CCAM). Par exemple, les deux premières lettres à gauche du code sont issues du dictionnaire des topographies.

[7] http://medecinelibre.nuxeo.org/nepal/

[8] http://membres.lycos.fr/pradeau/PMSI/outils/CCAM/CODAZ.htm

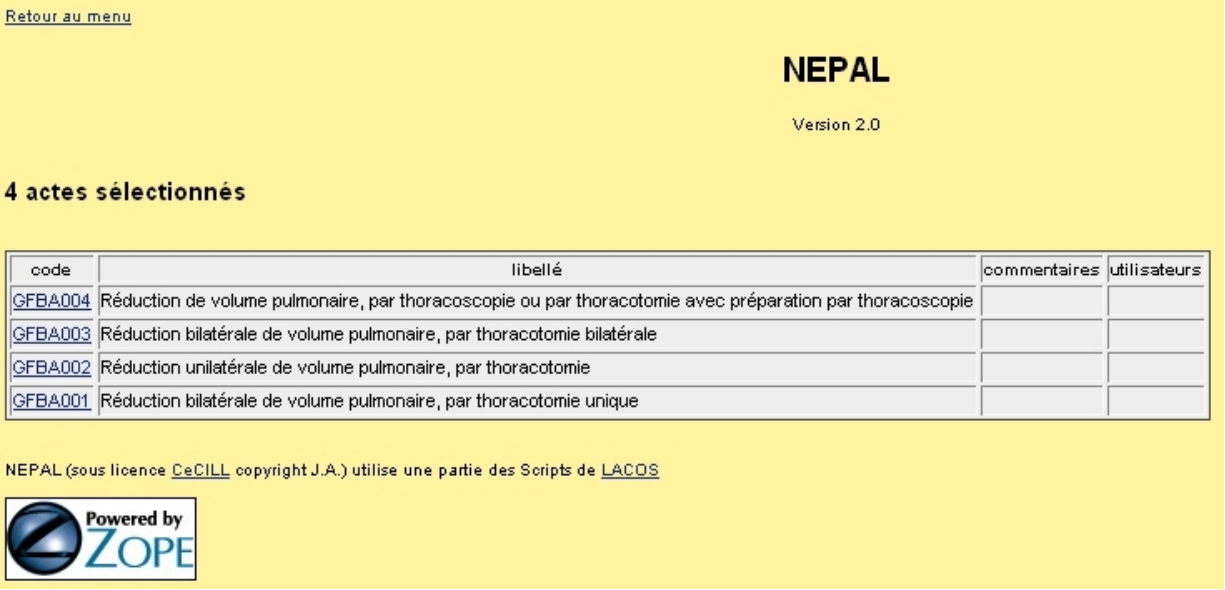


FIG. 5.3 – Copie d'écran de l'interface des résultats de l'outil NEPAL.

clés l'utilisateur dispose d'un dictionnaire de synonymes qu'il peut modifier (fonctions d'ajout et de suppression). La recherche par concepts facilite le repérage des codes. L'ajout d'un concept lié à une ou plusieurs actions permet de retrouver les termes qui se rapportent aux différentes combinaisons (topographie, type d'intervention, actions, accès ...). Cet outil propose également un mode de recherche topographique dans lequel l'utilisateur sélectionne sur un schéma du corps humain, une partie du corps. L'outil recherche alors les différentes notions en relation avec la partie du corps sélectionnée (*cf.* figure 5.4).

Nous nous sommes inspirés de SNOCODE[9], un outil de codage de phrases et de textes cliniques en SNOMED. Cet outil est commercialisé par la société MedSight qui développe des outils qui permettent aux professionnels de la santé de codifier des textes dans un anglais standardisé de la SNOMED-CT, des expressions de la SNOMED française ou de la CIM ou des expressions provenant d'autres classifications médicales. Cette société est aussi le distributeur mondial de la version électronique de la SNOMED française. SNOCODE codifie les textes soumis de façon automatique ou interactive selon le choix de l'utilisateur. Les textes à codifier peuvent parvenir d'éditeur de textes tel que Word, ou en ASCII directement de textes d'applications informatiques départementales (laboratoires) ou hospitalières. Cet outil peut aussi alimenter des bases de données de toutes sortes comme SQL SERVER, ORACLE, Access ... En mode automatique, SNOCODE établit les congruences entre des segments (jusqu'à 14 mots à la fois) de phrases et les énoncés semblables ou synonymes de la SNOMED. Les extrants de SNOCODE sont paramétrés. L'utilisateur choisit les codes, phrases, autres classifications à porter dans les fichiers en extrant.

[9] http://www.medsight-info.com/FR/Products/Snocode.html

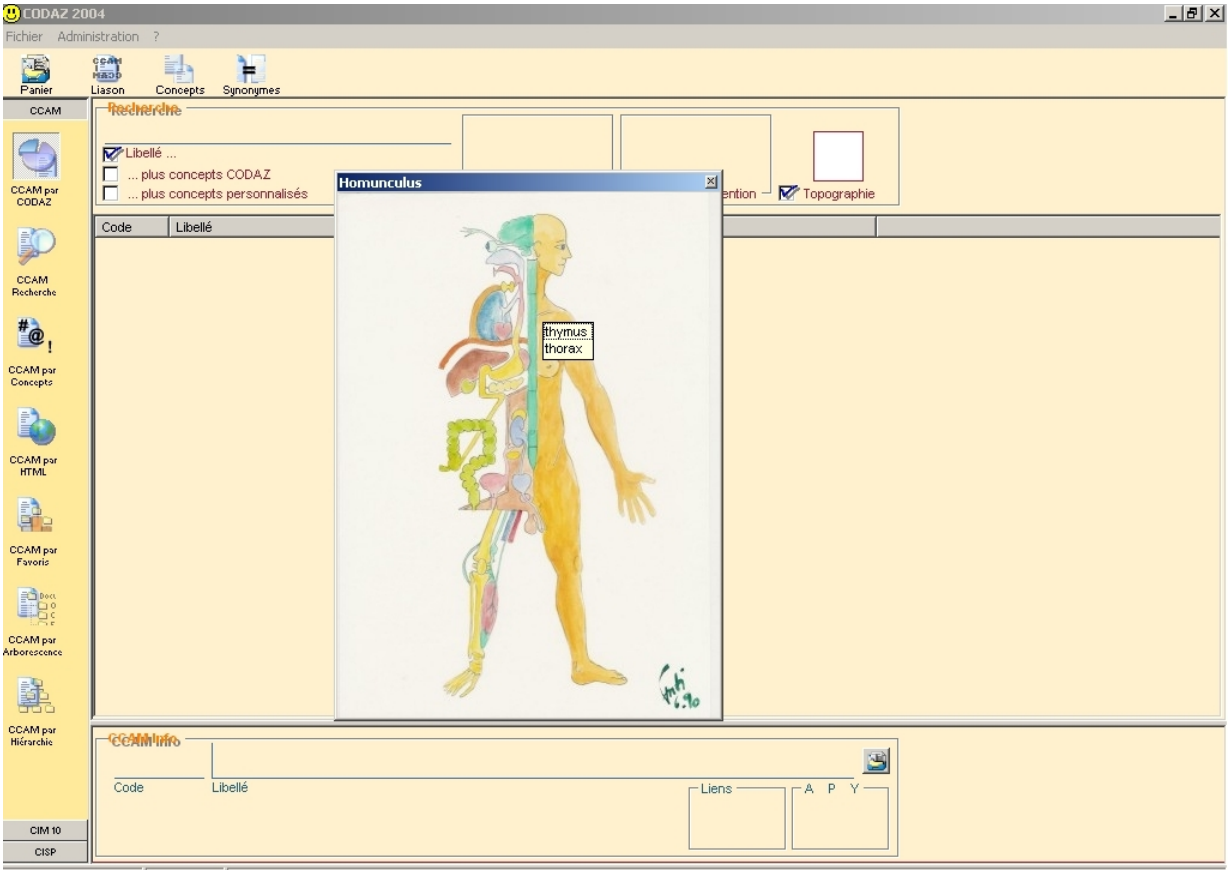


FIG. 5.4 – Copie d'écran de la recherche topographique dans l'outil CODAZ.

MEDLEE[10] (Medical Language Extraction and Encoding System) est un système de traitement automatique des langues qui fait ses preuves dans le domaine médical (Friedman *et al.*, 1994). Il est assez proche de ce que nous souhaitons faire avec MedCKARe. Il analyse les comptes rendus cliniques en utilisant une grammaire et un lexique identifiant des structures de phrases et en se servant de patrons lexico-syntaxiques. Sa méthode d'extraction repose sur la collecte, au préalable, de toutes les structures de phrases repérées dans des textes médicaux afin de pouvoir s'en servir pour retrouver d'éventuelles occurrences dans de nouveaux textes. Il encode chaque terme extrait en retrouvant le terme clinique standard associé, à savoir le terme équivalent contenu dans l'UMLS . Les informations ainsi obtenues peuvent être utilisées par différents types d'applications. MEDLEE offre une des fonctionnalités que notre outil propose qui est la représentation des informations mais il ne gère pas le codage PMSI, et il se base sur l'UMLS et non pas sur les terminologies usuelles dans différentes spécialités médicales.

Ces logiciels sont une aide importante pour le médecin mais l'investissement personnel et le temps d'utilisation qu'ils nécessitent sont un frein réel à leur emploi. De plus, ces différents types d'outils, à l'exception de SNOCODE, ne s'utilisent pas directement sur des textes, comme c'est

[10] `http://lucid.cpmc.columbia.edu/medlee/`

le cas avec MedCKARe. En effet, soit ils s'appuient sur des requêtes d'utilisateurs portant sur le code d'une expression médicale indiquée dans la requête, soit ils proposent une vue complète sur les données à coder pour permettre à l'utilisateur d'y naviguer. La plupart des médecins que nous avons rencontrés codent manuellement les CRH en utilisant une liste des pathologies les plus courantes extraites du thésaurus de leur spécialité. Nous proposons un outil où le codage sera semi-automatique et donc plus rapide, le médecin n'ayant plus qu'à valider les propositions faites par le système. Un système de sauvegarde des préférences sera implémenté dans le cadre du projet MedOC .

3 Ressources

Nous présentons ci-dessous l'ensemble des ressources qui constituent le cœur de MedCKARe.

3.1 Ontologie de la pneumologie

L'ontologie décrite au chapitre précédent est la ressource de base de MedCKARe. Elle compte 1 460 concepts primitifs auxquels s'ajoutent les 800 concepts repris du haut de l'ontologie du projet MENELAS . Soit, au total, 2 260 concepts. Nous utilisons l'ontologie au format OWL, nous l'analysons à l'aide de programme *ad hoc* et nous stockons les informations qui nous intéressent dans des tables MySQL. Elle nous fournit, *via* un terme préféré associé à chaque concept, un premier vocabulaire du domaine. Ce sont les termes de ce vocabulaire que nous cherchons à repérer dans les CRH.

3.2 Thésaurus de spécialité

Nous disposons également d'un thésaurus de spécialité de la pneumologie mis au point et fourni par la Société de Pneumologie de Langue Française[11]. Ce thésaurus prend en compte les cas de codages médico-économiques les plus couramment utilisés par les pneumologues. Il comprend deux parties : la première code les diagnostics selon le codage médico-économique issu de la CIM-10 (*cf.* figure 5.5) et la seconde code les actes avec la CCAM (*cf.* figure 5.6). Pour l'instant, nous avons uniquement pris en compte la partie concernant le codage des diagnostics. Nous ne disposons pas des ressources nécessaires pour évaluer notre travail sur le codage des actes. Nous prévoyons d'ajouter la gestion du codage des actes dans la version suivante de MedCKARe .

3.3 Corpus de référence

Nous disposons également de deux corpus de comptes rendus d'hospitalisation avec lesquels nous avons testé MedCKARe . Le premier de ces corpus est constitué de 400 CRH non codés, c'est-à-dire ne contenant pas de codes PMSI. Il provient de trois hôpitaux différents de l'Assistance Publique-Hôpitaux de Paris : l'hôpital du Kremlin-Bicêtre, l'hôpital de Créteil et l'hôpital

[11] http://www.splf.org

CATÉGORIE	LIBELLÉ DIAGNOSTIQUE	CODE validé PERNNS	CODE SPLF	REMARQUES
BPCO	Asthme allergique	J45.0		
	Asthme non allergique	J45.1		
	Asthme SAI	J45.9		
	Hyperréactivité bronchique	R94.2	R94.2 | | | 0	
	État de mal asthmatique, asthme aigu grave	J46		
	Bronchite aiguë SAI	J20.9		
	Bronchiolite SAI	J21.9		
	Bronchite chronique simple	J41.0		
	Surinfection de bronchite chronique	J41.1		
	Bronchite chronique obstructive	J44.8		
	Bronchite chronique SAI	J42		
	Dilatation des bronches	J47		
	Emphysème pan-lobulaire	J43.1		
	Emphysème SAI	J43.9		
	Mucoviscidose avec manifestations pulmonaires	E84.0		*à utiliser comme étiologie en diagnostic associé*
IRC - IRA	IRC obstructive	J96.1 +0		
	IRC restrictive	J96.1 +1		
	IRC SAI	J96.1		*rajouter l'étiologie en associé*
	IR aiguë	J96.0		*rajouter l'étiologie en associé*
	OAP lésionnel	J81		
	Syndrome d'apnée du sommeil avec *overlap syndrome*	J44.8 + G47.3		

FIG. 5.5 – Extrait du thésaurus de spécialité de la pneumologie concernant le codage des diagnostics.

Saint-Antoine. Dans la suite de ce chapitre, nous appellerons ce corpus [NONCODE]. Le second corpus est constitué de 100 CRH codés par des pneumologues. 50 CRH proviennent du CHU de Rouen et 50 CRH proviennent de l'hôpital du Kremlin-Bicêtre. Nous appellerons ce corpus [CODE]. Ces corpus sont utilisés au format texte.

3.4 Ressources lexicales

Nous utilisons des ressources lexicales de type médicales sur lesquelles a également travaillé N. Grabar (Grabar & Zweigenbaum, 2000). Ces tables sont issues de nos corpus [CRH] et [LIVRE] ainsi que d'autres ressources additionnelles de même type dont dispose notre laboratoire. Elles nous servent à enrichir le vocabulaire fourni par l'ontologie et, par conséquent, à améliorer la reconnaissance automatique des expressions pertinentes dans les CRH. Ces ressources lexicales comprennent :

06.01	ACTES DIAGNOSTIQUES SUR L'APPAREIL RESPIRATOIRE				
06.01.02	Échographie de l'appareil respiratoire				
	À l'exclusion de : échographie et/ou échographie-doppler de contrôle ou surveillance de pathologie (chapitre 19)				
GBQM001 (F, P, S, U)	**Échographie unilatérale ou bilatérale du sinus maxillaire et/ou du sinus frontal**	1	0	37,80	2
GFQM001 (F, P, S, U)	**Échographie transthoracique du médiastin, du poumon et/ou de la cavité pleurale** *Échographie transthoracique du thymus*	1	0	37,80	2
GFQJ002 (F, P, S, U)	**Échographie du médiastin et/ou du poumon, par voie œsophagienne ou par voie bronchique**	1	0	37,80	2
06.01.03	Radiographie de l'appareil respiratoire				
ZBQK002 (B, D, E, F, P, S, U, Y, Z)	**Radiographie du thorax** *Radiographie pulmonaire* *À l'exclusion de : radiographie du squelette du thorax (LJQK001)* *(YYYY030, ZZLP025)*	1	0	21,28	2
LJQK002 (B, D, E, F, P, S, U, Y, Z)	**Radiographie du thorax avec radiographie du squelette du thorax** *Radiographie pulmonaire avec gril costal* *(YYYY030, ZZLP025)*	1	0	45,22	2
ZBQK003 (E, F, P, S, U, Y, Z)	**Examen radiologique dynamique du thorax, pour étude de la fonction respiratoire et/ou cardiaque** *Étude radiologique de prothèse valvulaire cardiaque* *Étude radiologique de la cinétique des coupoles diaphragmatiques* *Avec ou sans : opacification* *(YYYY030)*	1	0	21,28	2
GEQH001 (E, F, P, S, U, Y, Z)	**Bronchographie** *(YYYY030, ZZLP025)*	1	0	33,25	2

FIG. 5.6 – Extrait du thésaurus de spécialité de la pneumologie concernant le codage des actes.

- une table de synonymes contenant à chaque ligne la forme canonique d'un terme de l'ontologie et son synonyme : *anesthésie - analgésie*,
- une table de flexions où l'on trouve la forme canonique d'un terme suivi de sa forme fléchie : *oncogène, Adjectif - oncogènes, Adjectif* | *consulter, Verbe - consultera, Verbe* | *aigu, Adjectif - aiguës, Adjectif* | *cellule, Nom - cellules, Nom*
- une table de dérivations où la forme canonique d'un terme est associée à sa forme dérivée : *pleural, Adjectif - pleurésie, Nom* | *thérapie, Nom - thérapeutique, Adjectif* | *parasitose, Nom - parasitaire, Adjectif*,
- une table de compositions qui permettent de faire des rapprochements : *médiastin, Nom - médiastinoscopie, Nom* | *bronchiolite, Nom - bronchodilatateur, Nom* | *bactérien, Adjectif - bactériologique, Adjectif*.

Ces données sont stockées dans une base de données MYSQL qui contient donc plusieurs descriptions pour un même mot.

4 Unitex, un outil pour l'extraction d'informations

L'extraction des informations jugées pertinentes dans les CRH est réalisée à l'aide d'un outil de Traitement automatique des langues nommé Unitex[12]. Cet outil a été réalisé à l'Institut Gaspard Monge de l'Université Marne-la-Vallée (UMR 8049), et permet d'étudier, dans un corpus textuel, les concordances d'expressions (co-référence). Autrement dit, il permet d'extraire toutes les phrases qui contiennent une ou plusieurs séquences reconnues. Ces séquences sont reconnues d'après les patrons lexico-syntaxiques que nous avons construits. Nous avons choisi d'utiliser cet outil plutôt qu'un autre (YAKWA par exemple) car nous en connaissions déjà le fonctionnement. Unitex est composé d'une interface graphique en JAVA et de programmes externes écrits en C. Les textes bruts qui lui sont donnés en entrée sont prétraités selon 3 étapes :

1. découpage en phrases, comme par exemple : *{S} Majoration de la détresse respiratoire*[13] ;
2. découpage en unités lexicales, comme par exemple : « Majoration », « de », « la », « détresse », et « respiratoire » ;
3. étiquetage des unités lexicales grâce au dictionnaire du français présent dans Unitex et en utilisant, éventuellement, le dictionnaire construit par l'utilisateur.

L'étape de traitement consiste à appliquer les patrons lexico-syntaxiques au texte. Ceux-ci sont créés sous forme de transducteurs et peuvent reconnaître les phrases souhaitées figurant dans le texte. Ainsi, ils permettent de repérer et d'analyser des segments textuels porteurs de sens et de les étiqueter, malgré les différentes formes qu'ils peuvent prendre. Ces transducteurs sont des graphes syntaxiques permettant de reconnaître certaines séquences au sein des textes. Les transitions sont représentées par des nœuds. Chaque graphe possède un nœud initial et un nœud terminal. Les séquences extraites correspondant aux patrons sont appelées des instances de patrons. Par exemple, le patron <N><DET>*<N><A>, qui peut être également construit sous forme de transducteur, permet d'obtenir la phrase (ou autrement dit l'instance) : *Pleuropneumopathie de la base gauche*[14]. Nous pouvons éventuellement attribuer des sorties aux transitions du graphe. Lors de l'application du transducteur, ces sorties vont s'insérer devant les motifs reconnus. Nous expliquons cette partie en détails dans la sous-section 5.2. Unitex permet également à l'utilisateur de construire son propre dictionnaire de termes. Ce dictionnaire personnel peut être appliqué dans le cas d'une recherche d'informations particulières. Les patrons peuvent ensuite faire référence à ces dictionnaires d'utilisateur. Les dictionnaires fournissent des descriptions de mots simples et composés. Le tableau 5.7 représente une entrée type[15] dans un dictionnaire Unitex.

[12] `http://www-igm.univ-mlv.fr/\%7Eunitex/index.html`

[13] Le symbole {S} marque un début de phrase.

[14] Dans le patron, N est un code grammatical désignant un nom, DET un déterminant et A un adjectif. Le signe * signifie la présence facultative de l'élément.

[15] Le premier élément est la forme fléchie de l'entrée. Elle doit être suivie d'une virgule, puis de la forme canonique de l'entrée (qui peut être omise). N est la séquence d'informations grammaticales, qui signifie ici que l'entrée est un nom. Elle doit être séparée de la forme canonique par un point et peut être remplacée par un code propre à l'utilisateur.

bronchioles,bronchiole.N

FIG. 5.7 – Exemple d'entrée dans un dictionnaire Unitex.

5 Développement et fonctionnement de l'outil

La méthode que nous avons suivie pour développer MedCKARe compte plusieurs étapes (*cf.* figure 5.8) :

1. traiter l'ontologie au format OWL généré par PROTÉGÉ pour récupérer les concepts et extraire les relations qui serviront pour le codage médical,
2. construire notre propre dictionnaire de concepts dans Unitex,
3. enrichir le dictionnaire à partir des trois ressources lexicales,
4. créer les patrons lexico-syntaxiques projetés sur les CRH,
5. modéliser le thésaurus de spécialité pour le codage médico-économique,
6. identifier, dans les CRH, les informations pertinentes à coder.

Ces étapes sont décrites en détails dans les sous-sections suivantes. MedCKARe est un outil semi-automatique, les étapes 1 à 5 listées ci-dessus, initialisent le système et se font une fois pour toute. L'étape 6 est également automatique et se répète pour chaque nouveau CRH. Les fonctionnalités qui dépendent de l'utilisateur sont listées dans la section 6.3 de ce chapitre.

5.1 Récupération des données de l'ontologie

Nous savons qu'il est difficile dans un système de terminologie de retrouver le terme qui exprime fidèlement le sens associé au libellé d'un diagnostic. L'utilisation de l'ontologie va permettre de lever les ambiguïtés. Elle décrit les concepts d'un domaine et également les relations ou propriétés entre ces concepts. Les concepts (par exemple : *DeficitImmunitaire*) et les relations (par exemple : *localiséDans*) sont organisés sous forme hiérarchique. Chaque concept est désigné par un terme préféré qui correspond au terme employé par le médecin. Nous considérons que ce terme préféré est la forme canonique du terme. Ainsi, au label de concept *DeficitImmunitaire* est associé le terme préféré *déficit immunitaire*. Le langage OWL définit les informations de l'ontologie sous la forme de triplets (sujet - prédicat /verbe - objet). L'extrait présenté dans la fugure 5.9 nous en donne un exemple. Il s'interprète comme « la classe *DeficitImmunitaire* a pour terme préféré en français *déficit immunitaire* et est une sous-classe de la classe *PathologieImmunitaire* ». À partir de l'ontologie, nous avons pu extraire automatiquement les informations nécessaires pour la construction d'un dictionnaire, à savoir les termes préférés associés aux concepts de l'ontologie et leurs synonymes. À partir de l'expression OWL du tableau 5.9, il est possible d'obtenir le triplet *DeficitImmunitaire-label-déficit immunitaire* qui permet d'enrichir le dictionnaire avec le terme *déficit immunitaire*.

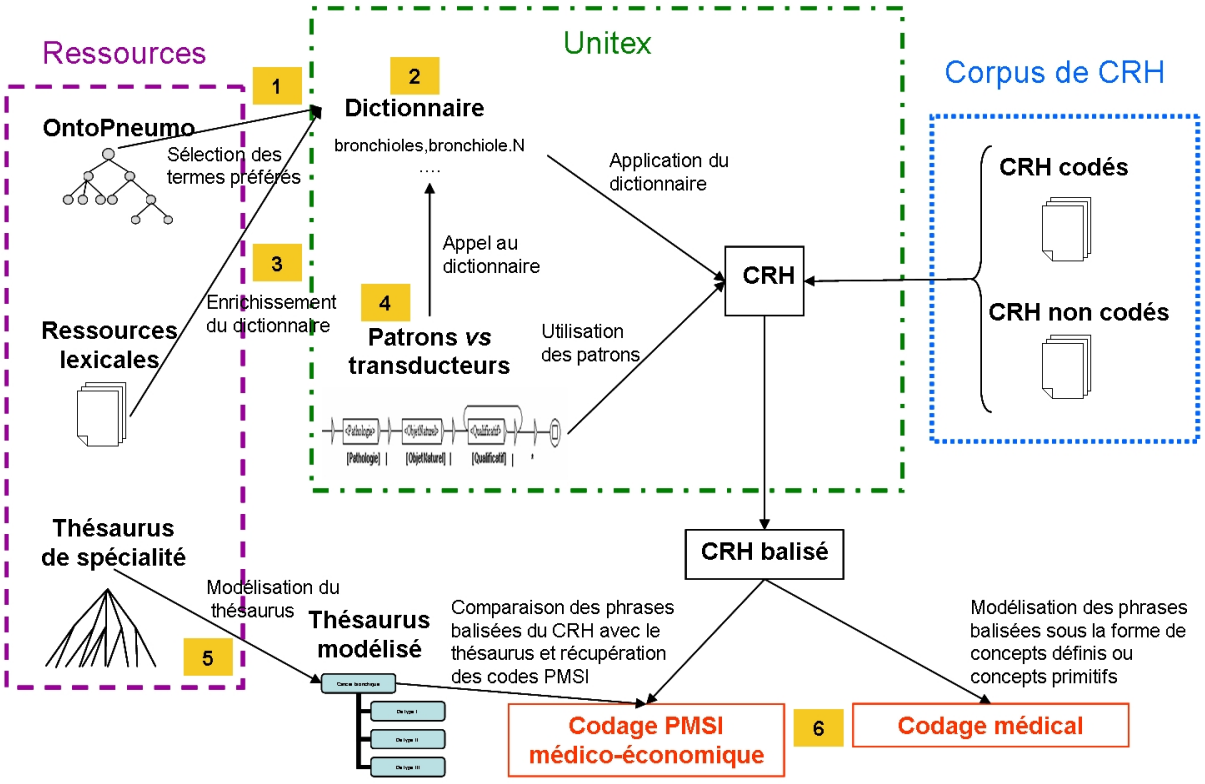


FIG. 5.8 – Schéma de fonctionnement de MedCKARe.

5.2 Construction du dictionnaire

Nous avons développé un programme qui permet de construire automatiquement le dictionnaire à partir de l'ontologie. Nous nous servons des principaux axes conceptuels du domaine qui ont été préalablement identifiés dans l'ontologie (*cf.* chapitre 4, section 4.2) : les actions médicales, les signes, les pathologies, les examens ... Nous nous servons également d'axes conceptuels de plus haut niveau, issus de la top-ontologie de MENELAS : ObjetNaturel, ObjetArtificiel ... Ce sont des axes de classification car la constitution de l'ontologie est telle qu'au sein d'un même axe sont regroupés des termes proches les uns des autres. Ainsi, pour la construction du dictionnaire, tous les termes subsumés par chacun de ces axes sont récupérés automatiquement, puis sont suivis d'un « ,. » comme l'impose le format d'un dictionnaire Unitex, et se voient affecter une catégorie qui correspond au label de l'axe auquel ils appartiennent (*cf.* figure 5.10).

tabagisme est le terme préféré associé au concept *Tabagisme* de l'ontologie et qui appartient à l'axe intitulé *OrigineDeLAffection*. Il est placé dans l'ontologie sous : *OrigineDeLAffection/Facteurs/Tabagisme*. Nous avons ensuite fait une analyse des tables lexicales afin d'enrichir le dictionnaire Unitex et ainsi augmenter le nombre de termes reconnus dans les CRH.

```
<owl:Class rdf:ID="DeficitImmunitaire">
  <rdfs:label xml:lang="TermFr">déficit immunitaire</rdfs:label>
  <rdfs:label xml:lang="TermAng">immune deficit</rdfs:label>
  <rdfs:subClassOf>
    <owl:Class rdf:ID="PathologieImmunitaire"/>
  </rdfs:subClassOf>
</owl:Class>
```

FIG. 5.9 – Extrait de l'ontologie de la pneumologie en OWL.

```
tabagisme,.OrigineDeLAffection
paroi antérieure,.ObjetAnatomique
paroi inférieure,.ObjetAnatomique
paroi,.ObjetAnatomique
avéré,.Qualificatif
acinetobacter,.AgentInfectieux
acquis,.Qualificatif
acquise,.Qualificatif
fièvre,.Signe
douleur,.Signe
douloureux,.Signe
cancer,.Pathologie
sonde,.ObjetArtificiel
```

FIG. 5.10 – Extrait du dictionnaire Unitex construit à partir d'OntoPneumo.

duodénal,.ObjetAnatomique
duodénale,.ObjetAnatomique
duodénum,.ObjetAnatomique

FIG. 5.11 – Extrait du dictionnaire Unitex après enrichissement par les ressources lexicales.

5.3 Traitement et utilisation des ressources lexicales

Les ressources lexicales (table de dérivation, de synonymes, de composition et de flexions) sont regroupées pour n'en faire qu'une. Ces tables comprennent deux champs : la forme canonique d'un mot et sa forme associée, qui peut être soit sa forme fléchie, soit sa forme dérivée, soit son synonyme. Dans ces tables, seules les formes canoniques des termes qui sont présents dans l'ontologie nous intéressent, leurs formes associées sont récupérées pour compléter le dictionnaire. Nous avons par exemple la forme fléchie de *duodénal* : *duodénal|duodénale*, et également la forme dérivé de *duodénum* : *duodénum|duodénal*. Puisque *duodénum* est le terme préféré du concept *Duodenum* présent dans l'ontologie, placé sous *ObjetAnatomique/ElementAnatomique/Organe/Duodenum*, *duodénal* et *duodénale* (ses deux formes dérivées) se voit alors affecter la catégorie à laquelle appartient *duodénum*, c'est-à-dire la catégorie *ObjetAnatomique*. Le tableau 5.11 présente un extrait du dictionnaire enrichi.

Nous avons également enrichi la table des lexiques. Si l'on trouve *rectal* comme dérivé du terme *rectus* mais que celui-ci n'est pas présent dans l'ontologie alors nous affectons à *rectal* une autre forme canonique qui elle sera contenue dans l'ontologie, par exemple *rectum*.

5.4 Mise au point des patrons lexico-syntaxiques

Les patrons que nous avons construits sous forme de transducteurs permettent de définir les différentes formes syntaxiques des phrases qui représenteraient des séquences informatives dans un CRH en pneumologie. Il a été nécessaire dans un premier temps d'étudier les CRH afin de pouvoir observer ces séquences. Cette observation a permis de schématiser le contexte lexical et syntaxique des unités lexicales en relation et de construire une synthèse de ce contexte sous la forme d'un patron lexico-syntaxique. Le patron est modélisé sous forme de transducteur (*cf.* figure 5.12), où un nœud du transducteur représente soit un code grammatical (comme DET) soit une des catégories qui a été associée à des entrées de notre dictionnaire (comme ObjetAnatomique) et qui correspond en réalité à un axe de l'ontologie. Une transition faisant appel à une catégorie est spécifiée comme suit : <catégorie>. Et c'est ainsi que <ObjetAnatomique> fait référence à toutes les entrées du dictionnaire ayant pour catégorie *ObjetAnatomique*, c'est-à-dire *duodénal*, *duodénum*, *paroi*, *poumon* ... Pour un état du transducteur faisant appel à une entrée du dictionnaire, nous spécifions entre crochets une sortie portant la catégorie de cette entrée, cela permettra, lors du traitement des instances, de connaître et de récupérer la catégorie du terme identifié. Afin de pouvoir traiter et analyser par la suite les instances repérées, il a fallu délimiter chaque entrée du dictionnaire et, pour ce faire, elle a été suivie d'un nœud vide dont la sortie est le symbole 'l'. À la fin de chaque patron, il a également fallu ajouter un délimiteur de patron, le

symbole '*'. Les résultats sont répertoriés dans un fichier texte où les instances de concepts sont mises en évidence dans le contexte de la phrase. L'expression *Adénopathie cervicale droite* est une instance du patron <Pathologie><ObjetNaturel[16]><Qualificatif>*. La figure 5.13 représente ce patron sous la forme d'un transducteur. La séquence balisée obtenue par ce transducteur est : « [Pathologie] adénopathie | [ObjetNaturel] cervicale | [Qualificatif] droite | * ».

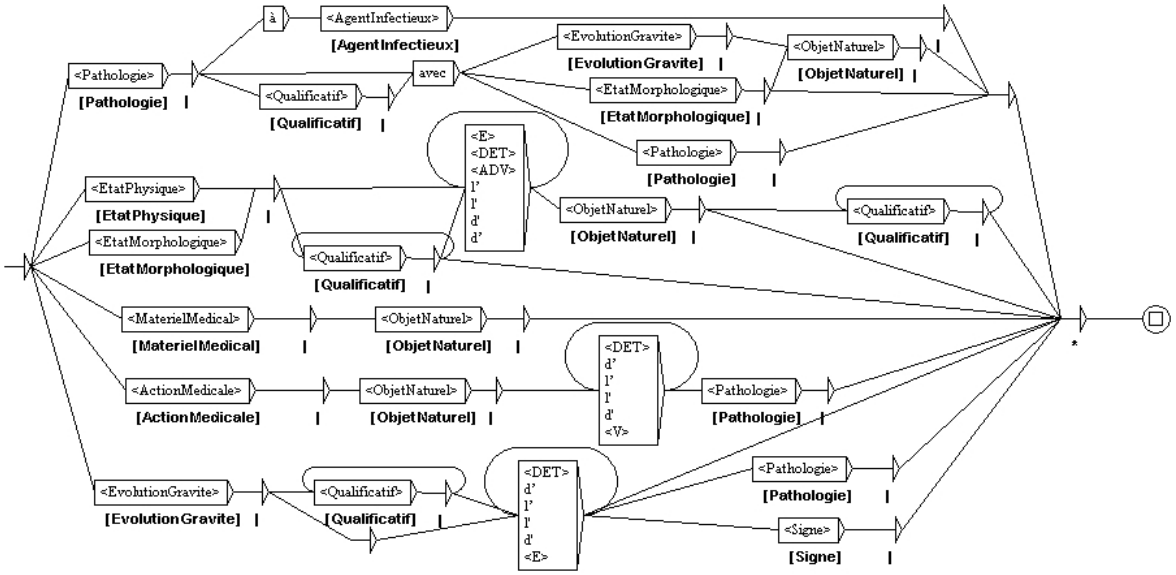


FIG. 5.12 – Exemple de transducteur dans Unitex.

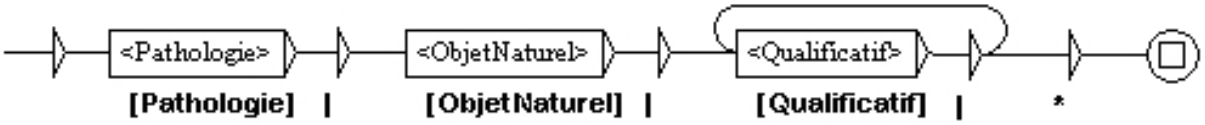


FIG. 5.13 – Exemple de transducteur dans Unitex qui reconnaît l'expression « adénopathie cervicale droite ».

5.4.1 Gestion des relations pour le codage médical

Un concept défini peut être composé de deux ou plusieurs concepts primitifs reliés entre eux par une ou plusieurs relations. La plupart du temps, une expression est modélisée par un concept défini. Cette combinaison sous-entend la présence de relations entre ces concepts, même si elles ne sont pas explicitées. Une définition explicite de ces liens doit donc être faite pour le

[16]ObjetNaturel est la catégorie d'ObjetAnatomique.

codage médical. Ces liens vont être définis dans l'ontologie. La formation des concepts définis provient d'un mode de composition des concepts et de relations régies par des contraintes. Ainsi, l'expression *hypertension artérielle pulmonaire* est modélisée par le concept défini suivant : *HypertensionArterielle-ObserveAuNiveauDe-Poumon* que nous faisons apparaître sous la forme de termes préférés (forme préférentielle, *cf.* figure 5.14). Sachant que *hypertension artérielle*

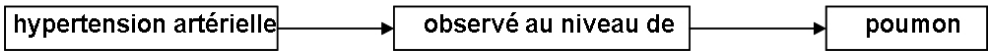


FIG. 5.14 – Représentation du concept défini *HypertensionArterielle-ObserveAuNiveauDe-Poumon*.

appartient à l'axe *Pathologie*, et *poumon* à l'axe *ObjetNaturel*. Le concept défini respecte la contrainte *Pathologie-ObserveAuNiveauDe-ObjetNaturel* avec l'unicité du lien existant entre les deux axes *Pathologie* et *ObjetNaturel*. L'étude de la composition de ce syntagme nominal nous a conduit à construire le patron <Pathologie><ObjetNaturel> puis son transducteur dans Unitex qui permettra de retrouver l'instance *hypertension artérielle pulmonaire* présente dans le CRH. Dans tous les cas, l'apparition consécutive de deux concepts appartenant à chacun de ces deux axes signifie qu'il y a une forte probabilité que ces deux axes aient un lien. Nous savons alors que ce patron n'entraînera pas de bruit. Un axe peut faire l'objet d'un patron qui a pour seul composant le label de l'axe. Il peut en effet être retrouvé dans la phrase sans qu'il y ait un lien avec les autres termes de la phrase. Par exemple, le patron <Pathologie>, avec lequel nous pouvons obtenir le terme *hypertension artérielle* désignant le concept *HypertensionArterielle* qui, lui, appartient à l'axe *Pathologie*, et qui peut figurer seul dans le CRH sans lien avec les termes qui l'entourent, par exemple : *Hypertension artérielle*, *arthrose*, et *dyspnée*. Nous avons construit environ 50 patrons.

5.4.2 Gestion de la négation

La prise en compte de la négation et son traitement est une étape importante et délicate. Elle permet d'exclure de l'étape de codage médico-économique, les pathologies ou les symptômes qui ont été éliminés par le pneumologue et qui, dès lors, figurent dans les CRH sous forme négative. Sans faire un traitement inférentiel de la négation, nous nous plaçons dans un monde fermé où tout ce qui est dit est vrai et seulement cela. À partir de l'étude des CRH, nous avons essayé de retrouver un ensemble de marqueurs de négation, c'est-à-dire des mots ou expressions révélateurs de négation (absence de, pas de consommation de, pas de), afin de les préciser dans les patrons d'extraction. La figure 5.15 est un extrait du transducteur que nous avons construit pour gérer ces négations. Il permet d'étiqueter les phrases et expressions négatives avec l'étiquette [negation], ce qui facilite nos traitements ultérieurs. Ce transducteur a permis notamment de reconnaître la phrase suivante : *pas de syndrome de Raynaud*, qu'il a balisé ainsi : [negation] pas de [Pathologie] syndrome de Raynaud | *.

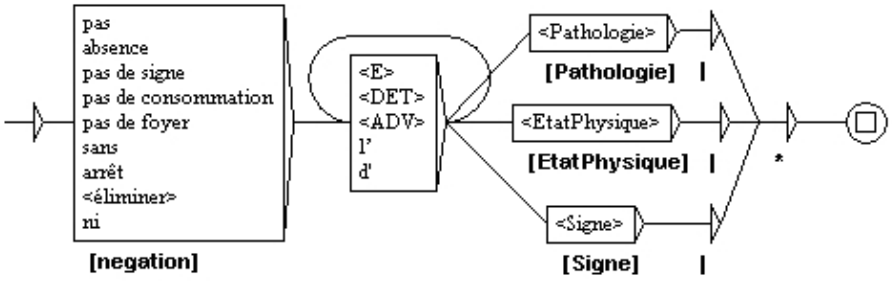


FIG. 5.15 – Transducteur Unitex pour la négation.

5.5 Modélisation du thésaurus pour le codage médico-économique

Le thésaurus de spécialité contient 337 expressions dans la partie dédiée au codage médico-économique des diagnostics. Les termes qui sont employés par les pneumologues dans les CRH et qui doivent aboutir à des codes PMSI, sont recensés dans le thésaurus. Il est donc nécessaire de tous les modéliser. Par modélisation, nous entendons au sens de notre présent travail, tout traitement appliqué à la donnée issue du thésaurus en vue de la représenter sous sa forme préférentielle. Chaque expression du thésaurus (par exemple : *dilatation des bronches*) est composée de termes ayant un lien avec les concepts de l'ontologie. Chacun de ces termes peut être :

- soit la forme dérivée d'un concept primitif contenu dans l'ontologie,
- soit sa forme fléchie,
- soit son synonyme,
- soit son terme préférentiel.

Or, pour pouvoir reconnaître les expressions en langue naturelle que nous devons coder dans le CRH, le concept défini modélisant chaque expression du thésaurus doit être composé de termes uniquement sous leur forme préférentielle, celle donnée dans l'ontologie par les termes préférés. Lorsque tous les termes de l'expression sont reconnus, le système accède au terme préféré associé à chacun des termes et en conserve la liste. Cette phase est essentielle. Non seulement elle permet de résoudre la reconnaissance de synonymes ou termes analogues, mais elle permet également d'associer des groupes de termes (composant une expression) à des termes préférés. Une première réduction est donc effectuée pour passer du large éventail de termes au groupe, plus restreint, de termes préférés reconnus. Les tableaux 5.16 et 5.17 montrent que quelle que soit la combinaison de termes choisie pour former une expression, sa représentation sous forme de termes préférés reste identique. Par conséquent, chacun de ces termes du thésaurus qui équivaut à la forme fléchie, la forme dérivée ou le synonyme, d'un terme préférentiel contenu dans l'ontologie, sera remplacé par le terme préférentiel en question. Cela permet au système de faire le lien entre les connaissances présentes dans le thésaurus de spécialité et l'ontologie du domaine qui agit alors comme pivot des connaissances. Ainsi, l'expression du thésaurus *tumeur maligne primitive du médiastin* est remplacée par *tumeur malin primitif médiastin*. Nous obtenons alors un nouveau thésaurus dans lequel toutes les expressions sont des regroupements de termes préférés.

Termes préférés			
A	B	C	D
tumeur	malin	primitif	médiastin

FIG. 5.16 – Termes préférés désignant des concepts de l'ontologie.

Expressions	Combinaisons de termes préférés
tumeur maligne primitive du médiastin	AxBxCxD
tumeur maligne primitive médiastinale	AxBxCxD

FIG. 5.17 – Représentation d'une expression par combinaison de termes préférés.

Un des points particulièrement délicat, tant pour l'ingénieur des connaissances que pour l'expert, a été de décider comment modéliser les différents termes qui constituent chacune des entrées du thésaurus de spécialité. Comme le montre la figure 5.18, l'expression « Mésothéliome pleural malin » a été précoordonnée[17] tandis que l'expression « Tumeur maligne secondaire de la plèvre » a été postcoordonnée[18]. Nous avons, autant que cela nous était possible, essayé de limiter la précoordination des éléments du thésaurus. Le fait de choisir une représentation précoordonnée ou postcoordonnée n'est pas anodin : quand on choisit une représentation précoordonnée, on peut être parfois plus près des modes d'expressions des professionnels mais dans notre cas, ce n'est pas très pertinent dans la mesure où ils n'auront accès à l'ontologie qu'à travers le thésaurus que l'on ne modifie pas de ce point de vue. Une représentation postcoordonnée est plus riche en termes de possibilités présentes et futures en permettant pas exemple une meilleure généralisation dans la recherche d'information. En revanche, plus on postcoordonne, plus l'information est granulaire, plus il y a de concepts, plus les temps de calcul liés à l'indexation et surtout à la recherche d'information sont long voire prohibitifs. Il faut donc trouver un équilibre en termes de représentation des connaissances et s'en tenir à la granularité – et donc la postcoordination – nécessaire.

L'analyse des termes du thésaurus permet également d'enrichir notre ontologie. En effet, dans l'étape de modélisation du thésaurus, les termes qui n'ont pas été reconnus par Unitex sont ceux qui sont absents de l'ontologie. Nous les prenons donc en compte pour enrichir l'ontologie. Cette procédure de modélisation du thésaurus sert pour attribuer des codes PMSI au CRH en cours de traitement dans MedCKARe. Pour garantir l'efficacité de la fonctionnalité de codage PMSI, nous avons évalué la couverture de l'ontologie par rapport au thésaurus. Les résultats de cette évaluation sont présentés dans la section 6.

[17]Principe suivant lequel les combinaisons entre les termes d'un langage s'effectuent au cours de son élaboration, par exemple la création des termes composés dans un thésaurus.

[18]Principe suivant lequel les combinaisons entre les descripteurs s'effectuent au cours de la recherche documentaire.

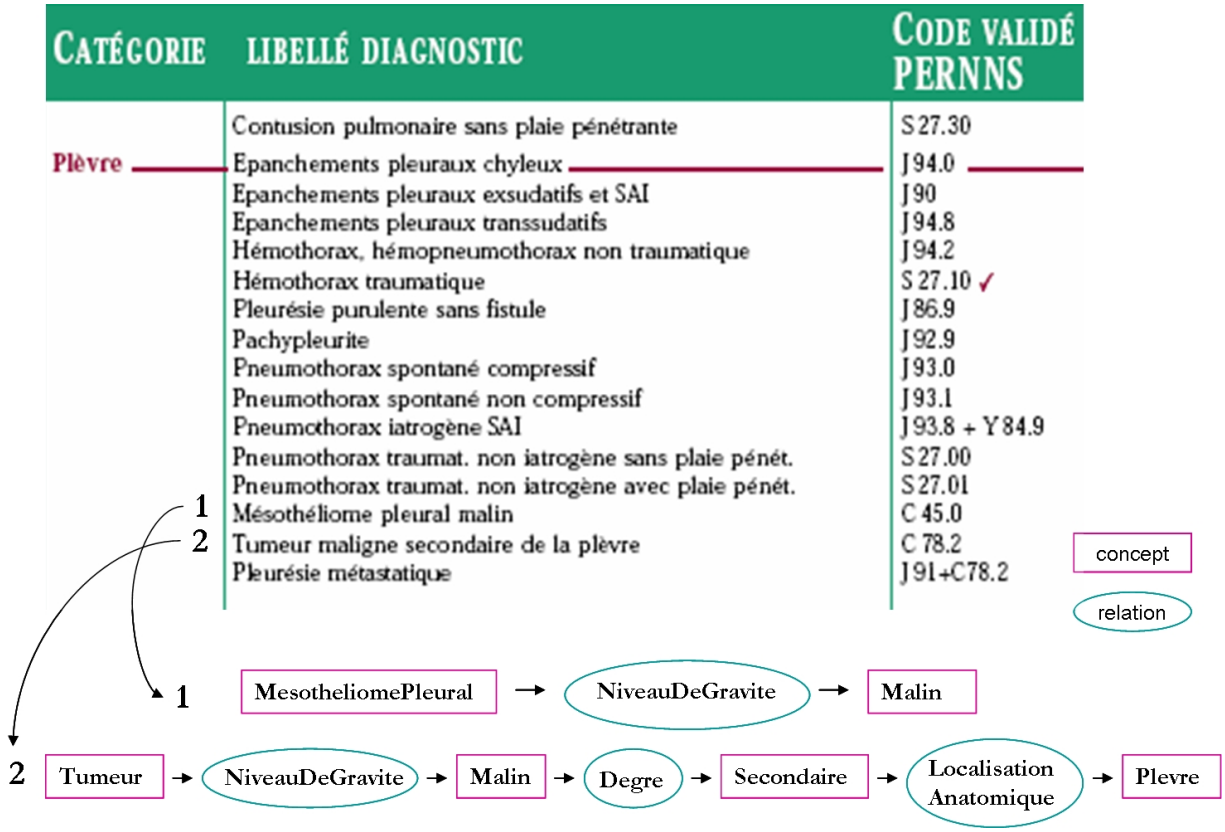


FIG. 5.18 – Modélisation de deux entrées du thésaurus de spécialité.

5.6 Identification des informations pertinentes pour le codage

Les instances de patrons repérées dans le texte sont traitées de façon à pouvoir, à la fois, les afficher graphiquement et générer leurs codes PMSI. Ainsi, pour une expression reconnue comme étant un diagnostic, l'interface affiche le concept défini décrivant cette expression et, si celle-ci correspond à une expression du thésaurus, le système propose au pneumologue le code PMSI associé.

5.6.1 Codage médical

Lors de l'analyse du CRH à l'aide d'Unitex, et précisément lors de l'application des patrons, chaque terme reconnu est précédé de sa catégorie. Nous recherchons alors la forme canonique, définie dans l'ontologie, de chacun des termes reconnus et des relations existant entre eux. Nous avons alors deux possibilités :

1. si le terme reconnu par le patron est une forme dérivée ou fléchie d'un terme préféré contenu dans l'ontologie, nous le retrouvons dans les tables lexicales et nous récupérons alors sa forme canonique ;

2. si le terme reconnu par le patron n'est pas retrouvé dans les tables lexicales alors c'est un terme préféré de l'ontologie et il est récupéré tel quel.

Le graphe correspondant au concept défini identifié par les patrons dans le CRH (*cf.* figure 5.19) est constitué d'une ou plusieurs relations existant entre le concept primitif droit et les autres concepts primitifs gauche qui le composent.

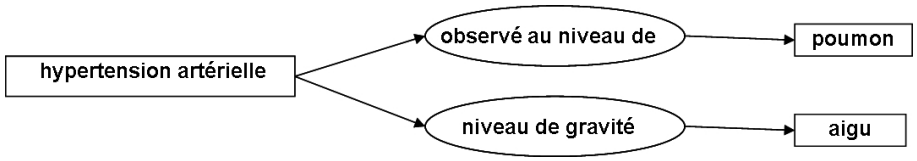


FIG. 5.19 – Graphe correspondant à l'expression « hypertension artérielle pulmonaire aiguë » affiché avec les termes préférés.

Les termes récupérés – *hypertension artérielle*, *poumon* et *aigu* – sont des descripteurs du CRH qui serviront, lors d'un prochain développement, pour la recherche d'informations. Le graphe 5.19 est visualisable dans l'interface par le pneumologue qui a également la possibilité d'ajouter une relation entre deux concepts ou de supprimer les concepts qui ne le satisfont pas. Par ailleurs, nous avons décidé que les termes d'une expression négative, qui sous-entendent des affections (pathologies, signes, symptômes) ayant été éliminées, vont être précédés, lors de leur le codage médical, de l'étiquette [pas de]. Par exemple l'expression *pas de consommation de tabac* n'est pas traitée au niveau du codage PMSI mais est représentée au niveau du codage médical.

5.6.2 Codage médico-économique

Chaque expression du thésaurus est composée d'une conjonction de termes préférés que nous projetons sur la conjonction de termes préférés obtenus par l'analyse du CRH. Le système compare les termes médicaux extraits du CRH avec ceux du thésaurus. Si dans le thésaurus modélisé nous avons la conjonction de termes *embolie pulmonaire* et que nous la projetons sur la conjonction de termes *embolie pulmonaire droit*, contenue dans le CRH, une correspondance est observée. Lorsque notre programme identifie un groupe de termes du CRH qui correspondent aux termes du thésaurus de spécialité modélisé, le code PMSI associé est alors récupéré et affiché dans l'interface utilisateur. Le pneumologue doit ensuite valider la sélection automatique des codes. Comme nous l'avons dit dans la section précédente, les séquences négatives détectées par nos patrons seront exclues de l'étape de codage PMSI. Nous prévoyons de développer, dans le cadre du projet MedOC , un système de préférences qui en fonction du profil utilisateur du pneumologue accordera plus de poids à certains codes plutôt qu'à d'autres.

6 Résultats

Nous présentons dans cette section les résultats de plusieurs évaluations.

1. La première évaluation consiste à vérifier la possibilité de construire une représentation conceptuelle de toutes les expressions du thésaurus de spécialité en combinant les concepts primitifs dont nous disposons dans l'ontologie.
2. Ensuite, à partir des résultats récupérés à la fin du traitement de chaque CRH, c'est-à-dire les résultats du codage médical et ceux du codage PMSI, nous avons établi un bilan quantitatif et un bilan qualitatif. Pour ce faire, nous avons divisé l'évaluation en deux phases :
   – une première évaluation est faite sur les résultats bruts des deux codages ;
   – une seconde évaluation est faite sur les résultats des deux codages après analyse et correction d'un certain nombre d'erreurs.

6.1 Résultats de la modélisation du thésaurus de spécialité

L'enrichissement manuel de l'ontologie effectué à l'aide des termes contenus dans le thésaurus garantit le fait que toutes les expressions du thésaurus sont modélisées. En effet, les 337 expressions que compte la partie PMSI du thésaurus sont constituées de termes qui ont un lien (de synonymie, de variances lexicales) avec les labels des concepts de l'ontologie. La couverture de l'ontologie par rapport au thésaurus est donc assurée.

6.2 Résultats qualitatifs et quantitatifs pour les deux types de codage

Ces résultats se calculent en terme de rappel et de précision. Sachant que le rappel est le rapport du nombre de réponses pertinentes trouvées au nombre total de réponses pertinentes et que la précision étant le rapport du nombre de réponses pertinentes trouvées au nombre total de réponses sélectionnées par l'outil, un rappel de 100% signifie que toutes les réponses pertinentes ont été trouvées et une précision de 100% que toutes les réponses trouvées sont pertinentes.

6.2.1 Résultats du codage médical

En ce qui concerne le codage médical, nous sommes capables de prévoir à peu près quels sont les termes médicaux pertinents qui devraient apparaître ainsi que leur nombre, et ce, en analysant le CRH et en retrouvant les termes médicaux importants. L'évaluation[19] a lieu sur l'ensemble de notre corpus de référence (*cf.* section 3.3), c'est-à-dire sur l'ensemble du corpus [NONCODE] et du corpus [CODE], soit 500 CRH au total. La figure 5.20 présente nos résultats sous la forme d'un histogramme. À la première étape de l'évaluation, le rappel pour le codage descriptif est de 43%. Afin d'améliorer ce chiffre, nous avons enrichi l'ontologie par les termes médicaux importants figurant dans le CRH et qui n'ont pas été reconnus, comme par exemple : *hémothorax*, *thymome*, *rhinite* ... Le rappel passe à 88% après améliorations et enrichissement de l'ontologie.

[19] L'évaluation du codage médical se fonde sur notre propre expertise car ce type de codage n'a, à notre connaissance, jamais été fait. Nous ne disposons donc pas de CRH codés médicalement.

En ce qui concerne la précision, elle est, lors de la première phase d'évaluation, de 70%. Après la correction d'un certain nombre d'erreurs et donc la diminution du bruit (par exemple, l'amélioration de la prise en compte de la négation), la précision passe à 85%.

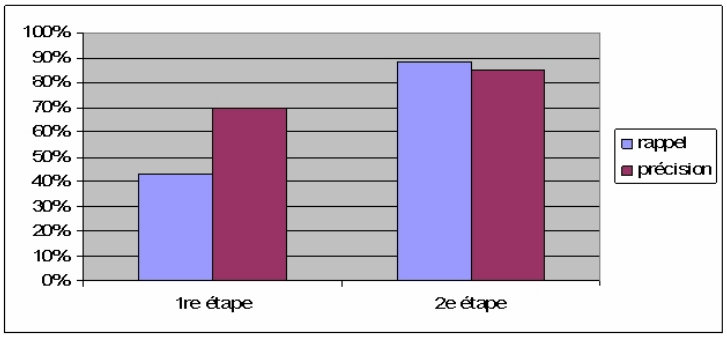


FIG. 5.20 – Histogramme des résultats obtenus pour le codage médical.

6.2.2 Résultats du codage médico-économique

Pour évaluer[20] le rappel sur le codage PMSI, nous avons utilisé le corpus [CODE] qui contient 100 CRH qui fournissent, d'une part les réponses pertinentes attendues et, d'autre part leur nombre. La figure 5.21 présente nos résultats sous la forme d'un histogramme. La première évaluation consiste à analyser les résultats obtenus lors de l'attribution des codes PMSI aux CRH et à les comparer avec les codes présents dans les CRH. Le rappel est de 25%. L'analyse de ces résultats nous a aidés à enrichir nos ressources (l'ontologie et le thésaurus). Nous avons tout d'abord recherché les termes associés aux codes PMSI, sachant que ces codes étaient déjà présents dans les CRH du [CODE]. Nous avons ensuite étudié les CRH afin de savoir à quel endroit ces termes apparaissent et pourquoi notre système ne les a pas détectés. Nous avons essayé de corriger certains problèmes, dus notamment à l'absence, au niveau de l'ontologie, de termes issus du thésaurus ou à d'autres problèmes cités dans la section 6.4. Après amélioration, le rappel est de 80%.

La précision est de 87% et reste à peu près la même pendant les deux phases d'évaluation. Ce résultat s'explique par le fait que, puisque tous les codes PMSI identifiés correspondent aux termes cités dans les CRH, ces codes sont généralement corrects. Les erreurs rencontrées sont souvent dues aux pathologies ou symptômes mentionnés dans les antécédents mais que le malade ne présente plus. En effet, ceux-ci sont tout de même codés par notre outil puisqu'ils apparaissent dans le CRH. Nous avons décidé de régler ce problème de choix des codes en offrant à l'utilisateur la possibilité de sélectionner parmi les codes identifiés ceux qui lui semblent être nécessaires.

[20] L'évaluation du codage médico-économique est fondée sur un ensemble de CRH codés par des médecins pneumologues.

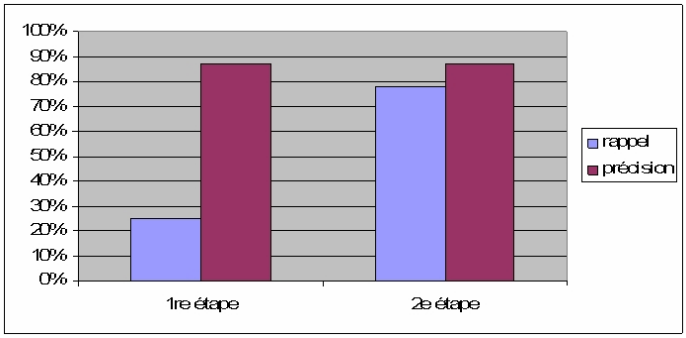


FIG. 5.21 – Histogramme des résultats obtenus pour le codage médico-économique.

6.3 Interface utilisateur

L'interface est, nous semble t'il, facile d'utilisation et offre au pneumologue un certain nombre de fonctionnalités :

- charger un compte rendu,
- visualiser les expressions identifiées et présélectionnées par le système lors du chargement du CRH,
- visualiser le codage médical et le codage PMSI associé au CRH chargé,
- choisir les codes PMSI souhaités parmi ceux proposés automatiquement par le système,
- intervenir sur le codage médical,
- enregistrer le CRH traité ainsi que les représentations issues du codage médical et les codes PMSI validés par le pneumologue.

Dans la figure 5.22, nous voyons que :

1. les informations extraites à l'aide de patrons lexico-syntaxiques sont présélectionnées par le système et surlignées en jaune ;
2. le codage médical associé au CRH est représenté grâce aux termes préférés des concepts primitifs et définis représentant les informations présélectionnées ;
3. les codes PMSI déduits sont affichés et le médecin peut cocher ceux qu'il juge pertinents ;
4. le médecin a la possibilité d'ajouter ou de supprimer des relations et des concepts pour adapter la représentation des connaissances du CRH à son expertise.

Les avantages d'une telle interface sont de permettre au médecin de visualiser rapidement les diagnostics fait sur un patient et la codification médico-économique associée.

6.4 Problèmes à résoudre et pistes d'amélioration

Nous proposons ci-dessous quelques remarques concernant les erreurs – bruit ou silence – produites par notre système. Quand cela est possible, nous proposons des pistes d'améliorations qui seront prises en compte dans la prochaine version de MedCKARe.

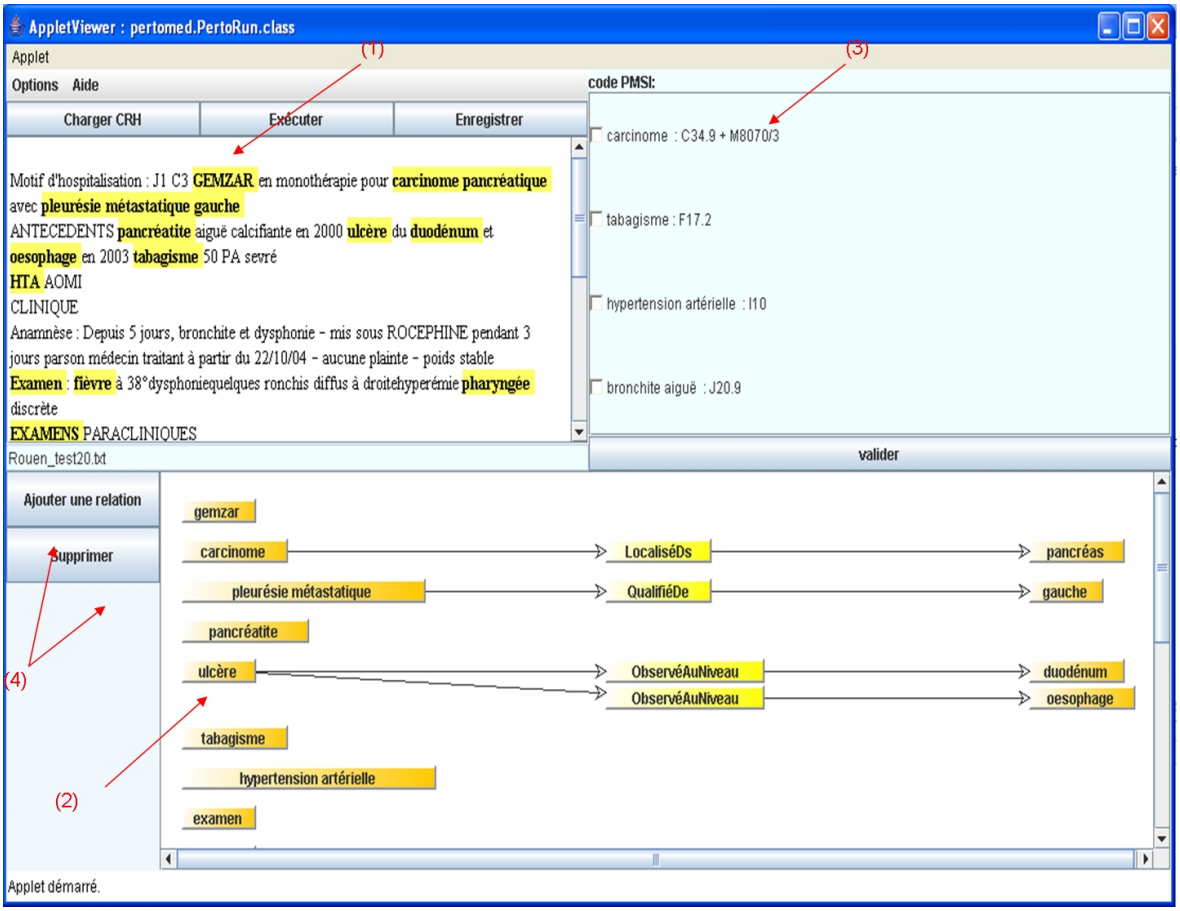


FIG. 5.22 – Copie d'écran de l'interface utilisateur de MedCKARe.

6.4.1 Insuffisances terminologiques

Les insuffisances terminologiques se traduisent par une lacune observée au niveau de nos ressources. Au niveau de l'ontologie, on note une absence de termes préférés, et donc de concepts qui, pourtant, sont souvent retrouvés dans les CRH. En ce qui concerne le thésaurus, l'insuffisance se caractérise par un manque de relations qui pourraient exister entre des termes, telle que la relation de synonymie, de spécialisation ou de généralisation.

Manque de concepts dans l'ontologie

Un certain nombre de concepts présents dans les CRH sont absents de l'ontologie et ne peuvent donc pas être identifiés dans les corpus. Chacune des deux phases de l'évaluation nous a permis de découvrir de nouveaux concepts. L'ontologie est essentiellement une ontologie de la pneumologie. Dans une expression, certains termes qui ne concerne pas la pneumologie, mais qui sont toutefois présents dans les CRH, se retrouvent non codés. Ce problème doit être étudié et corrigé en réfléchissant à la prise en compte des termes qui ne sont pas de la spécialité. Dans le cas de termes spécifiques au domaine et manquants, la

solution est simple et consiste à compléter l'ontologie. Ainsi, dans l'instance *recollement du [ObjetNaturel] parenchyme |* [ObjetNaturel] pulmonaire |*, nous pouvons voir que *parenchyme* et *pulmonaire* sont deux termes qui sont reconnus par le système et qui font bien partie du domaine de la pneumologie, tandis que *recollement* est manifestement un terme pertinent qui n'a pas été détecté. Cette erreur doit être prise en considération car le sens de la phrase est faussé tant que *recollement* n'est pas reconnu. La solution est donc d'ajouter le concept *recollement* à l'ontologie.

Synonymie non prise en compte

Nous avons rencontré des problèmes d'insuffisance terminologique au niveau du thésaurus. Par exemple, le terme *métastase* peut être considéré comme synonyme de *tumeur secondaire* et ne figure pas dans le thésaurus de spécialité. Or *métastase* est plus souvent employé dans les CRH. Nous avons ajouté au thésaurus l'expression *métastases pleurales*, avec pour code médico-économique celui de *tumeur secondaire de la plèvre*. Il en est de même pour tous les autres types de tumeurs secondaires.

Lien de spécialisation ou de généralisation non établi dans le thésaurus

Par ailleurs, nous avons également ajouté au thésaurus les termes qui ont un lien de spécialisation avec ceux déjà présents. Par exemple, l'expression *bruits du cœur assourdis* est une sorte d'*arythmie*, terme présent dans le thésaurus. L'expression est alors ajoutée dans le thésaurus et portera donc le code d'*arythmie*. Le terme *pleuropneumopathie*, qui ne figure pas dans le thésaurus, est un terme médical qui comprend les épanchements pleuraux, il doit être par conséquent ajouté au thésaurus et doit avoir pour code celui d'*épanchement pleural* qui, lui, est déjà présent dans le thésaurus. Comme dans le point précédent, un ajout de ce type doit bien entendu être validé, dans un premier temps, par un médecin puis, dans un deuxième temps, par la Société de pneumologie de langue française, responsable du thésaurus.

6.4.2 Manque d'expertise médicale

L'expertise médicale est, dans le cadre de ce travail, la compétence permettant de déterminer la nature d'une maladie d'après les renseignements donnés par le malade, ou l'étude des signes et symptômes. Le système, contrairement au médecin, ne peut pas déduire le code PMSI d'une maladie à partir de la liste de ses signes. Comme il procède par analyse syntaxique, la présence du nom de cette maladie dans le CRH lui est indispensable.

Relations signe-pathologie

Nous avons étudié dans l'ensemble du corpus [CODE] les termes reliés aux codes PMSI qui devaient être retrouvés par le système mais qui ne l'ont pas été. Nous avons remarqué que certains de ces termes ne figuraient pas en tant que tel dans le CRH mais qu'ils avaient été déterminés par le médecin codeur d'après les signes mentionnés dans le CRH. Ces termes correspondent donc aux signes d'une pathologie qui est présente dans le thésaurus. Par exemple, le code de pneumopathie, J18.9, figure dans un des CRH codés alors que le terme *pneumopathie* n'est pas mentionné dans ce CRH. Dans un cas comme celui-ci, notre système n'est pas capable de produire le codage médico-économique. Si nous

étudions le CRH « à la main », nous sommes capables de voir que l'expression *murmure vésiculaire diminué* est un signe de *pneumopathie*, on comprend alors que c'est pour cette raison que le médecin affecte le code J18.9 de pneumopathie au CRH. Une solution à ce problème consiste à mettre en œuvre un système d'inférences pour relier les signes et les symptômes aux pathologies. Cette amélioration est cependant difficilement réalisable car elle nécessite une réflexion de fond sur les inférences acceptables par les médecins. Il faut prendre garde ici à ne pas répéter les erreurs du passé où des systèmes experts inféraient trop de conclusions et avaient, de ce fait, un comportement inacceptable par les praticiens.

Autres relations

Il existe également des termes qui peuvent se combiner donnant ainsi une seule expression portant le sens de ces deux termes. Par exemple, un malade qui présente de l'asthme et de l'allergie, souffre en fait d'un asthme allergique. Dans ce type de cas, notre outil va coder les deux termes *asthme* et *allergie* séparément, alors que le code approprié serait celui d'*asthme allergique*.

6.4.3 Séquences non identifiées

Des séquences n'ont pas été identifiées car elles n'ont pas été modélisées par un de nos patrons d'extraction. Pour pallier ce problème, il suffit de construire le patron modélisant la structure de la phrase qui n'a pas été reconnue.

7 Perspectives et conclusion

L'enjeu de ce travail a été de montrer que le codage des connaissances médicales pouvait s'effectuer efficacement à l'aide d'un outil informatique, MedCKARe , faisant intervenir des outils de Traitement automatique des langues ainsi que des ressources sur la terminologie et les connaissances du domaine médical concerné. Par ailleurs, dans un domaine particulier tel que la pneumologie, on ne peut envisager de réaliser des outils informatiques pour l'aide au codage qu'en intégrant préalablement une structure permettant d'organiser les objets du domaine en fonction de la tâche à réaliser. Cette structure est l'ontologie du domaine et nous a permis d'optimiser le rappel. Les phrases pertinentes contenues dans les CRH sont identifiées grâce aux concepts de l'ontologie et sont ensuite mises en correspondance avec les concepts issus de la modélisation du thésaurus de spécialité. Bien que très simple, le patron que nous avons construit pour gérer les négations nous a permi de détecter la majorité des occurrences de phrases négatives. La détection de la négation est une étape primordiale dans l'extraction des informations et il nous faudra encore améliorer ces patrons pour éviter de coder, notamment, les antécédents. Grâce aux patrons implémentés dans Unitex, les processus d'identification et d'extraction des expressions pertinentes se sont montrés très efficaces quant à la précision des résultats attendus. Le fait d'avoir intégré des ressources lexicales dans le système nous a permis d'avoir plusieurs entrées pour un même terme. Cela accroît le nombre de termes que nous sommes en mesure d'identifier et permet de représenter avec plus de détails les informations relatives aux diagnostics faits sur le patient. Les résultats de MedCKARe sont satisfaisants d'un point de vue qualitatif

et quantitatif, notamment par rapport à ceux qu'obtiennent S. Pereira *et al.* (2006). Cependant, des améliorations peuvent et doivent être apportées au niveau du codage médical. Nos directions de recherches et de développements sont valides aux vues des résultats obtenus pour l'instant. L'outil MedCKARe en est encore au stade de prototype et nécessite des améliorations tant linguistiques qu'informatiques avant de pouvoir envisager un fonctionnement en routine. L'efficacité réelle de l'outil sera à prouver dans le futur.

Pour améliorer le codage médical, nous devons améliorer la gestion des relations qui, pour l'instant, est plus que sommaire dans le système. Ainsi, les relations entre les concepts doivent être précisées et définies de manière plus spécifique. En raffinant la sélection des axes conceptuels associés aux entrées qui composent notre dictionnaire nous pourrons enrichir nos représentations. Ainsi, au lieu de voir apparaître une même relation *AuNiveauDe* à la fois dans le concept défini *dissection-AuNiveauDe-aorte*, instanciant <ActionMedicale><ObjetNaturel> et modélisant le syntagme *dissection aortique*, et à la fois dans le concept défini *fibroscopie-AuNiveauDe-bronche*, modélisant *fibroscopie bronchique*, nous aurions pu établir différentes relations entre ces concepts. Par exemple, nous aurions pu attribuer au concept *Dissection* (ActionMedicale/ProcedureTechnique/Dissection) la catégorie de l'axe *ProcedureTechnique* au lieu de lui attribuer la catégorie *ActionMedicale*. Nous aurions alors pu définir que la relation reliant *ProcedureTechnique* et *ObjetNaturel* serait *RealiseSur*, d'arité 2. Le concept défini ainsi obtenu serait : *dissection-RealiseSur-aorte*. Par ailleurs, le terme *Fibroscopie*, qui est un examen et qui est placé dans l'ontologie sous ActionMedicale/Examen/ExamenParaClinique/Imagerie/Endoscopie/, aurait été associé à l'axe *Examen* à la place d'*ActionMedicale*. Ces modifications se révèleront sans doute nécessaires pour la future tâche d'indexation des CRH. Elles vont nécessairement accroître la complexité des traitements et donc le temps de calcul, mais elles permettront à l'outil de faire un bond qualitatif. Nous les implémenterons dans le cadre du projet MedOC et de la version 2 de MedCKARe. L'outil sera mis à disposition du corps médical courant 2008 dans le cadre du projet DaFOE4App . Il passera des tests en routine dans deux hôpitaux de l'Assistance Publique-Hôpitaux de Paris auprès de pneumologues appartenant à la SPLF afin d'évaluer d'une part l'utilité et l'utilisabilité de l'outil dans le contexte d'usage et d'autre part l'ergonomie de l'interface.

Nous avons vu que plus le processus de construction de modèle se dote d'outils qui améliorent les processus de construction à partir de textes, plus il est efficace de revenir des modèles vers les textes. Il y a un cycle intéressant à exploiter entre la construction de l'ontologie à partir de textes et son utilisation, au cœur d'un système informatique, pour identifier des expressions dans des textes. Ainsi, les relations trouvées à l'aide des patrons lexico-syntaxiques dans la phase de construction peuvent être de bonnes pistes pour bâtir des concepts définis pour le codage médical. Réciproquement, les relations identifiées dans les patrons Unitex peuvent apporter un plus lors de la construction de l'ontologie.

La réalisation de MedCKARe nous a également permis de montrer et de mesurer l'impact de l'application cible sur le contenu du modèle et sur la manière de le construire. Le travail d'adaptation et celui de prise en compte des besoins requis pour intégrer l'utilisation de l'ontologie dans l'application cible est loin d'être négligeable.Il serait intéressant de caractériser plus précisément et de manière plus générique ce travail supplémentaire d'adaptation. Dans le même ordre d'idée, il faudrait spécifier les compétences nécessaires à l'articulation entre textes, terminologies et

connaissances.

CHAPITRE 6

Évaluation, évolution et maintenance d'une ontologie en médecine

« Oui, je me demande parfois si l'homme, tout bien pesé, n'a pas fait faire à la connaissance un énorme pas en arrière en renonçant à l'imagination et à la poésie comme moyens d'investigation scientifique... »
Jean Anouilh
L'Hurluberlu (1959)

Ce chapitre s'intéresse aux questions, étroitement liées, d'évaluation, d'évolution et de maintenance des ressources termino-ontologiques (RTO) dans le domaine médical. Dans la section 2, nous proposons un état de l'art des critères d'évaluation des RTO. Dans la section 3, nous décrivons notre expérimentation concernant l'évolution et la maintenance de notre ontologie de la pneumologie et du thésaurus de spécialité. Enfin, dans la section 4, nous proposons des directions de recherche pour l'évaluation et la maintenance de RTO et, particulièrement, d'ontologies médicales.

1 Introduction

Les ressources terminologiques ou ontologiques en médecine ont la spécificité d'être nombreuses, d'abord parce qu'il existe de nombreux thésaurus, ensuite parce que les réflexions et les développements sur les ontologies sont très avancés dans le milieu médical.

La production de documents électroniques en tout genre a fait apparaître la nécessité de ressources terminologiques pour produire, diffuser, rechercher, exploiter et traduire ces documents. On voit ainsi apparaître de nouveaux types de ressources terminologiques ou ontologiques adaptées aux nouvelles applications de la terminologie en entreprise et particulièrement dans les établissements de soins : thésaurus pour les systèmes d'indexation, pour des motivations

bibliographiques, médicales ou médico-économiques, terminologies de référence pour l'aide à la rédaction, ontologies pour l'indexation, la gestion des connaissances, les systèmes d'aide à la décision ou pour les systèmes d'extraction d'information, glossaires de référence et liste de termes pour les outils de communication interne et externe, etc.

En dehors du fait que la multiplication des types de ressources terminologiques met à mal le principe théorique de l'unicité et de la fixité d'une terminologie pour un domaine donné (Slodzian, 1999; Bourigault *et al.*, 2004), cela nous amène à réfléchir à la construction de ces ressources et aux relations qu'elles entretiennent avec les corpus, pendant leur élaboration et, ensuite, leur maintenance. De plus, nous verrons spécifiquement en médecine, les relations qu'entretiennent entre eux les thésaurus et les ontologies, en termes de spécificité, de validation et d'évolution.

Par ailleurs, le besoin d'ontologies et leur développement amènent à réfléchir à des méthodologies efficientes et reproductibles. Dans cette optique, les travaux de B. Bachimont (2000) et du groupe Terminologie et Intelligence Artificielle ont mené à la réalisation de la méthodologie ARCHONTE. Comme nous l'avons présenté dans le chapitre 4, nous avons mis au point et expérimenté une méthode, adaptée d'ARCHONTE, pour développer des ontologies à partir de corpus. Notre méthode permet de fonder la construction et la structuration d'ontologies médicales sur les corpus de textes écrits par les professionnels considérés. Nous voulons, dans ce chapitre, étudier la question de la validation et de la maintenance des ontologies médicales au regard des méthodologies de construction mises en œuvre, particulièrement par rapport aux corpus textuels de référence. Nous ne pouvons manquer de citer le rapport de l'action spécifique « Corpus et terminologie » qui met noir sur blanc les résultats des débats qui ont animé le groupe sur la plupart de ces questions, en particulier les corpus, les genres textuels et leurs liens avec les ontologies (Aussenac-Gilles & Condamines, 2003).

2 Critères pour évaluer une RTO

L'évaluation des RTO, tout comme les réflexions menées autour des questions de leur évolution et de leur maintenance, est un domaine relativement récent. Des critères spécifiques à l'évaluation de ressources de type ontologique apparaissent peu à peu dans les communautés de recherche liées à la terminologie et l'Ingénierie des connaissances. Ici, il nous semble primordial de lier évalution, évolution et maintenance car nous pensons qu'il ne peut y avoir d'évolution sans évaluation de la ressource en amont. Dans cette section, nous voulons dresser un rapide état de l'art des travaux existants et plus particulièrement des critères permettant d'évaluer les ressources ontologiques.

L'évaluation d'une RTO peut avoir plusieurs objectifs : effectuer un bilan d'ensemble, guider la prochaine étape de construction, découvrir l'origine de difficultés, rendre compte de résultats … Il est clair que ces objectifs peuvent s'opposer et qu'ils doivent être organisés. La question n'est pas simple et de nombreuses communications indiquent la nécessité d'évaluer les RTO (Bourigault & Aussenac-Gilles, 2003), leur méthode de construction aussi bien que leur utilité, sans toujours préciser comment cette tâche peut être menée à bien. Faut-il envisager une méthode qualitative, quantitative, mixte ? Faut-il évaluer toutes les étapes de construction ou bien

seulement le résultat obtenu ?

Ces dernières années, des progrès considérables ont été accomplis dans les domaines du partage et de la réutilisation des connaissances pour des ressources telles que les RTO. Les ontologies font partie de ces systèmes et indiquent généralement des conceptualisations ayant un degré élevé de formalisation. Cependant, leur processus de construction est encore trop souvent personnel : chaque « ontologue » suit son propre ensemble de principes de conception et d'étapes dans le procédé de développement. Partant de ce constat, l'évaluation des ontologies est, au mieux, exécutée différement dans chaque cas et il en va de même des protocoles d'évolution et de maintenance. C'est pourquoi il est intéressant de dresser un état de l'art des critères d'évaluation les plus représentatifs, tant du point de vue quantitatif que qualitatif.

Une ontologie, à l'instar d'une connaissance ou d'un logiciel, n'est en rien figée et possède un cycle de vie propre (*cf.* figure 3.7, page 38). Il existe donc plusieurs niveaux d'évaluation d'une telle ressource, chacun contenant ses critères (Tarhuni *et al.*, 2005) :

1. évaluation des ressources servant à construire l'ontologie. Nous nous intéresserons spécifiquement aux corpus textuels,
2. évaluation du contenu de l'ontologie,
3. évaluation de la qualité taxinomique de l'ontologie,
4. évaluation de l'ontologie en situation, c'est-à-dire dans son contexte d'usage, et de sa réutilisabilité.

Pour chacun de ces quatre niveaux d'évaluation, nous allons essayer d'exprimer les critères qui nous semblent les plus pertinents dans la littérature.

2.1 Élaboration et évaluation des corpus textuels

Dans un projet de construction d'ontologies à partir de textes, le corpus, son statut et sa collecte sont d'une importance primordiale à la fois comme source de connaissances pour construire le modèle et comme source de référence tout au long du processus d'élaboration. Comme le souligne Didier Bourigault et Nathalie Aussenac-Gilles (2003), *« Dans ce domaine, il n'est hélas pas encore possible de définir à priori des instructions méthodologiques très précises pour encadrer la tâche de sélection des sources textuelles qui viendront constituer le corpus »*. Pourtant, nous commençons à avoir certains éléments de réponse concernant l'évaluation de la qualité d'un corpus textuel pour l'élaboration d'une ontologie (Aussenac-Gilles *et al.*, 2003; Bourigault & Lame, 2002; Le Moigno *et al.*, 2002a; Baneyx *et al.*, 2005a) :

- Il faut bien identifier les différents genres de textes disponibles et nécessaires, et sélectionner ceux qui sont pertinents par rapport aux objectifs.
- Les textes constituant le corpus doivent être collectés avec l'aide de spécialistes du domaine à modéliser et les connaissances qu'ils contiennent doivent avoir fait l'objet d'un consensus à l'intérieur de cette communauté.
- Ils doivent ensuite être triés et organiser en groupes homogènes. Cependant, si le corpus doit conserver une certaine homogénéité dans le choix du genre textuel, il doit également

conserver une part d'hétérogénéité dans la mesure où les divers textes qui le composent doivent, si possible, provenir de sources variées. Il est ainsi plus représentatif. C'est le cas de nos deux corpus : l'un rassemble des comptes rendus d'hospitalisation provenant de six hôpitaux et l'autre est constitué d'un ensemble de documents correspondant à un livre de cours sur la pneumologie.

- Les textes sont ensuite analysés en connaissance de cause, en tenant compte de leur nature, de leur origine, ...

  L'élaboration du corpus est une tâche délicate qui nécessite du temps, aussi il paraît important de ne pas remettre en cause la qualité et l'adéquation du corpus une fois la collecte des ressources achevée. Le choix des éléments du corpus se fait donc en fonction de l'application visée par l'ontologie.

- Enfin, la taille du corpus varie d'un travail à l'autre, il semble difficile de donner des indications précises quant à un nombre de mots optimum. À titre indicatif, le corpus nécessaire à l'élaboration d'une ontologie du droit par Didier Bourigault compte 1 596 583 mots, celui nécessaire à l'élaboration d'une ontologie de la réanimation chirurgicale par Sophie Le Moigno compte 350 000 mots et nous nous sommes servis de deux corpus différents, respectivement 417 000 et 823 000 mots pour construire une ontologie dans le domaine de la pneumologie (*cf.* chapitre 4). G. K. Zipf, dans les années 30, a montré que si l'on dresse une table de l'ensemble des mots différents d'un texte quelconque, classés par ordre de fréquences décroissantes, on constate que la fréquence d'un mot est inversement proportionnelle à son rang dans la liste, ou, autrement dit, que le produit de la fréquence de n'importe quel mot par son rang est constant : ce que traduit la formule f * r = C, où f est la fréquence et r le rang. Cette égalité, qui n'est vraie qu'en approximation, est indépendante des locuteurs, des types de textes et des langues. Il semble ainsi qu'il s'agisse véritablement d'un trait général des énoncés linguistiques (Zipf, 1949). Ainsi, la loi de Zipf stipule que la fréquence du second mot le plus fréquent est la moitié de celle du premier, la fréquence du troisième mot le plus fréquent, son tiers, ... Cette constatation, permet de déterminer statistiquement le nombre de « mots-clés » qu'il est intéressant d'étudier dans un corpus et représente un moyen, ou en tout cas une indication, pour évaluer la taille nécessaire du corpus. S. Le Moigno *et al.* ont ainsi appliqué la loi de Zipf à leur corpus en utilisant cette loi pour 200, 400, 600 et 800 comptes rendus d'hospitalisation répartis de façon homogène pour limiter tout risque d'effet centre. D'après l'évolution des résultats obtenus en fonction du nombre de comptes rendus d'hospitalisation analysés, 600 comptes rendus d'hospitalisation représentant 350 000 mots ont été retenus (Le Moigno *et al.*, 2002b).

2.2 Évaluation du contenu de l'ontologie

2.2.1 Critères pour la modélisation versus critères pour l'évaluation

À l'heure actuelle, il n'existe pas encore de consensus au sein de la communauté d'Ingénierie ontologique à propos des meilleures pratiques à adopter lors du processus de développement d'ontologies. Cependant, des contributions allant dans ce sens sont déjà disponibles dans la lit-

térature, en particulier les travaux de R. Mizoguchi (Mizoguchi, 1998) et de M. Uschold et M. Grüninger (Uschold & Grüninger, 1996).

Nous pensons que les critères pertinents pour évaluer une ressource terminologique et ontologique sont ceux qui doivent être respectés lors de la construction de cette même ressource. Ainsi, il est important de respecter un certain nombre de contraintes pour obtenir une modélisation de qualité, à la fois adaptable et maintenable. Nous présentons ci-dessous une série de principes, issus de la littérature, qui nous semblent particulièrement pertinents :

- Clarté et Objectivité (Gruber, 1993) : l'ontologie doit fournir la signification des termes définis en donnant des définitions objectives. Les ambiguïtés doivent être réduites, quand une définition peut être axiomatisée, elle doit l'être. Dans tous les cas, des définitions en langage naturel doivent être fournies.
- Perfection (Gruber, 1993) : une définition exprimée par des conditions nécessaires et suffisantes est préférable à une définition partielle.
- Cohérence et extensibilité (Gruber, 1993) : une ontologie doit être cohérente pour permettre des inférences conformes aux définitions. Les axiomes doivent être consistants. La cohérence des définitions en langage naturel doit être vérifiée autant que faire se peut. Ainsi, l'ajout de nouveaux concepts à l'ontologie ne devrait pas entraîner la révision des définitions existantes. Plus généralement, l'ontologie doit être construite de telle manière que l'on puisse l'étendre facilement, sans remettre en cause ce qui a déjà été fait.
- Biais d'encodage minimal (Gruber, 1993) : l'ontologie doit être conceptualisée indépendamment de tout langage d'implémentation. Le but est de permettre le partage des connaissances, contenues dans l'ontologie, entre différentes applications utilisant des langages de représentation différents.
- Engagements ontologiques minimaux (Gruber, 1993) : une ontologie doit faire un minimum d'hypothèses sur le monde : elle doit contenir un vocabulaire partagé mais ne doit pas être une base de connaissances comportant des connaissances supplémentaires sur le monde à modéliser. Autrement dit, elle doit choisir de représenter un point de vue en particulier sur le domaine qui l'intéresse.
- Principe de distinction ontologique (Borgo *et al.*, 1996) : à chaque fois que l'on peut identifier et isoler un noyau de propriétés considérées comme invariables pour une classe (critère d'identité), il faut créer une nouvelle classe de concepts.
- Minimiser la distance sémantique entre les concepts enfants de mêmes parents (Arpirez *et al.*, 1998; Bachimont, 2000) : il s'agit de la distance minimale entre les concepts enfants de mêmes parents. Les concepts similaires sont groupés et représentés comme des sous-classes d'une classe, et doivent être définis en utilisant les mêmes primitives, sachant que les concepts qui sont moins similaires sont représentés plus loin dans la hiérarchie.
- Normaliser les termes chaque fois que c'est possible (Arpirez *et al.*, 1998).

D'autres principes du même type sont proposés par d'autres auteurs (*cf.* article de A. Gómez-Pérez (2000)). Nous montrerons que la méthode ARCHONTE (Bachimont, 2000), décrite chapitre 4 section 1, et employée pour développer l'ontologie de la pneumologie, fournit des moyens d'appliquer ces principes au niveau sémantique ou au niveau formel selon les cas.

2.2.2 Notre proposition

Nous avons tenté de respecter les critères présentés ci-dessous dans la construction de l'ontologie de la pneumologie. Nous les envisageons également comme critères d'évaluation :

- le modèle doit être structuré avec des principes rigoureux, explicites et propagés dans l'ensemble du modèle ;
- le modèle doit être adapté en fonction des besoins et, tout particulièrement, en fonction des besoins des applications informatiques cibles, il doit donc être cohérent ;
- le modèle doit s'attacher à représenter de manière explicite les connaissances ;
- le modèle construit doit être maintenable et extensible, c'est pourquoi il faut garder constamment à l'esprit, tout au long de l'élaboration, les principes de structuration que l'on décide d'appliquer ;
- le modèle doit être une ressource à part entière et être, en tant que tel, séparé des applications qui l'utilisent ;
- si possible, nous pensons qu'il faut utiliser des langages standardisés, largement diffusés et diffusables, pour exprimer ce modèle.

Le respect de ces critères facilite l'ajout de nouvelles connaissances et la maintenance du modèle. Le respect systématique des principes qui guident la construction de l'ontologie et déterminent la position de chaque concept dans la hiérarchie, permet de savoir comment placer un nouveau concept. Par exemple, nous avons besoin dans notre modèle de représenter précisément les signes qui permettent aux médecins d'établir un diagnostic. Pour classer ces signes, nous devons les distinguer les uns des autres. Les *signes cliniques* sont visibles sans matériel médical, par exemple la toux, tandis que les *signes paracliniques* ne sont visibles qu'à l'aide de matériel médical, par exemple l'atélectasie (un affaissement des alvéoles du poumon) qui est visible grâce à l'imagerie médicale. Cette classification est bien évidemment discutable mais elle répond à nos besoins, plus précisément, ceux de notre application. L'ajout d'un nouveau signe dans l'arborescence se fait en déterminant s'il est ou n'est pas visible sans matériel médical.

2.3 Évaluation de la qualité taxinomique

La plupart des ontologies existantes sont organisées sous la forme d'une hiérarchie de subsomption respectant en cela les propositions d'Aristote pour l'Ontologie philosophique. Cette organisation taxinomique s'est ainsi imposée depuis des siècles comme la plus efficace pour structurer des modèles du monde, en particulier les ontologies. Nous présenterons ici un ensemble d'erreurs qui peuvent être faites en établissant des taxonomies et qui donneront des ontologies problématiques (Gruber, 1993). Ces biais de création ressortent lors de l'utilisation d'éditeurs d'ontologie, PROTÉGÉ [1] par exemple, et de raisonneurs, RACER[2] par exemple. À cette étape, on vérifie si le contenu de l'ontologie est correctement structuré, c'est-à-dire que l'on s'assure que ses définitions (en langage naturel et en langage semi-formel) respectent les spécifications

[1] `http://protege.stanford.edu/`
[2] `http://www.racer-systems.com/`

de l'ontologie ou que les modèles de l'ontologie sont conformes à la réalité (le domaine). D'un point de vue pratique, il s'agit ici de vérifier la structure de l'ontologie semi-formelle. On vérifie l'ensemble des erreurs pouvant être faites au moment de la construction de l'ontologie initiale (Gómez-Pérez, 2004). Il faudra rechercher ces erreurs en priorité lors de l'évaluation car elles ont, bien évidemment, un impact important sur l'évolution et la maintenance du modèle construit :

- Erreurs de circularité : elles se produisent quand une classe est définie comme spécialisation ou généralisation d'elle-même. Selon le nombre de relations impliquées, ces erreurs peuvent être classées comme : erreurs de circularité à distance zéro (une classe avec elle-même), à distance 1 et à distance n.
- Erreurs de partition (ou erreurs de bords) : les partitions peuvent définir des classifications de concept d'une façon disjointe ou complète. Des erreurs pourraient apparaître quand la définition de la partition entre un ensemble de classes est omise. Elles se produisent quand, par exemple, l'ontologue définit une partition d'une classe dans un ensemble de sous-classes qui ne sont pas disjointes et devraient l'être.
- Erreurs de redondance : elles se produisent quand on redéfinit des expressions qui ont déjà été explicitement définies ou qui peuvent être déduites à partir d'autres définitions. Par exemple, une « fracture du tibia » est une « fracture de la jambe » mais c'est également une « fracture du membre inférieur ». Or, si l'on veut garantir la cohérence et donc la non redondance, de la hiérarchie de l'ontologie, il ne faut pas que « fracture du tibia » hérite des deux classes.
- Erreurs sémantiques : elles se produisent habituellement quand l'ontologue classe un concept comme sous-classe d'une classe auquel il n'appartient pas réellement.
- Erreurs d'incomplétude (ou erreurs d'imperfection) : généralement, cette erreur se produit chaque fois que des concepts sont classés sans les expliquer complètement et lorsque l'ontologue n'a pas une vue d'ensemble des concepts existants dans le domaine. Une erreur de ce type serait, par exemple, de créer différentes classes pour « fièvre bactérienne », « fièvre virale », et « fièvre parasitaire » là où il serait préférable de créer une seule classe « fièvre » et la préciser via les propriétés du concept.

Nous pensons qu'un travail bien pensé en amont sur la structure différentielle de l'ontologie, tel que présenté dans la méthode ARCHONTE, évite en grande partie les erreurs listées ci-dessus.

2.4 Évaluation de l'ontologie en situation

Ne l'oublions pas, une ontologie est avant tout un outil. C'est donc en situation qu'il convient d'évaluer sa qualité et son intérêt. Selon A. Gómez-Pérez cette étape sert à juger de la valeur ajoutée (compréhension, utilisabilité...) et de la qualité de l'ontologie du point de vue de l'utilisateur (Gómez-Pérez, 2004). Il devient nécessaire de penser l'évaluation de l'ontologie en testant ses performances sur des tâches spécifiques. Cette idée a été discutée dans le chapitre précédent, section 6.

Une question reste en suspend : s'il paraît nécessaire de valider et d'évaluer l'ontologie, qui doit le faire ? La réponse qui vient en premier est l'utilisateur. Mais l'expérience tend à prouver

que les utilisateurs finaux des systèmes à base ontologique, s'ils sont capables de donner leur avis sur tel ou tel concept, sur une partie de l'arborescence, n'ont que très rarement les compétences requises pour avoir une vision d'ensemble et juger la totalité de la construction et de son schéma (Navigli & Velardi, 2004). Il est donc nécessaire de prévoir également une évaluation au niveau des concepteurs (Porzel & Malaka, 2004). R. Porzel et R. Malaka distinguent trois critères pour mesurer l'efficacité d'une ontologie en situation, selon une tâche donnée : 1) la pertinence du vocabulaire décrivant les concepts, 2) la pertinence de la hiérarchie *is-a* et 3) la pertinence des relations sémantiques. Chacun de ces critères est typé selon trois sortes d'erreurs possiblement rencontrées : a) les erreurs d'addition, b) les erreurs d'omission et c) les erreurs d'ajustage. De nombreux auteurs semblent avoir utilisé ces critères pour évaluer leurs ontologies en situation, parmi lesquels nous citerons les travaux de C. Brewster *et al.* (2004), M. Stevenson (2002) et D. Gildea et D. Jurafsky(2002).

2.5 La question de la réutilisabilité

Sur internet, les ontologies se multiplient et se partagent. Il existe un grand nombre de serveurs qui mettent à disposition et recensent RTO et outils : Ontolingua[3], Ontosaurus[4], Onthology[5]... Lorsque les concepteurs de systèmes à base ontologique recherchent des ontologies pour leurs applications, ils sont confrontés à un choix difficile. Il s'agit bien sûr de choisir une formalisation et une conceptualisation qui leur conviennent. La méthode OntoMetric, mise au point par A. Lozano-Tello *et al.*, propose de quantifier la réutilisabilité d'une ontologie et de mesurer son adéquation à un nouveau projet (Lozano-Tello & Gómez-Pérez, 2004). Cependant, l'extrême spécificité d'une ontologie de domaine, construite dans un but bien précis, contraint de manière significative son usage (Charlet, 2002). Cela dit, un certain nombre de travaux sur les top-ontologies, c'est-à-dire des ontologies ayant un haut niveau de conceptualisation, se poursuivent.

Ces ontologies se veulent génériques, comme par exemple la top-ontologie du projet MENELAS (Zweigenbaum *et al.*, 1995), voire même universelles, nous pensons en particulier à DOLCE, élaborée par l'équipe de N. Guarino (LOA, Italie) (Gangemi *et al.*, 2002).

3 Expérimentation

3.1 De l'évaluation à l'évolution

La médecine fait partie de ces domaines pour lesquels les concepts évoluent vite, en particulier, en termes de nouveaux médicaments, de nouvelles procédures d'imagerie, de nouveaux protocoles de soins. Ainsi, dans le domaine de la pneumologie qui nous intéresse principalement ici, des connaissances nouvelles sur l'allergologie ou la médecine du travail avec exposition à l'amiante doivent être prises en compte. On constate, dans ce cas, que ces connaissances ne

[3] http://www.ksl.stanford.edu/software/ontolingua/
[4] http://www.isi.edu/isd/ontosaurus.html
[5] http://www.onthology.org/

sont pas liées uniquement à des avancées de la médecine mais aussi à des prises de conscience liées à la santé publique. Les chiffres qui caractérisent cette évolution et sont communément admis laissent penser qu'en dix ans, la moitié des connaissances médicales ont été renouvelées. À l'inverse, faire une ontologie, c'est figer la modélisation à un temps t, celui d'une modélisation statique.

Des travaux envisagent de construire et de faire évoluer automatiquement des ontologies (Maedche & Staab, 2004) mais nous pensons qu'avant cela, il faut se donner une bonne méthodologie de construction qui, de toute façon, sera efficiente, même si insuffisante, pour gérer cette évolution. Ainsi, les principes différentiels utilisés pour construire l'ontologie (*cf.* chapitre 4, section 1.1) sont les premiers garants de la cohérence de celle-ci quand elle évolue.

3.2 Représentation conceptuelle d'un thésaurus médical et évolution

La représentation conceptuelle du thésaurus de la Société de pneumologie de langue française (*cf.* figures 5.5 et 5.6) répond à une double contrainte, d'une part, laisser les utilisateurs s'exprimer en langue, vecteur principal de l'information et de la connaissance et, d'autre part, permettre une représentation formelle précise de la connaissance pour qu'elle soit traitée dans le cadre d'un logiciel.

Il est important de noter que, d'un certain point de vue, ontologies et thésaurus sont proches mais que les thésaurus ayant été développés dans un contexte linguistique – des termes pour indexer ou pour exprimer la médecine – ils n'ont pas les propriétés formelles des ontologies qui permettent leur utilisation par un système informatique, ici un système de codage de comptes rendus d'hospitalisation. Inversement, le sens des termes d'un thésaurus est accessible par quelqu'un du domaine, là où les concepts d'une ontologie le sont plus difficilement. C'est principalement pour cela qu'une ontologie ne peut remplacer un thésaurus. En effet, un thésaurus fournit une liste de termes utilisables, au moins dans le contexte pour lequel il a été développé, là où l'ontologie fournit des labels de concepts – à bien différencier des *termes préférés* – dont la signification dépend de la place dans l'arborescence. Par exemple, le concept *asthme* de notre ontologie ne recouvre pas toutes les significations et contextes d'interprétation du terme *asthme* par un pneumologue.

Dans un domaine comme la médecine où les thésaurus préexistent souvent à l'ontologie, celle-ci est développée *a posteriori* et et peut être utilisée pour construire une représentation conceptuelle du thésaurus. Dans notre application, comme dans les autres soumises aux mêmes contraintes, cette représentation est obligatoirement sujette à validation récurrente, l'ambiguïté de la langue étant irréductible à la formalisation. C'est pour cela qu'il est impossible, dans ce domaine, de mettre en place des systèmes automatiques et que les fonctionnalités proposées dans MedCKARe (*cf.* chapitre 5) appellent une validation par l'expert.

Dans ce contexte, l'ontologie devra évoluer en parallèle avec le thésaurus de spécialité. Cette évolution va dépendre de l'évolution de la médecine au sens large. Cette évolution va générer de nouvelles nominalisations qui ont, à plus ou moins long terme, une place au sein des thésaurus. Ces nouveaux termes doivent, pour être utilisés par les médecins dans notre application, avoir

une représentation conceptuelle qui peut entraîner la création de nouveaux concepts. Nous avons ainsi rencontré deux situations qui ont nécessité des évolutions :

- Il y a des cas où des termes sont utilisés par les médecins pour décrire des maladies. Ils sont donc importants pour le codage et ne sont pourtant pas dans le thésaurus. C'est le cas de *pneumopathie* : des termes plus spécifiques comme *pneumomathie d'inhalation* étaient présents et il nous a donc fallu rajouter le terme qui le subsume avec son code J18.9. Pour cet exemple, le terme avait une correspondance directe avec un concept de l'ontologie mais ce n'est pas toujours le cas.
- À l'inverse, des termes sont présents dans le thésaurus de spécialité mais ne sont plus assez spécifiques en raison d'avancées médicales ou, dans l'exemple qui suit, en raison de nouvelles préoccupations de santé publique. Ainsi, le terme *exposition professionnelle SAI*[6] avec le code Z57.9 est présent dans le thésaurus de spécialité mais pas le terme *exposition professionnelle à l'amiante*. Cela a nécessité de compléter le thésaurus ainsi que l'ontologie au sein de laquelle l'amiante, en tant que produit toxique, n'était pas présente.

Cette question de l'évolution du thésaurus doit être prise en compte en fonction de la question de la validation. Ce point est détaillé au chapitre 7, section 4.1.1.

3.3 Évolution de l'ontologie due à l'usage

Dans le cadre de notre application, nous avons évalué les résultats que nous obtenons concernant, d'une part, le codage PMSI et, d'autre part, le codage médical. Nous avons vu que ces résultats se calculent en terme de rappel et de précision. Nous distinguons deux phases d'évaluation. Entre ces deux phases, nous avons analysé les erreurs et imprécisions de notre outil et avons fait évoluer l'ontologie pour améliorer sensiblement nos résultats (*cf.* figures 5.20 et 5.21).

Concernant le codage PMSI, le rappel passe de 25% à 80%. L'analyse de ces résultats nous a aidé à enrichir nos ressources (l'ontologie et le thésaurus). Nous avons recherché les termes associés aux codes PMSI et nous avons étudié les CRH afin de savoir à quel endroit ces termes apparaissent et pourquoi notre système ne les a pas détectés. Entre les deux phases d'évaluation, nous avons essayé de corriger certains problèmes, dus le plus souvent à l'absence, au niveau de l'ontologie, de termes issus du thésaurus, comme par exemple *nausée* sous *signe fonctionnel digestif*. Nous avons également enrichi l'ontologie avec des synonymes, comme par exemple *greffe* synonyme de *transplantation* ou *ABPA* synonyme de *aspergillose broncho-pulmonaire allergique*. La précision reste stable entre les deux évaluations. Concernant le codage médical, le rappel est de 43% lors de la première phase d'évaluation. Afin d'améliorer cette mesure, nous avons enrichi l'ontologie en identifiant les termes médicaux importants figurant dans le CRH et n'ayant pas été reconnus par notre système, comme par exemple *hémiplégie* sous *paralysie*, *adénocarcinome* sous *carcinome* et *vomissement* sous *signeFonctionnelDigestif*. Lors de la deuxième phase d'évaluation, nous constatons que le rappel passe à 88%.

En ce qui concerne la précision, elle est, lors de la première phase d'évaluation, de 70%. Après la correction d'un certain nombre d'erreurs ayant engendré du bruit – comme par exemple une forme de négation non prise en compte –, la précision passe à 85%.

[6] Sans autre indication.

Pour l'instant l'évolution de l'ontologie se fait essentiellement par l'ajout manuel de nouveaux concepts. Mais notre méthode, et particulièrement la structuration de la hiérarchie à l'aide des axes différentiels, ainsi que les principes rigoureux que nous nous sommes donnés, facilitent grandement l'évolution et la maintenance de notre modèle et permettent d'envisager des interfaces pour supporter cette évolution. Ces développements font l'objet d'un projet récemment lancé.

4 Perspectives et conclusion

Dans ce chapitre, nous avons présenté notre expérience de construction d'une ontologie de la pneumologie au regard de trois problématiques étroitement liées : 1) l'évaluation de la pertinence de notre modèle ontologique, 2) l'évolution des connaissances au sein de notre ontologie et du thésaurus de spécialité associé, notamment par rapport aux corpus textuels de référence, et 3) la maintenance de notre ontologie et son application.

La question de la maintenance n'a, pour l'instant, été expérimentée que par rapport aux premiers résultats fournis par notre outil d'aide aux codages et aux améliorations qui s'en sont suivies. Nous n'avons pas encore été réellement confrontés à des besoins d'évolution et de maintenance qui seraient dus à de nouvelles découvertes dans le domaine de la pneumologie. Cela ne manquera pas de nous arriver en travaillant à la modélisation des connaissances dans le domaine médical.

Nous voudrions ici remettre en avant les axes de réflexions qu'il nous semble important de développer pour faire entrer l'ingénierie ontologique dans les domaines maîtrisés au même titre qu'un logiciel. Certains de ces axes (1, 4) ont été spécifiquement développés dans les propositions de ASSTICCOT (Aussenac-Gilles & Condamines, 2003) :

Évaluation des corpus textuels
: Les corpus textuels qu'il faut collecter pour construire les ontologies posent encore de nombreux problèmes. En particulier, la faculté de préciser leur genre est importante pour nos méthodologies. Certains critères de classement semblent être intéressants comme, par exemple, date et lieu de production, niveau de compétence du rédacteur, objectif initial de la production, public visé – *e.g.* patient ou médecin –, nature de l'activité documentarisée par ces textes...

Principes de construction de l'ontologie
: Les travaux de Bruno Bachimont sur les principes différentiels permettent de fournir des critères de différenciations constructifs et constructivistes pour l'ontologie. La méthodologie longuement mise au point dans ce contexte doit encore être complétée pour s'intégrer totalement dans l'étape suivante de construction de l'ontologie formelle qui doit être opérationnelle.

Évaluation du contenu de l'ontologie
: Nous synthétisons ici les propositions de la section 2.2 en insistant sur la rigueur des principes de modélisation à appliquer tout au long du processus, de la construction de l'ontologie à sa maintenance. La problématique qui reste encore à approfondir est la possibilité

d'expliciter des critères qui permettraient d'affirmer la bonne adéquation de l'ontologie à l'application. Ces critères peuvent être à chercher du côté de la formalisation ou avec des visées plus empiriques.

Maintenance de l'ontologie avec le corpus

Les travaux à ce sujet sont le pendant de la construction d'ontologies à partir de corpus qui commence à faire ses preuves. Ils doivent être développés et approfondis pour que le niveau d'efficience de la maintenance d'ontologies rejoigne celui de la construction.

Maintenance du thésaurus avec l'ontologie

La question des thésaurus, si elle n'est pas spécifique à la médecine, est prégnante dans ce domaine. On peut, et l'on doit très probablement, imaginer des applications combinant les deux. C'est ce que nous faisons avec l'application MedCKARe. Ce type d'application oblige à penser conjointement l'évolution du thésaurus et de l'ontologie.

CHAPITRE 7

Perspectives et conclusion

FIG. 7.1 – Illustration de Denis Pessin

Ce dernier chapitre, consacré aux perspectives et conclusion, revient sur les différents axes de recherche qui nous tiennent à cœur. Ainsi, nous reprenons la question de la réutilisabilité de la top-ontologie du projet MENELAS *en section 1. Nous faisons le point, en section 2, sur les différentes fonctionnalités implémentées dans le cadre du serveur de terminologie* PERTOM*ed. Les sections 3 et 4 présentent le projet RNTL 2006 Da*FOE*4App puis notre propre projet postdoctoral intitulé Med*OC *qui est une des applications de la plateforme Da*FOE*. Enfin, la section 5, conclut ce mémoire et dresse le bilan de nos contributions méthodologiques et techniques.*

1 Réutilisabilité de la top-ontologie de MENELAS

L'utilisation de la « top-ontologie » de MENELAS doit être validée. Cela se fera au cours de l'année 2007, en collaboration avec le réseau d'excellence « Semantic Mining »[1]. En effet, une des tâches de ce réseau est précisément la définition d'une (ou de la) top-ontologie de la médecine. Les partenaires du réseau font partie des équipes impliquées dans les principaux projets de top-ontologies médicales (DOLCE/ROME (Gangemi *et al.*, 2002), BOF (Basic Formal Ontology, (Smith, 2004; Grenon, 2003)), GALEN (Rector, 1998; Rector *et al.*, 1992)...).

Dans le cadre de ces collaborations, les ontologies sont découpées en trois niveaux, l'ontologie du domaine telle que nous l'avons définie dans ce mémoire, la top-ontologie donnant les grandes classifications nécessaires au découpage du monde et comprenant moins d'une trentaine de concepts et la « middle-ontologie », de niveau intermédiaire entre la top-ontologie et l'ontologie du domaine. On peut noter que même la top-ontologie des grandes classifications ne peut pas être indépendante du domaine au sens large – ici la médecine – et que les travaux de N. Guarino *et al.* consistent, en ce domaine, à adapter DOLCE à la médecine.

Nous n'avons pas d'*a priori* sur le choix définitif de la top-ontologie et de la middle-ontologie pour la médecine. Il est d'ailleurs rarement évident de distinguer avec précision leurs contours. Nous ne sommes pas fixés par principe sur la structure retenue pour MENELAS. Nous espérons, et ce sera un test, que quel que soit le choix fait, le raccord entre ce travail et OntoPneumo sera aisé. Notre première expérience, entre MENELAS et OntoPneumo développées à des périodes différentes et avec des méthodologies différentes – fondées sur les études de corpus pour OntoPneumo et fondées sur des réflexions philosophiques et pratiques pour MENELAS– est, en l'espèce, encourageante.

Enfin, de l'avis même des partenaires avec lesquels nous travaillons et bien que nous réduisions notre champ d'investigation à la médecine, il n'est pas sûr que nous convergions à court terme même si c'est le but affiché.

2 Serveurs de terminologies et services associés

Dans le cadre du projet PERTOMed et de son serveur de terminologie[2] (*cf.* figure 7.2), nous souhaitions offrir un certain nombre de services à l'utilisateur. L'ensemble de ces services s'appuient sur une ontologie du domaine. La plupart ont pu être implémentés, soit par nos soins, soit par des collaborateurs, et mis à disposition sur internet à l'adresse suivante : `http://baneyx.net/SPIP/`.

– navigation dans l'ontologie,
– édition/modification de l'ontologie (selon autorisation),

[1] `http://www.semanticmining.org/`

[2] Un serveur de terminologie est une plateforme logicielle qui fournit, dans un domaine d'expertise donné, un certain nombre de services, a minima conceptuels, ensuite linguistiques. Ces services sont à destination d'un utilisateur ou d'un programme (requêtes normalisées). La délégation de ces services au serveur permet une prise en charge centralisée des problèmes terminologiques pour une communauté de travail ou un ensemble d'applications.

– spécification et construction d'une expression conceptuelle complexe d'éléments de l'ontologie en fonction des signatures des relations (les contraintes de combinaison),
– vérifications portant sur une expression conceptuelle (validité, canonicité, mise sous forme canonique),
– distance sémantique entre concepts,
– indexation ontologique.

Ces services nécessitent la mise en œuvre d'outils de classification et d'inférence sur les ontologies. Les services linguistiques sont rendus par la mise en relation des concepts (ontologies) avec des terminologies ou des thésaurus du domaine et permettent, dans un premier temps, la maintenance, la comparaison ou la fusion de thésaurus. Par manque de temps, l'ensemble de

FIG. 7.2 – Serveur PERTOMed

ces services n'ont pas pu être implémenté, en particulier le service d'indexation ontologique des comptes rendus d'hospitalisation de pneumologie. Les réflexions et les développements en cours vont se poursuivre dans le cadre du projet RNTL DaFOE 4app et de notre projet postdoctoral MedOC qui commencent début 2007. Nous les décrivons ci-dessous.

3 Projet DaFOE4App

Le projet DaFOE4App au sein duquel nous allons poursuivre nos recherches est un projet financé par le Réseau National des Technologies Logicielles[3] (RNTL 2006). J. Charlet, rattaché au laboratoire Santé publique et Informatique médicale de l'INSERM UMRS 729, le pilote.

3.1 Contexte et enjeux

Le but du projet DaFOE4App est de fournir des méthodologies et des outils permettant de construire des ontologies pour des systèmes à base de connaissances utilisés dans des environnements professionnels où la connaissance est complexe et nécessite, par conséquent, un effort de modélisation important. La plateforme DaFOE (*Differential and Formal Ontologies Editor*) a pour objectif de créer et/ou de faire évoluer ces ontologies. Elle regroupe un ensemble d'outils, dont un éditeur d'ontologies qui prendra en charge la question de la sémantique de ces ontologies, à travers des questions épistémologiques liées aux concepts formels de haut niveau - i.e. la top-ontologie. La question de la composante métier sera prise en charge au travers de travaux sur les corpus textuels. Le but est de rendre disponibles, à travers la plateforme, les méthodologies qui vont permettre de construire l'ontologie. On obtient ainsi une ontologie référentielle qui pourra être traitée dans un éditeur respectant les standards des langages d'ontologies référentielles du W3C (OWL) avant d'être utilisée dans un système à base de connaissances.

Nos propres recherches méthodologiques pourront servir à établir le cahier des charges de l'éditeur et à participer à sa réalisation (phases de test, d'évaluation et d'évolution). L'évaluation d'une telle plateforme et de ses fonctionnalités se fait au regard de la qualité des ontologies développées. Nous en profiterons pour rééditer OntoPneumo dans ce nouvel environnement et y apporter quelques améliorations. Cela nous permettra également de mesurer la valeur ajoutée de l'éditeur de DaFOE par rapport à son « ancêtre » DOE.

La qualité des ontologies développées dans ce nouvel environnement est quantifiable à travers un certain nombre de critères mais, de façon privilégiée, elle sera prouvée via le fonctionnement de l'application dans laquelle elle va servir. Le projet prévoit la réalisation de trois applications dans trois domaines différents : 1) l'aide au codage médical, 2) l'indexation patrimoniale et 3) l'indexation d'images satellitaires. Dans les domaines traités, la qualité de l'ontologie se traduit par la capacité de justifier et motiver les choix de modélisation ontologique, tant sur le contenu des concepts (aspect métier pris en charge dans l'ontologie différentielle) que sur leur forme logique (aspect épistémologique pris en charge dans l'ontologie formelle). En effet, les systèmes visés ont pour but d'assister le spécialiste : l'enjeu est donc l'utilisabilité et pas seulement l'effectivité.

3.2 Intérêts scientifiques

Dans ce contexte, la plateforme DaFOE doit fournir un certain nombre de services. Ces services sont explicités à travers les expériences d'un certain nombre de partenaires impliqués

[3] http://www.rntl.org/

dans la construction d'ontologies ou dans l'élaboration d'éditeurs d'ontologies. Ce sont ces expériences ainsi que les difficultés rencontrées au niveau des fonctionnalités offertes par les différents logiciels dédiés à la création d'ontologies qui motivent ce projet. Les fonctionnalités étudiées ou utilisées dans DaFOE font appel à des méthodes ou algorithmes proposés dans d'autres approches, plus « automatiques ». C'est justement les conditions d'utilisation des méthodes/algorithmes étudiés qui impliquent de les intégrer au sein de DaFOE en tenant compte de l'interaction forte avec l'utilisateur. C'est là que se situe la première originalité du projet.

3.2.1 Recherche et extraction des constituants d'une ontologie

Dans la suite des travaux et réflexions du groupe TIA , il est apparu que les corpus textuels étaient d'intéressantes sources de connaissances pour construire des ontologies, donc pour repérer les concepts et les relations. C'est ce qui justifie, entre autres, depuis quelques années, le développement d'outils d'analyse de textes robustes et à large couverture. Leurs résultats servent d'entrées à des outils d'analyse distributionnelle qui proposent de structurer les concepts. Les résultats de ces analyseurs dans le cadre de construction de ressources terminologiques et ontologiques ont été exploités dans des domaines expérimentaux et a amené les premiers développements d'interfaces de modélisation pour construire ces ressources terminologiques et ontologiques. Ce type d'interface doit être entièrement re-réfléchi et ré-implémenté pour être intégré dans la plateforme. Par ailleurs, les méthodes de fouille de données à base de techniques de classification peuvent s'avérer utiles pour extraire les concepts et relations d'une ontologie et compléter les techniques linguistiques décrites ci-dessus.

Dans le cadre de nos travaux, nous avons été confronté aux divers problèmes posés par la construction de corpus textuels (pertinence, niveau de langue, taille ...), la structuration des concepts en hiérarchie (limite du domaine à représenter, cohérence, régularité ...). Nous avons utilisé, pour OntoPneumo et pour MedCKARe , des outils d'analyse de textes et avons des idées d'améliorations à proposer : ergonomie, fonctionnalités ...

Au cœur de la recherche de relations se trouve l'exploration des textes par projection de patrons lexico-syntaxiques. Cette exploration est d'autant plus efficace si le corpus a été analysé syntaxiquement et étiqueté. L'analyse distributionnelle des expressions extraites et l'identification de relations grammaticales entre les termes au sein de ces expressions augmentent fortement la puissance des patrons. De plus, une étude statistique des termes et formes syntaxiques identifiées devrait permettre de se focaliser plus rapidement sur des contextes pertinents. Or, à ce jour, les différents types d'analyse sont menés par des outils indépendants et prenant peu en compte des traitements statistiques. Nous pensons augmenter l'efficacité du processus en couplant ces approches au sein d'une même plateforme de modélisation. Une partie du processus fera appel à des techniques d'apprentissage, pour suggérer de nouveaux patrons à partir d'un ensemble de patrons définis ou la révision de certains patrons par la confrontation avec l'ensemble des exemples couverts. Comme suggéré dans le chapitre 4 et comme le montre le repérage des concepts définis au chapitre 5, les relations non taxinomiques peuvent également être extraites en recherchant des motifs fréquents à partir de patrons.

Nous espérons que cette partie du projet nous permettra de développer nos connaissances sur les processus d'extraction de relations. Nous pourrons alors envisager d'un œil critique les

relations que nous avons reprises pour notre propre usage du projet MENELAS. Nous souhaitons également que les réflexions issues de ces recherches nous permettent de développer des patrons plus performants pour détecter le phénomène de négation dans MedCKARe et gérer la non détection des antécédents familiaux aussi bien pour le codage médical que pour le codage PMSI.

3.2.2 Évolution des ontologies

L'évolution des ontologies peut être appréhendée de différentes manières.

- Elle peut s'effectuer par rapport à des changements au niveau de l'application, et donc l'intégration nécessaire de nouvelles sources de données et leur implication sur l'ontologie. Comme nous l'avons vu dans le chapitre 6 de ce mémoire, la difficulté est alors de spécifier la façon de relier l'évolution des ontologies à l'évolution des corpus, des thésaurus ou, plus précisément dans les cas sur lesquels nous voulons mettre l'accent, des corpus de textes traçant l'activité des professionnels du domaine.
- Elle peut être nécessaire parce qu'on dispose de deux versions de l'ontologie. On doit alors rechercher les différences par des techniques assez semblables à celles utilisées pour établir des correspondances sémantiques entre ontologies. Cette question de l'évolution des ontologies est centrale pour la pérennité des applications mais n'est pas traitée par les éditeurs actuels.

Nous avons rencontré les deux cas de figure dans notre thèse. Jusqu'à présent, nous avons centré nos réflexions sur l'évolution des ontologies par rapport aux thésaurus et aux corpus dans le domaine de la médecine. Nous aimerions, à l'occasion du projet DaFOE4App, reprendre notre collaboration avec D. Pisanelli du LOA à Rome portant sur la fusion d'ontologies et les évolutions que cela suppose. Nous avons déjà amorcé la fusion d'OntoPneumo avec ROME sous PROTÉGÉ à l'aide de la suite PROMPT. Les premiers résultats au niveau de la middle ontologie sont intéressants et nous ont permis de faire évoluer, par exemple, la représentation des médicaments. Les résultats au niveau de la top-ontologie sont plus difficilement exploitables car ROME et MENELAS reposent sur des présupposés philosophiques très différents.

Le projet fait donc la synthèse entre plusieurs types de recherches encore séparés : l'extraction terminologique de corpus, l'extraction de relations, la modélisation linguistique des concepts, l'exploitation des ontologies, leur évolution et leur maintenance. Ces sujets de réflexion appartiennent déjà à notre champ d'investigation.

4 Projet MedOC

Le projet MedOC est une des applications concrètes du projet DaFOE4App Il s'agit également de notre projet postdoctoral, c'est pourquoi la plupart des objectifs de développement sont dans la continuité de ceux développés au cours de notre doctorat dans le cadre du projet PERTOMed.

4.1 Objectifs

4.1.1 Environnement d'aide au codage

Ce projet s'attache donc à développer un environnement d'aide au codage, combinant étroitement une ontologie de la pneumologie et le thésaurus de cette même spécialité. Ceci peut se faire dans le cadre d'une représentation de chacun des termes du thésaurus dans les concepts de l'ontologie, élaborant ce que nous appelons un codage médical avec MedCKARe. Cette recherche est déjà bien avancée à la fin de ce doctorat. Cependant, il reste des améliorations à apporter pour compléter l'ontologie et la modélisation du thésaurus de spécialité. Une réflexion de fond devra être menée dans MedOC sur la validation des changements apportés dans le thésaurus qu'elle soit le fait d'un expert ou bien de la Société de pneumologie de langue française, responsable du thésaurus. Le statut des entrées ajoutées au thésaurus doit être précisé et conservé de manière à garder la trace des entrées officielles par rapport à celles que nous ajoutons pour les besoins de notre outil MedCKARe. Nous prévoyons également de comparer nos résultats avec ceux de l'outil de codage SNOCODE, présenté chapitre 5 section 2.

4.1.2 Aide à l'indexation des dossiers médicaux et à la recherche d'information

Le développement du dossier médical informatisé, qu'il soit effectif à certains endroits ou embryonnaire à d'autres, pose le problème de son indexation, c'est-à-dire de la possibilité, d'enregistrer avec ce dossier, un certain nombre d'informations qui permettront de le retrouver, durant l'hospitalisation ou dans les archives, en fonction des besoins exprimés par les praticiens. Jusqu'à présent les médecins que nous avons rencontrés prennent en notes manuscrites les noms des patients présentant des pathologies intéressantes pour la recherche et les études épidémiologiques. Il est bien entendu impossible pour eux de faire des recoupements intéressants sur un grand nombre de cas. Nous souhaitons apporter dans MedOC une solution à ce problème. L'indexation que nous proposons se sert des concepts et des relations comme descripteurs et doit permettre :

- le codage médical des dossiers médicaux, en vue de leur archivage et surtout de leur récupération en fonction de critères médicaux (par exemple, une requête telle que la recherche de tous les patients pour lesquels on a diagnostiqué telle pathologie par tel examen),
- le codage PMSI.

Cette application sera développée dans le contexte de la pneumologie, spécialité dans laquelle nous disposons déjà d'une ontologie et d'une première version de l'application.

4.1.3 Collaborations

MedOC est un programme de recherche pluridisciplinaire qui repose sur la mise en place de collaborations pour réaliser l'application et la tester dans deux services de pneumologie de l'Assistance Publique - Hôpitaux de Paris avec lesquels nous sommes en contact : l'unité de Pneumologie, Service de Médecine Interne, CHU Bicêtre (Dr F.-X. Blanc) et le Service de Pneumologie, Hôpital Saint-Antoine (Pr C. Chouaid).

Nous souhaitons également collaborer avec la société Mondeca et le LISI pour expérimenter les bases de données à base ontologique, ONTODB, (Pierra *et al.*, 2005; Jean *et al.*, 2006). Il s'agit de représenter l'ontologie en base de données selon un modèle de données spécifiques dans lequel le LISI a développé des compétences. On sait par expérience que des grandes ontologies sont fréquemment inexploitables au sein d'applications réelles : que l'ontologie soit spécifiée en OWL (langage par ailleurs indispensable pour exprimer l'ontologies dans des termes normalisés) ou en RDF, dès une certaine taille, environ 2000 concepts typiquement. Il est bien souvent difficile de faire des requêtes sur l'ontologie – subsomption ou raisonnement – qui soient satisfaites en des temps acceptables. Nous espérons que la mise en place du modèle OntoDB se révèlera fructueux.

4.2 Intérêt scientifique

4.2.1 Originalité du projet de recherche

Nous proposons dans le projet MedOC que le noyau de l'environnement d'aide au codage soit une ontologie du domaine de la pneumologie. C'est là que réside une partie de l'originalité du projet par rapport à la réalisation d'outils de codage industriels. Une ontologie décrit de manière générique et formelle les connaissances propres à un domaine donné, ici la pneumologie, et offre de celui-ci une compréhension consensuelle pour un groupe de professionnels. Cette même ontologie est formalisée dans un langage de représentation des connaissances qui permet son exploitation par une application informatique. Cette ontologie a été développée dans le cadre du projet PERTOMed et servira de matériel pour la réalisation de notre environnement. Une seconde source d'originalité réside dans les services qu'une telle application sera en mesure de fournir par rapport aux outils de codage PMSI existants : recherche sémantique, indexation multicritères des données du patient, optimisation du codage PMSI ... Enfin, dans la continuité du travail de thèse, je choisis de poursuivre mon activité dans un cadre pluridisciplinaire.

4.2.2 Utilité médicale

Comme nous l'avons déjà dit, le codage des diagnostics et actes médicaux est une obligation légale des structures de soins (hôpitaux, cliniques et bientôt tout professionnel de santé). Ce codage est la description normalisée au travers de thésaurus des actes pratiqués sur un patient et des diagnostics posés. Ces thésaurus sont la CIM-10 pour les diagnostics et la CCAM pour les actes. Cette activité est faite de façon manuelle ou à l'aide d'outils commerciaux semi-automatiques par le praticien et est source de nombreuses erreurs. Les travaux sur l'aide au codage se concentrent en grande partie sur la représentation du patient (actes et diagnostics) à travers les 2 thésaurus réglementaires. Une telle démarche est forcément limitée, en particulier parce que ces thésaurus sont loin des préoccupations médicales du médecin : la CIM-10 et, encore plus, la CCAM ont été développées pour des besoins de représentation médico-économique. Les besoins du médecin sont de natures différentes et concernent plus la recherche, l'analyse, l'échange et l'exploitation clinique des informations médicales. Ces activités doivent impérativement s'envisager à partir d'une représentation (unique et consensuelle) du patient en termes d'actes et de diagnostics.

5 Conclusion

Concernant notre participation au projet PERTOMed , nous avons travaillé activement à la réalisation de ses trois principaux objectifs :

1. Production de ressources termino-ontologiques : nous avons construit deux corpus, le corpus [CRH] et le corpus [LIVRE], qui ont pu servir de base de travail à d'autres membres de l'équipe et une ontologie dans le domaine de la pneumologie comptant 2 260 concepts primitifs.
2. Partage et développement d'expertise : Nous avons collaboré avec V. Malaisé – alors doctorante à l'Institut National de l'Audiovisuel – pour mettre en application des patrons lexico-syntaxiques pour la recherche d'énoncés définitoires dans nos corpus. L'objectif de ce travail était d'aider l'ingénieur des connaissances à préciser, en langue naturelle, les principes différentiels qui structurent la hiérarchie de l'ontologie. Nous avons également travaillé avec N. Grabar pour traiter les lexiques nécessaires au développement du dictionnaire de MedCKARe et nous avons collaboré avec D. Bourigault pour utiliser SYNTEX-UPERY et extraire les candidats termes intéressants pour notre modélisation.
3. Développement de méthodes et d'outils pour l'appariement de terminologie : Notre participation à la réalisation de cet objectif se situe au niveau technique. Nous avons assuré la traduction des lexiques dans un format permettant de les visualiser en ligne et leur mise à disposition sur la plateforme PERTOMed [4] .

La construction d'ontologies à partir de textes constitue un enjeu important aussi bien pour la communauté des chercheurs en Traitement automatique des langues que pour celle de l'Ingénierie des connaissances. D'après notre expérience, les systèmes de traitement de l'information qui doivent fonctionner dans des domaines de connaissances spécialisés comme la médecine ne peuvent être efficaces que s'ils s'appuient sur des ressources terminologiques et ontologiques, construites pour le domaine concerné et en vue d'une application particulière. L'enjeu, dès lors, est d'élaborer des méthodes d'acquisition des connaissances à partir de textes qui spécifient (1) comment utiliser les outils de Traitement automatique des langues, nécessaires à l'analyse de corpus (SYNTEX-UPERY), et (2) les environnements de modélisation des connaissances, nécessaires à la construction d'ontologies (DOE, PROTÉGÉ). À cet effet, nous avons repris la méthode ARCHONTE et l'avons complété et précisé en ce sens. Ce travail de modélisation des connaissances à partir de textes nous a également permis de mesurer la nécessité d'utiliser conjointement des outils d'analyse du langage et de modélisation. Il semble intéressant d'intégrer ces différents outils pour faciliter le passage des candidats termes à la représentation des concepts tout en assurant de pouvoir revenir aux textes. C'est ce qui est fait notamment dans TERMINAE .

Si les ontologies construites à partir de textes peuvent représenter et saisir les objectifs d'un domaine de connaissances, encore faut-il savoir comment délimiter ce domaine. Il doit être délimité aussi précisément que possible, et découpé, si besoin est, en plusieurs aspects : les connaissances du domaine, les connaissances de raisonnement et les connaissances de haut niveau qui,

[4] Disponible pour l'instant à l'url suivante, `http://baneyx.net/SPIP/`, et dans quelques mois à l'url : `http://pertomed.spim.jussieu.fr`

par leur degré d'abstraction supérieur, peuvent être communes à plusieurs domaines. Nous avons vu dans le chapitre 4 que cette question est loin d'être triviale. De mauvais choix peuvent compromettre l'intérêt du modèle et son utilité dans une application informatique. Il n'existe pas, à notre connaissance, de critères ou de méthodes qui répondent de manière fiable à ce problème. Il serait intéressant de développer cette recherche en Ingénierie des connaissances et particulièrement dans le domaine des connaissances médicales.

La question des genres textuels n'est pas nouvelle en Ingénierie des connaissances. Cependant, l'approche comparative que nous avons suivie en analysant les hiérarchies terminologiques issues de l'analyse ditributionnelle sur le corpus [CRH] et du repérage de relations par patrons lexico-syntaxiques sur le corpus [LIVRE] est plus rarement exploitée. Bien que les traitements aient porté sur des corpus différents, il est intéressant d'observer la relative compatibilité des deux ensembles terminologiques extraits. Au vue des résultats que nous avons obtenus, il serait intéressant de revenir sur la question des genres textuels.

La divergence des structures terminologiques est également un point prometteur. Elle tend à prouver l'existence d'organisations conceptuelles différentes au sein d'un même domaine de connaissances. Ce point va à l'encontre d'un certain nombre de travaux sur la modélisation d'ontologies universelles. En effet, une ontologie est une modélisation conceptuelle, non contextuelle et non ambiguë. Elle n'a donc qu'un seul contexte d'interprétation possible. Nous sommes convaincus qu'il existe de nombreuses modélisations possibles pour un domaine donné, en fonction de la tâche à réaliser, c'est-à-dire en fonction du contexte. Ainsi, le travail de structuration ontologique vise à expliciter les choix faits parmi l'ensemble des modélisations potentielles. Que faut-il alors penser des travaux qui cherchent à automatiser entièrement la construction de l'ontologie ? Il peut être intéressant d'automatiser une partie de ce processus, notamment pour gagner du temps, mais l'intervention humaine, de l'ingénieur des connaissances comme de l'expert du domaine, nous semble irremplaçable.

Le contenu, la forme, la couverture et le degré de formalisation d'une ontologie sont choisis en fonction du rôle qu'elle doit jouer dans l'application cible. Une fois construite et acceptée par une communauté particulière, cette ontologie traduit un consensus explicite et un certain niveau de partage, deux aspects essentiels pour permettre son exploitation par différentes applications ou agents logiciels. Ce point de vue soulève un paradoxe important : d'une part, dans un souci de réutilisabilité, l'ontologie gagne à cultiver une certaine indépendance vis-à-vis des différentes applications dans lesquelles elle peut être utilisée et, d'autre part, sa construction elle-même doit être guidée par l'usage dans l'application cible. Se pose alors le problème de l'utilisation opérationnelle des ontologies, c'est-à-dire de leur mise en œuvre pratique. L'utilisation d'Onto-Pneumo au cœur de MedCKARe a nécessité un travail d'adaptation conséquent mais donne de bons résultats. Les chapitres 4 et 5 de ce mémoire offrent des pistes de réponses en montrant comment, dans un cas spécifique, passer de l'expression linguistique des connaissances à une représentation formelle, calculable, propre à l'exploitation informatique. Nous tentons d'extraire les résultats de cette expérience de leur cadre spécifique et de réfléchir à un ensemble de principes généralisables pour guider l'ingénieur des connaissances dans sa démarche. Nous proposons, à cet effet, un guide méthodologique en annexe A de ce mémoire.

Le chapitre 2 souligne les limites du codage médico-économique et pose le problème de la standardisation du langage médical. Nous avons essayé, dans l'ensemble des étapes de construc-

tion de l'ontologie, dans la modélisation du thésaurus de spécialité, et dans leurs utilisations au sein de MedCKARe, de garder à l'esprit : (1) les besoins des médecins en terme de codage de leur spécialité ; (2) les connaissances médicales nécessaires pour « parler » du diagnostic et de l'acte, et plus généralement pour rendre compte de l'activité pneumologique ; (3) les connaissances pertinentes pour le codage PMSI. L'expression de ces connaissances doit être transformée en un codage qui, lui, est toujours réducteur en termes de représentation du sens. L'idée est, avec MedCKARe, de permettre aux médecins de se réapproprier le processus de codage, en aucun cas de remplacer leurs expertises. C'est pourquoi, MedCKARe est semi-automatique et intègre un système de représentation des connaissances avec lequel le médecin peut interagir pour construire la représentation du patient qu'il désire, en tenant compte de ses propres capacités de choix et d'interprétation. Les débats d'actualité concernant la gestion de la Santé et l'intérêt croissant du grand public pour les questions attenantes nous font penser que les recherches en informatique médical et des outils tels que le notre sont d'une utilité réelle. Concernant la question de la réutilisabilité, nous pensons que l'outil MedCKARe, en lui-même, est facilement réutilisable car son fonctionnement est générique. Par contre, il suppose l'utilisation de textes en entrée et la production de ressources textuelles (ontologie, thésaurus, dictionnaire, ...) adaptées à cette nouvelle utilisation. Il ne faut pas négliger la charge de travail que nécessite le développement de ces ressources.

Pour conclure ce chapitre, nous revenons sur nos principales contributions, d'ordre méthodologique, technique et pratique :

- Nous avons mis au point une méthodologie d'Ingénierie ontologique unifiée (tenant compte des principes et des méthodes de l'Ingénierie ontologique, de l'Ingénierie des connaissances, du Traitement automatique des langues, de la logique et de la sémantique différentielle) pour la construction d'ontologies, à partir de textes. À cet effet, nous avons complété et précisé la méthode ARCHONTE mise au point par B. Bachimont en montrant comment s'enchaînent les différentes étapes : 1) analyse terminologique fondée sur l'analyse des textes, 2) identification, sélection et extraction des termes pertinents, 3) normalisation, 4) formalisation et 5) opérationalisation. L'enchaînement des processus d'extraction, de sélection et de choix des candidates termes du domaine ainsi que l'aide fournie par les patrons lexico-syntaxiques pour renseigner les principes différentiels la rende relativement facile d'emploi (ou moins difficile qu'une autre) pour un ingénieur des connaissances. Nous avons généralisé la méthodologie, autant que faire ce peut, dans un guide pratique qui se trouve en annexe A de ce mémoire.
  Nous avons expérimenté, en collaboration avec V. Malaisé, la complémentarité de deux modes d'analyse de la langue en usage dans les textes, l'analyse distributionnelle et l'approche par patrons lexico-syntaxiques et nous avons montré que l'utilisation conjointe de ces méthodes facilite la structration hiérarchique des concepts de l'ontologie.
  Nous avons réutilisé une ontologie de haut niveau de la médecine et expliqué comment elle guide la réorganisation d'une première structuration des concepts.
- Pour articuler l'utilisation de plusieurs outils et de plusieurs approches, nous avons mis au point un certain nombre de programmes informatiques permettant d'aider le passage des uns aux autres, notamment en ce qui concerne la conversion des formats. À cette occasion, nous avons pu cerner quels étaient leurs atouts et leurs limites. Cela nous a fait réfléchir

aux moyens d'enchaîner ces outils, à leur complémentarité et aux meilleurs moments pour les utiliser.
Nous avons suivi le cycle de vie complet d'une ontologie, de sa création à son évaluation dans une application.

– L'ontologie OntoPneumo est un résultat en soi. Elle peut être réutilisée pour des tâches approchantes et nécessitera alors d'être complétée et en partie remaniée.
Nous avons développé MedCKARe, un outil de codage médical semi-automatique, qui propose deux types de codages : (1) un codage médical qui représente graphiquement, pour l'instant, les informations relatives aux pathologies du patient et, à terme, servira de descripteur pour indexer intelligement les CRH ; (2) un codage médico-économique qui propose au médecin pneumologue une liste de codes PMSI générée en fonction des pathologies pertinentes identifiées. Cet outil ne s'intéresse pour le moment qu'aux pathologies – *i.e.* aux diagnostics – mais nous comptons, dans un avenir proche, gérer également les actes. MedCKARe utilise OntoPneumo et fait ainsi la démonstration de l'utilité d'une telle modélisation dans un système informatique.
La réalisation de cette application nous a également permis de montrer et de mesurer l'impact de l'application cible sur le contenu du modèle et sur la manière de le construire. Le travail d'adaptation et celui de prise en compte des besoins requis pour intégrer l'utilisation de l'ontologie dans l'application cible est loin d'être négligeable.

Les résultats que nous avons obtenus jusqu'à présent concernant l'ontologie et l'outil nous encouragent à poursuivre nos recherches et à améliorer les solutions que nous proposons, ce que nous ferons dans le cadre de notre projet postdoctoral MedOC.

Références

ARPIREZ J., GÓMEZ-PÉREZ A., LOZANO A. & PINTO S. (1998). (onto)2agent : An ontology-based www broker to select ontologies. In *Workshop on Applications of Ontologies and Problem Solving Methods (ECAI)*, Brighton, United Kingdom.

ASSADI H. & BOURIGAULT D. (2000). Analyses syntaxique et statistique pour la construction d'ontologies à partir de textes. In J. CHARLET, M. ZACKLAD, G. KASSEL & D. BOURIGAULT, Coordinateurs, *Ingénierie des connaissances - Evolutions récentes et nouveaux défis*. Eyrolles.

AUSSENAC N. (1989). *Conception d'une méthodologie et d'un outil d'acquisition de connaissances expertes*. PhD thesis, Université Paul Sabatier, Toulouse III.

AUSSENAC N. & SEGUELA P. (2000). Les relations sémantiques : du linguistique au formel. In A. CONDAMINES, Coordinateur, *Cahiers de grammaire, Numéro spécial sur la linguistique de corpus*, volume 25, p. 175–198. Toulouse, France : Presse de l'UTM.

A.-G. N. AUSSENAC-GILLES, B. BIEBOW & S. SZULMAN, Coordinateurs (2000). *Proceedings of the EKAW'2000 Workshop on Ontologies and Texts*, volume 51. CEUR Workshop Proceedings. Available at `http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-51/`.

AUSSENAC-GILLES N. (2005). *Méthodes ascendantes pour l'Ingénierie des connaissances*. Rapport interne, Habilitation à diriger des recherches, Institut de Recherche en Informatique de Toulouse (IRIT), Université Paul Sabatier, Toulouse III. Disponible à `http://tel.archives-ouvertes.fr/action/open_file.php?url=http://tel.archives-ouvertes.fr/docs/00/08/91/65/PDF/HDR13fevrier2006.pdf&docid=89165`.

AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S. (2000). Corpus analysis for conceptual modelling. In N. AUSSENAC-GILLES, B. BIÉBOW & S. SZULMAN, Coordinateurs, *Workshop "Ontologies and Text" associated to the 12^{th} International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, p. 13–27, Juan-les-Pins, France.

AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S. (2002). Terminae. In A. GÓMEZ-PÉREZ & V. BENJAMINS, Coordinateurs, *Workshop on Evaluation of Ontology Engineering Environments, Knowledge Engineering and Knowledge Management : Methods, Models and Tools, 13^{th} International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, volume 62 of *Lecture Notes in Artificial Intelligence*, Sigüenza, Espagne : Springer.

AUSSENAC-GILLES N., BIÉBOW B. & SZULMAN S. (2005). *Ingénierie des connaissances*, chapitre Modélisation du domaine par une méthode fondée sur l'analyse de corpus, p. 49–71. L'Harmattan.

AUSSENAC-GILLES N., BOURIGAULT D. & TEULIER R. (2003). Analyse comparative de corpus : cas de l'ingénierie des connaissances. In *14^{e} journées Francophones d'Ingénierie des Connaissances (IC)*, p. 67–84, Laval, France.

AUSSENAC-GILLES N. & CONDAMINES A. (2003). *Rapport de l'action spécifique ASSTICCOT*. Rapport interne, CNRS. Action Spécifique STIC « Corpus et Terminologie » (AS 34), rattachée au RTP 33 (RTP-DOC). Disponible à `http://rtp-doc.enssib.fr/article.php3?id_article=40`.

AUSSENAC-GILLES N. & SÖRGEL D. (2005). Text analysis for ontology and terminology engineering. *Applied Ontology*, **1**, 35–46.

BAADER F., HORROCKS I. & SATTLER U. (2003a). *Description logics as ontology languages for the semantic web.* Lecture Notes in Articial Intelligence. Springer-Verlag. Festschrift in honor of Jörg Siekmann. Available at `http://www.cs.man.ac.uk/~horrocks/Publications/download/2003/BaHS03.pdf`.

F. BAADER, D. MCGUINESS, D. NARDI & P. PATEL-SCHNEIDER, Coordinateurs (2003b). *The Description Logic Handbook : Theory, Implementation and Applications.* Cambridge University Press.

BAADER F. & NUTT W. (2003). *The Description Logic Handbook : Theory, Implementation and Applications*, chapitre Basic description logics, p. 47–100. Cambridge University Press.

BACHIMONT B. (2000). Engagement sémantique et engagement ontologique : conception et réalisation d'ontologies en ingénierie des connaissances. In R. TEULIER, J. CHARLET & P. TCHOUNIKINE, Coordinateurs, *Ingénierie des connaissances*, chapitre 19. Paris : L'Harmattan. Article réédité en 2005 dans le cédérom associé au livre.

BACHIMONT B. (2001). Modélisation linguistique et modélisation logique des ontologies : l'apport de l'ontologie formelle. In *12^{e} Journées francophones d'Ingénierie des Connaissances (IC)*, p. 349–368, Grenoble, France : Presses Universitaires de Grenoble.

BACHIMONT B., ISAAC A. & TRONCY R. (2002). Semantic commitment for designing ontologies : A proposal. In A. GOMEZ-PÉREZ & V. BENJAMINS, Coordinateurs, *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, volume LNAI 2473 of *Lecture Notes in Artificial Intelligence*, p. 114–121 : Springer.

BANEYX A., CHARLET J. & JAULENT M. (2005a). Building medical ontologies based on terminology extraction from texts : an experimentation in pneumology. In R. ENGELBRECHT, A. GEISSBUHLER, C. LOVIS & G. MIHALAS, Coordinateurs, *Connecting Medical Informatics and Bio-Informatics – Proceedings of the XIXth International Congress of the European Federation for Medical Informatics (MIE)*, volume 116 of *Studies in Health Technology and Informatics*, p. 659–664, Geneva, Switzerland : IOS Press. ISBN : 1-58603-549-5.

BANEYX A., LY B. & CHARLET J. (2005b). *Cahier des charges et scénarios d'usage de MedCKARE*. Rapport interne, Université Pierre et Marie Curie, Paris 6.

BANEYX A., MALAISÉ V., CHARLET J., ZWEIGENBAUM P. & BACHIMONT B. (2005c). Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles. In *6^{e} rencontres - Terminologie et Intelligence Artificielle (TIA)*, p. 31–42, Rouen, France.

BECHHOFER S., HORROCKS I., GOBLE C. & STEVENS R. (2001). OilEd : a reason-able ontology editor for the semantic web. In *Proceedings of KI2001, Joint German/Austrian conference on Artificial*

Intelligence, number 2174 in Lecture Notes in Computer Science, p. 396–408, Vienna : Springer-Verlag. Available at http://potato.cs.man.ac.uk/papers/ki2001.pdf.

BELLATRECHE L., XUAN D., PIERRA G. & DEHAINSALA H. (2006). Contribution of ontology-based data modeling to automatic integration of electronic catalogues within engineering databases. *Computers in Industry Journal*. To appear.

BENSADOUN H. (2001). Pmsi et chirurgiens : pourquoi les chirurgiens doivent-ils coder, comment bien coder ? *Journal de Chirurgie, Masson*, **138**(1).

BERNERS-LEE T., HENDLER J. & LASSILA O. (2001). The semantic web - a new form of web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American*.

BLASQUEZ M., FERNANDEZ M., GARCIA-PINAR J. & GOMEZ-PEREZ A. (1998). Building ontologies at the knowledge level using the ontology design environment. In *Banff Workshop on Knowledge Acquisition for Knowledge-based Systems*.

BLOIS M. (1984). *Information and Medicine : The Nature of Medical Descriptions*. Univ of California Press.

BORGO S., GUARINO N. & MASOLO C. (1996). Stratified ontologies : the case of physical objects. In *Workshop on Ontological Engineering (ECAI)*, Budapest.

BORST W. (1997). *Construction of Engineering Ontologies*. PhD thesis, University of Twente, Enschede, The Netherlands. Available at http://purl.org/utwente/fid/1392.

BOUAUD J., BACHIMONT B., CHARLET J. & ZWEIGENBAUM P. (1994). Acquisition and structuring of an ontology within conceptual graphs. In *Proceedings of ICCS'94 Workshop on Knowledge Acquisition using Conceptual Graph Theory*, p. 1–25, University of Maryland, College Park, MD.

BOUAUD J. & ZWEIGENBAUM P. (1992). A reconstruction of conceptual graphs on top of a production system. In *Workshop on Conceptual Graphs*, p. 127–136.

BOURIGAULT D. (1994a). Extraction et structuration automatiques de terminologie pour l'aide à l'acquisition de connaissances à partir de textes. In *9ème congrès Reconnaissance des Formes et Intelligence Artificielle (RFIA'94)*, p. 1123–1132, Paris.

BOURIGAULT D. (1994b). *LEXTER, un logiciel d'extraction de terminologie. Application à l'acquisition de connaissances à partir de textes*. PhD thesis, École des hautes études en sciences sociales, Paris.

BOURIGAULT D. (2002). Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues*, p. 75–84, Nancy, France. http ://www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/TALN02-Bourigault.doc.

BOURIGAULT D. & AUSSENAC-GILLES N. (2003). Construction d'ontologies à partir de textes. In *Actes de la conférence sur le traitement automatique des langues (TALN)*, p. 27–50, Batz-sur-Mer, France.

BOURIGAULT D., AUSSENAC-GILLES N. & CHARLET J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. In J. PIERREL & M. SLODZIAN, Coordinateurs, *Revue d'Intelligence Artificielle.*, volume 18, p. 87–110. Hermès. Numéro spécial sur les Techniques Informatiques et Structuration de Terminologies, disponible à http://www.univ-tlse2.fr/erss/textes/pagespersos/bourigault/RIA-bourigault-aussenac-charlet.doc.

BOURIGAULT D., FABRE C., FRÉROT C., JACQUES M.-P. & OZDOWSKA S. (2005). Syntex, analyseur syntaxique de corpus. In *Actes des 12èmes journées sur le Traitement Automatique des Langues Naturelles*, Dourdan, France. Disponible à http://w3.univ-tlse2.fr/erss/membres/bourigault/TALN05-bourigault-Syntex.pdf.

BOURIGAULT D. & FRÉROT C. (2004). Ambiguïté de rattachement prépositionnel : introduction de ressources exogènes de sous-catégorisation dans un analyseur syntaxique de corpus endogène. In *Actes des 11èmes journées sur le Traitement Automatique des Langues Naturelles*, Fès, Maroc. Disponible à http://www.univ-tlse2.fr/erss/textes/pagespersos/frerot/Publications/BourigaultFrerot_taln04.pdf.

BOURIGAULT D. & JACQUEMIN C. (2000). Construction de ressources terminologiques. In J.-M. PIERREL, Coordinateur, *Ingénierie des langues*, Traité IC2, chapitre 9, p. 215–233. Paris : Hermès.

BOURIGAULT D. & LAME G. (2002). Analyse distributionnelle et structuration de terminologie : application à la construction d'une ontologie documentaire du droit. In *Traitement automatique des langues (TAL)*, volume 43, p. 129–150.

BRACHMAN R. & LEVESQUE H. (1984). The tractability of subsumption in frame-based description languages. In *The National Conference of the American Association on Artificial Intelligence (AAAI)*, p. 34–37, Austin (Texas), United-States.

BREWSTER C., ALANI H., DASMAHAPATRA S. & WILKS Y. (2004). Data driven ontology evaluation. In *International Conference on Language Resources and Evaluation*, Lisbon, Portugal.

CARABALLO S. A. (1999). Automatic construction of a hypernym-labeled noun hierarchy from text. In *Meeting of the Association for Computational Linguistics (ACL)*, p. 120–126, Maryland, USA.

CHABERT-RANWEZ S. (2000). *Composition automatique de documents hypermédia adaptatifs à partir d'ontologies et de requêtes intentionnelles de l'utilisateur*. PhD thesis, Université Montpellier II - Sciences et Techniques du Languedoc, Montpellier, France. Disponible à http://archive-edutice.ccsd.cnrs.fr/edutice-00000381/en/.

CHARBONNEL P., HEBBRECHT G., MAUREL-ARRIGHI E., COUTANT D., LEGENDRE J., GOREL J., ROLAND M., JAMOULLE M., DE PALMA A. & REISS M. (1996). Le codage en médecine : les enjeux. *Pratiques*, **46**, 4–30.

CHARLET J. (2002). *L'Ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales*. Rapport interne, Habilitation à diriger des recherches, Université Paris 6. Disponible à http://archivesic.ccsd.cnrs.fr/sic_00001064 le 20.09.06.

CHARLET J. (2005). L'ingénierie des connaissances, une science de gestion ? Recherches, chapitre 11. La découverte. Version longue disponible à http://archivesic.ccsd.cnrs.fr/sic_00000805.

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

CHARLET J., BACHIMONT B. & JAULENT M. (2006). Building medical ontologies by terminology extraction from texts : An experiment for the intensive care units. *Computer in Biology and Medicine*, **36**(7-8), 857–870. ISSN : 0010-4825.

J. CHARLET, P. LAUBLET & C. REYNAUD, Coordinateurs (2005). *Le Web sémantique*, volume 4. Toulouse : Cépaduès. Hors série diponible à http://www.revue-i3.org/. ISBN : 2.85428.666.9.

J. CHARLET, C. REYNAUD & P. LAUBLET, Coordinateurs (2004). *Web sémantique*. Hors série de la revue I3 : information - interaction - intelligence. Cépaduès. Disponible à http://www.revue-i3.org/hors_serie/annee2004/index.html.

CHEIN M. & MUGNIER M.-L. (1992). Conceptual graphs : Fundamental notions. *Revue d'Intelligence Artificielle*, **6**(4), 365–406. Available at http://www.lirmm.fr/~mugnier/ArticlesPostscript/RIA92ChMu.ps.

CIMINO J., CLAYTON P., HRIPCSAK G. & JOHNSON S. (1994). Knowledge-based approaches to the maintenance of a large controlled medical terminology. *J Am Med Inform Assoc.*, **1**(1), 35–50.

CLEMMER T. (1995). The role of medical informatics in telemedecine. *Journal of Medical System*, **19**, 45–58.

COLLEGE OF AMERICAN PATHOLOGISTS (1993). *Systematized Nomenclature of Human and Veterinary Medicine : SNOMED International*. College of American Pathologists.

CONDAMINES A. (2003). *Sémantique et corpus spécialisé : constitution de bases de connaissances terminologiques*. Habilitation à diriger des recherches, Université de Toulouse Le Mirail.

CORCHO O., FERNÁNDEZ-LÓPEZ M., GÓMEZ-PÉREZ A. & LÓPEZ-CIMA A. (2005). *Law and the Semantic Web. Legal Ontologies, Methodologies, Legal Information Retrieval, and Applications*, chapitre Building Legal Ontologies with METHONTOLOGY and WebODE, p. 142–157. Springer-Verlag. Available at http://www.cs.man.ac.uk/~ocorcho/documents/LawSemWeb2004_CorchoEtAl.pdf.

CORCHO O., FERNÁNDEZ-LÓPEZ M., GÓMEZ-PÉREZ A. & VICENTE O. (2002). Webode : an integrated workbench for ontology representation, reasoning and exchange. In *The 13th International Conference on Knowledge Engineering an Knowledge Management (EKAW)*, volume 2473 of *Lecture Notes on Artificial Intelligence*, p. 138–153 : Springer-Verlag. Available at http://webode.dia.fi.upm.es/WebODEWeb/Documents/EKAW2002_CorchoEtAl.pdf.

CORCHO O. & GÓMEZ-PÉREZ A. (2000). Evaluating knowledge representation and reasoning capabilities of ontology specification languages. In *Workshop on Applications of Ontologies and Problem Solving Methods (ECAI)*, Berlin, Allemagne. Available at http://www.cs.man.ac.uk/~ocorcho/documents/ECAI00WS_CorchoGomezPerez.pdf.

R. A. CÔTÉ, D. J. ROTHWELL, J. L. PALOTAY, R. S. BECKETT & L. BROCHU, Coordinateurs (1993). *The Systematised Nomenclature of Human and Veterinary Medicine : SNOMED International*. Northfield : College of American Pathologists.

CRISTIANI M. & CUEL R. (2005a). *Encyclopedia of Knowledge Management*, chapitre Ontology Development Methodologies. Idea Group Reference.

CRISTIANI M. & CUEL R. (2005b). Ontology methodologies : a survey. *International Journal on Semantic Web and Information Systems*, **1**(2), 49–69. Invited paper.

DAMERON O. (2003). *Modélisation, représentation et partage de connaissances anatomiques sur le cortex cérébral*. PhD thesis, Faculté de médecine de Rennes I. Disponible à http://sim3.univ-rennes1.fr/~odameron/thesis/dameronThesis.pdf.

DARLU P. & TASSY P. (1993). *La reconstruction phylogénétique. Concepts et méthodes.* Paris, France : Masson. Disponible à http://lis.snv.jussieu.fr/sfs/pdf/Darlu_Tassy_online.pdf, ISBN : 2-225-84229-9.

DARMONI S., JARROUSSE E., ZWEIGENBAUM P., LE BEUX P., NAMER F., BAUD R., JOUBERT M., VALLEE H., COTE R., BUEMI A., BOURIGAULT D., RECOURCE G., JEANNEAU S. & RODRIGUES J. (2003). Vumef : Extending the french involvement in the umls metathesaurus. In *Proc AMIA Symp.*, p. 824. Available at http://www.vidal.fr/vumef/fichiers/PublicationsDiffusees/AMIA2003posterVUMeF\%202.pdf.

DARMONI S.-J., LEROY J.-P., BAUDIC F., DOUYÈRE M., PIOT J. & THIRION B. (2000). CISMeF : a structured health resource guide. *Methods of Information in Medicine*, **39**(1).

DARMONI S.-J. & THIRION B. (1996). Indexing the web ? a comparative study of three medical web servers on the internet : Cliniweb, "diseases, disorders and related topics", omni. In *European Congress of the Internet in Medicine (Mednet)*, Brighton, Royaume-Uni. Available at http://www.chu-rouen.fr/dsii/publi/mdntdl4.html.

DARMONI S.-J. & THIRION B. (2000). A standard metadata scheme for health resources. *J Am Med Inform Assoc*, **7**(1), 108–109.

DAVIS R., SHROBE H. & SZOLOVITS P. (1993). What is a knowledge representation ? *AI Magazine*, **14**(1), 17–33. Available at http://groups.csail.mit.edu/medg/ftp/psz/k-rep.html.

DEGOULET P. & FIESCHI M. (1991). *Traitement de l'information médicale : Méthodes et applications hospitalières*, chapitre Informatisation des dossiers médicaux. Manuels Informatiques. Masson Entreprise.

DIENG R., CORBY O., GANDON F., GIBOIN A., GOLEBIOWSKA J., MATTA N. & RIBIÈRE M. (2001). *Méthodes et outils pour la gestion des connaissances : une approche pluridisciplinaire du knowledge management*. Dunod, 2 edition.

DOLIN R., SPACKMAN K., ABILLA A., CORREIA C., GOLD-BERG B., KONICEK D., LUKOFF J. & LUNDBERG C. (2001). The SNOMED-RT procedure model. In *Proc AMIA Symp*, p. 139–143.

DONINI F., LENZERINI M., NARDI D. & NUTT W. (1991). Tractable concept languages. In *The International Joint Conference on Artificial Intelligence (IJCAI)*, p. 458–463. Best Paper Award.

ERMINE J. (2000). *Les systèmes de connaissances*. Hermès-Lavoisier, 2^e édition edition. ISBN : 2-7462-0159-3.

FARQUHAR A., FIKES R. & RICE J. (2000). Ontolingua server : a tool for collaborative ontology construction. *International Journal of Human-Computer studies*, **46**, 707–727. Available at http://www.cs.umbc.edu/771/papers/KSL-96-26.pdf.

FERNANDEZ M., GÓMEZ-PÉREZ A. & JURISTO N. (1997). Methontology : from ontological art towards ontological engineering. In *Spring Symposium Series on Ontological Engineering, National Conference of the American Association on Artificial Intelligence (AAAI)*.

FOSKETT D. (1997). *Readings in Information Retrieval*, chapitre Thesaurus, p. 111–134. Morgan Kaufmann : San Francisco, California, USA.

FOURNIER-VIGER P. (2005). Un modèle de représentation des connaissances à trois niveaux de sémantique pour les systèmes tutoriels intelligents. Master's thesis, Université de Sherbrooke, Sherbrooke, Canada. Adaptation du chapitre 4 disponible à http://www.philippe-fournier-viger.com/introduction_logiques_de_description.htm.

FRIEDMAN C., ALDERSON P., AUSTIN J., CIMINO J. & JOHNSON S. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, **1**(2), 161–74.

FRIEDMAN C., SHANIGA L., LUSSIER Y. & HRIPSACK G. (2004). Automated encoding of clinical documents based on natural language processing. *JAMIA*, **11**, 392–402.

FÜRST F. (2004a). *Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationalisation*. PhD thesis, École Polytechnique de l'Université de Nantes (EPUN), Nantes, France. Disponible à `http://www.sciences.univ-nantes.fr/info/perso/permanents/furst/papers/theseFurst.pdf`.

FÜRST F. (2004b). Opérationalisation des ontologies : une méthode et un outil. In *Ingénierie des Connaissances (IC)*, p. 199–210, Lyon, France. Disponible à `http://www.sciences.univ-nantes.fr/lina/fr/documents/ra-LINA.pdf`.

GANDON F. (2006). Ontologies informatiques. *Interstices, Journal en ligne de l'INRIA*. Disponible à `http://interstices.info/display.jsp?id=c_17672`.

GANGEMI A., GUARINO N., MASOLO C., OLTRAMARI A. & SCHNEIDER L. (2002). Sweetening ontologies with dolce. In *Ontologies and the Semantic Web : 13th International Conference (EKAW)*.

GENE ONTOLOGY CONSORTIUM (2001). Creating the gene ontology resource : design and implementation. *Genome Research*, **11**, 1425–1433.

GILDEA D. & JURAFSKY D. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, **28**(3), 245–288.

GÓMEZ-PÉREZ A. (2000). Développements récents en matière de conception, de maintenance et d'utilisation d'ontologies. *Terminologies Nouvelles*, **19**, 9–20. Traduit de l'anglais par S. Descotte.

GÓMEZ-PÉREZ A. (2004). In S. STAAB & R. STUDER, Coordinateurs, *Handbook on Ontologies*, chapitre Ontology Evaluation, p. 251–275. Handbooks in Information Systems. Springer.

GÓMEZ-PÉREZ A., FERNANDEZ M. & DE VICENTE A. (1996). Towards a method to conceptualize domain ontologies. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, p. 41–52.

GÓMEZ-PÉREZ A., FERNANDEZ-LOPEZ M. & CORCHO O. (2004a). *Ontological Engineering*. Advanced Information and Knowledge Processing. Madrid, Spain : Springer, 2 edition.

GÓMEZ-PÉREZ A., FERNANDEZ-LOPEZ M. & CORCHO O. (2004b). *Ontological Engineering*, chapitre Methodologies and Methods for Building Ontologies, p. 107–196. Springer.

GÓMEZ-PÉREZ A., FERNANDEZ-LOPEZ M., CORCHO O., AHN T., AUSSENAC-GILLES N., BERNARDOS S., CHRISTOPHIDES V., CORBY O., CROWTHER P., DING Y., ENGELS R., ESTEBAN M., GANDON F., KALFOGLOU Y., KARVOUNARAKIS G., LAMA M., LOPEZ A., LOZANO A., MAGKANARAKI A., MANZANO D., MOTTA E., NOY N., PLEXOUSAKIS D., RAMOS J. & SURE Y. (2002). *OntoWeb Technical Roadmap*. Rapport interne OntoWeb deliverable 1.1.2, Universidad Politecnia de Madrid. Available at `http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/OntoWeb_Del_1-1-2.pdf`.

GRABAR N. & ZWEIGENBAUM P. (2000). Automatic acquisition of domain-specific morphological resources from thesauri. In *Actes de RIAO 2000 : Accès à l'Information Multimédia par le Contenu*, p. 765–784, Paris, France.

GRENON P. (2003). *BFO in a Nutshell : A Bi-categorial Axiomatization of BFO and Comparison with DOLCE*. Rapport interne, Faculty of Medicine, Institute for Formal Ontology and Medical Information Science (IFOMIS). Available at `http://www.ifomis.org/Research/IFOMISReports/IFOMIS\%20Report\%2006_2003.pdf`.

GRUBER T. (1993). A translation approch to portable ontology specifications. *Knowledge Acquisition*, **5**(2), 199–220.

GUARINO N. (1996). Understanding, building and using ontologies. In *Workshop on Knowledge Acquisition for Knowledge-Based Systems*. Available at http://ksi.cpsc.ucalgary.ca/KAW/KAW96/guarino/guarino.html.

GUARINO N. (1997). Some organizing principles for a unified top-level ontology. In *National Conference of the American Association on Artificial Intelligence (AAAI)*, p. 57–63, Stanford, United-States.

GUARINO N. (1998). Formal ontology in information systems. In N. GUARINO, Coordinateur, *1st International Conference on Formal Ontology in Information Systems (FOIS)*, p. 3–15, Trento, Italy : IOS Press.

GUARINO N. (1999). The role of identity conditions in ontology design. In V. BENJAMINS, B. CHANDRASEKARAN, A. GOMEZ-PEREZ, N. GUARINO & M. USCHOLD, Coordinateurs, *Proc. of the IJCA'99 Workshop on Ontologies and Problem-Solving Methods*, p. 2/1–2/7, Sweden. Available at http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-18/2-guarino.pdf.

GUARINO N., CARRARA M. & GIARETTA P. (1994). Formalizing ontological commitments. In *National Conference of the American Association on Artificial Intelligence (AAAI)*, p. 560–567.

GUARINO N., GANGEMI A., MASOLO C. & OLTRARI A. (1995). Understanding top-level ontological distinctions. In *Workshop on Basic Ontological Issues in Knowledge Sharing, The International Joint Conference on Artificial Intelligence (IJCAI)*.

GUARINO N. & WELTY C. (2000a). A Formal Ontology of Properties. In R. DIENG & O. CORBY, Coordinateurs, *12th International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, volume (1937) of *Lecture Notes in Computer Science*, p. 97–112, Juan-les-Pins, France : Springer Verlag.

GUARINO N. & WELTY C. (2000b). Identity, unity, and individuality : Towards a formal toolkit for ontological analysis. In W. HORN, Coordinateur, *The 14th European Conference on Artificial Intelligence*, p. 219–223 : Berlin : IOS Press. Available at http://www.ladseb.pd.cnr.it/infor/Ontology/Papers/LADSEB02-2000.pdf.

HARRIS Z. (1968). *Mathematical Structures of Language*. New-York, USA : John Wiley and Sons.

HEARST M. (1992). Automatic acquisition of hyponyms from large text corpora. In *The 14th conference on Computational linguistics*, volume 2, p. 539–545, Nantes - France : Association for Computational Linguistics. Available at http://www.cs.mu.oz.au/acl/C/C92/C92-2082.pdf.

HORROCKS I., PATEL-SCHNEIDER P. & VAN HARMELEN F. (2003). From shiq and rdf to owl : The making of a web ontology language. *Journal of Web Semantics*, **1**(1), 7–26. Available at http://www.cs.man.ac.uk/~horrocks/Publications/download/2003/HoPH03a.pdf.

HOUSSET B. (1999). *Abrégé de Pneumologie*. Abrégés. Paris : Masson.

JEAN S., PIERRA G. & AÏT-AMEUR Y. (2006). Domain ontologies : a database-oriented analysis. In *Web Information Systems and Technologies (WEBIST'2006)*. Disponible à http://www.lisi.ensma.fr/ftp/pub/documents/papers/2006/2006-WEBIST-Jean.pdf.

JOHANSSON I. (2005). Qualities, quantities, and the endurant-perdurant distinction in top-level ontologies. In *Wissensmanagement 2005 : Professional Knowledge Management Experiences and Vision*, p. 543–550. Available at http://hem.passagen.se/ijohansson/information2.PDF.

KASSEL G. (2002). Ontospec : une méthode de spécification semi-formelle d'ontologies. In *13e journées francophones d'Ingénierie des Connaissances (IC)*, p. 75–87, Rouen, France.

KAYSER D. (1997). *La représentation des connaissances*. Informatique. Paris, France : Hermes. ISBN 2-86601-647-5.

KNUBLAUCH H., FERGERSON R., NOY N. & MUSEN M. (2004). The protégé owl plugin : An open development environment for semantic web applications. In S. MCILRAITH, D. PLEXOUSAKIS & F. VAN HARMELEN, Coordinateurs, *The Third International Semantic Web Conference (ISWC)*, volume 3298 of *Lecture Notes in Computer Science*, p. chapter p. 229, Hiroshima, Japan : Springer. Available at `http://smi-web.stanford.edu/auslese/smi-web/reports/SMI-2004-1011.pdf`.

KOHLER F., MAYEUX D., MUSSE J. & LOMBRAIL P. (1990). Informatique et aide au codage du p.m.s.i. In *Gestions hospitalières*, number 299, p. 662–666.

LAME G. (2002). *Construction d'ontologies à partir de textes. Une ontologie du droit dédiée à la recherche d'information sur le Web*. Thèse de doctorat, École des Mines. Disponible à `http://www.cri.ensmp.fr/classement/doc/A-345.ps`.

LE BOZEC C. (2001). *Gestion des connaissances multi-experts en imagerie médicale. IDEM : images et diagnostics par l'exemple en médecine*. PhD thesis, Université Paris 6.

LE MOIGNO S., CHARLET J., BOURIGAULT D., DEGOULET P. & JAULENT M. (2002a). Terminology extraction from text to build an ontology in surgical intensive care. In *Proc AMIA Symp*, p. 430–435. Available at `http://www-test.biomath.jussieu.fr/~jc/Files/LemoignoAMIA2002.pdf`.

LE MOIGNO S., CHARLET J., BOURIGAULT D. & JAULENT M.-C. (2002b). Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale. In *13e journées francophones d'ingénierie des Connaissances (IC)*, Rouen.

LECOINTRE G. & LE GUYADER H. (2001). *La Classification phylogénétique du vivant*. Belin. ISBN : 2-7011-2137-X.

LEFÈVRE P. (2000). *La Recherche d'informations. Du texte intégral au thésaurus.* Paris, France : Hermès-Lavoisier, 1er édition edition. ISBN : 2-7462-0173-9.

LINDBERG D. A. B. & HUMPHREYS B. L. (1990). The UMLS knowledge sources : tools for building better user interfaces. In *Proceedings of the 14th Annual Symposium on Computer Applications in Medical Care*, p. 121–125.

LOPEZ M. F., GÓMEZ-PÉREZ A., SIERRA J. P. & SIERRA A. P. (1999). Building a chemical ontology using methontology and the ontology design environment. *IEEE Intelligent Systems*, **14**(1), 37–46. Available at `http://www.aifb.uni-karlsruhe.de/Lehrangebot/Sommer2001/SemanticWeb/papers/chemical_ontology.pdf`.

LOZANO-TELLO A. & GÓMEZ-PÉREZ A. (2004). Ontometric : A method to choose the appropriate ontology. *Journal of Database Management*, **15**(2).

MAEDCHE A. & STAAB S. (2001). Ontology learning for the semantic web. In *IEEE Intelligent Systems*, volume 16.

MAEDCHE A. & STAAB S. (2004). *Handbook on Ontologies*, chapitre Ontology Learning, p. 173–190. Handbooks in Information Systems. Springer.

MALAISÉ V. (2005). *Méthodologie linguistique et terminologique pour la structuration d'ontologies différentielles à partir de corpus textuels*. PhD thesis, Université Paris 7 - Denis Diderot.

MALAISÉ V., ZWEIGENBAUM P. & BACHIMONT B. (2004). Repérage et exploitation d'énoncés définitoires corpus pour l'aide à la construction d'ontologie. In *11e Conférence Annuelle sur le Traitement Automatique des Langues (TALN)*, p. 269–278, Fès, Maroc.

MINSKY M. (1981). *Mind Design.*, chapitre A framework for representing knowledge., p. 95–128. The MIT Press. ISBN : 0262081105.

MIZOGUCHI R. (1998). A step towards ontological engineering. In *12th National Conference on AI of JSAI*, p. 24–31. Traduit du japonais. Available at `http://www.ei.sanken.osaka-u.ac.jp/english/step-onteng.html`.

MIZOGUCHI R. & IKEDA M. (1997). Towards ontology engineering. In *The Joint 1997 Pacific Asian Conference on Expert systems - International Conference on Intelligent Systems*, p. 259–266, Singapore.

MIZOGUCHI R., VANWELKENHUYSEN J. & IKEDA M. (1995). Task ontology for reuse of problem solving knowledge. In N. MARS, Coordinateur, *Towards Very Large Knowledge Bases : Knowledge Building and Knowledge Sharing (KBKS)*, p. 46–57, The Netherlands : University of Twente IOS Press.

MORIN E. (1999). Des patrons lexico-syntaxiques pour aider au dépouillement terminologique. *Traitement Automatique des Langues*, **40**(1).

MUGNIER M.-L. & CHEIN M. (1996). Représenter des connaissances et raisonner avec des graphes. *Revue d'Intelligence Artificielle*, **10**(1), 7–56. Disponible à `http://www.lirmm.fr/~mugnier/ArticlesPostscript/RIA96MuCh.ps`.

NAMER F. & ZWEIGENBAUM P. (2004). Acquiring meaning for french medical terminology : contribution of morphosemantics. In M. FIESCHI, E. COIERA & Y.-C. J. LI, Coordinateurs, *10^{th} World Congress on Medical Informatics*, p. 535–539, San Francisco. Available at `http://www-test.biomath.jussieu.fr/~pz/FTPapiers/Namer_MEDINFO2004.pdf`.

NARDI D. & BRACHMAN R. (2003). *The Description Logic Handbook : Theory, Implementation and Applications.*, chapitre An introduction to description logics., p. 544. Cambridge University Press.

NAVIGLI R. & VELARDI P. (2004). Automatic ontology learning : Supporting a per-concept evaluation by domain experts. In *Workshop on Ontology Learning and Population (ECAI)*, Valence, Espagne.

NELSON S. J., JOHNSTON D. & HUMPHREYS B. L. (2001). Relationships in medical subject headings. In C. A. BEAN & R. GREEN, Coordinateurs, *Relationships in the organization of knowledge*, New York : Kluwer Academic Publishers.

NOY N. & MUSEN M. (2003). The prompt suite : Interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, **59**(6), 983–1024. Available at `http://smi-web.stanford.edu/auslese/smi-web/reports/SMI-2003-0973.pdf`.

NÉVÉOL A. (2005). *Automatisation des tâches documentaires dans un catalogue de santé en ligne*. PhD thesis, Institut National des Sciences Appliquées de Rouen. Disponible à `http://aurelie.neveol.free.fr/theseAN.pdf`.

ORGANISATION MONDIALE DE LA SANTÉ (1995). *Manuel d'utilisation de la CIM-10- Classification statistique internationale des maladies et des problèmes de santé connexes*. Genève, Suisse, 10^{me} révision edition. volume 2, 167 pages. Disponible à `http://www.spieao.uhp-nancy.fr/~kohler/CIM10/CIM10Som.htm`.

PEREIRA S., MASSARI P., JOUBERT M. & DARMONI S. (2007). Utilisation de métatermes pour la recherche d'information dans les dossiers médicaux. In C. O. BAGAYOKO, Coordinateur, *12^{e} Journées Francophones Informatique Médicale (JFIM)*, Bamako, Mali. À paraître.

PEREIRA S., NÉVÉOL A., MASSARI P., JOUBERT M. & DARMONI S. (2006). Construction of a semi-automated icd-10 coding help system to optimize medical and economic coding. In A. HASMAN, R. HAUX, J. VAN DER LEI, E. DE CLERCQ & F. ROGER-FRANCE, Coordinateurs, *Ubiquity : Technologies for Better Health in Aging Societies – Proceedings of the* XXth *International Congress of the European Federation for Medical Informatics (MIE)*, volume 124 of *Studies in Health Technology and Informatics*, p. 845–850, Maastricht, Pays-Bas : IOS Press. Available at `http://cybertim.timone.univ-mrs.fr/recherche/doc-recherche/informatique/JOUBERTMIE2006/publication_file`, ISBN 978-1-58603-647-8.

PIERRA G., DEHAINSALA H., AÏT-AMEUR Y. & BELLATRECHE L. (2005). Base de données à base ontologique : principe et mise en oeuvre. In *Ingénierie des systèmes d'information*, volume 10. Hermès. Disponible à `http://www.lisi.ensma.fr/ftp/pub/documents/papers/2005/2005-I-Pierra.pdf`.

PORZEL R. & MALAKA R. (2004). A task-based approach for ontology evaluation. In *Workshop on Ontology Learning and Population (ECAI)*. Available at `http://olp.dfki.de/ecai04/final-porzel.pdf#search=%22robert%20porzel%20%22ontology%20evaluation%22%22`.

RASTIER F., CAVAZZA M. & ABEILLÉ A. (1994). *Sémantique pour l'analyse - De la linguistique à l'informatique*. Enseignement de. Paris, France : Masson.

REBEYROLLE J. (2000). Utilisation de contextes définitoires pour l'acquisition de connaissances à partir de textes. In *Journées Francophones d'Ingénierie des Connaissances (IC'2000)*, p. 105–114, Toulouse.

RECTOR A. (1998). Thesauri and formal classifications : Terminologies for people and machines. *Methods Inf Med*, **37**(4-5), 501–509.

RECTOR A. L., NOWLAN W. A. & KAY S. (1992). Conceptual knowledge : the core of medical information systems. In K. C. LUN, P. DEGOULET, T. PIEMME & O. RIENHOFF, Coordinateurs, *Proc MEDINFO 92*, p. 1420–1426, Geneva : North Holland.

RENARD J., BEUSCARD R., DELERUE D. & GEIB J. (2000). Le réseau ville-hôpital : une nouvelle forme de communication entre professionnels de santé. In P. DEGOULET & M. FIESCHI, Coordinateurs, *Revue européenne de biotechnologie médicale*, volume 21 of *Innovation et technologie en biologie et médecine*, p. 275–280. Springer-Verlag.

RODRIGUES J.-M., TROMBERT-PAVIOT B., BAUD R., WAGNER J. & MEUSNIER F. (1998). Galen-In-Use : Using artificial intelligence terminology tools to improve the linguistic coherence of a national coding system for surgical procedures. In *Proc 9 th World Congress on Medical Informatics*, Studies in Health Technology and Informatics, p. 623–627, Amsterdam.

RODRIGUES J.-M., TROMBERT-PAVIOT B., RECTOR A., BAUD R., CLAVEL L., ABRIAL V., IDIR H. & VERY J. (1999). GALEN, il existe quelque chose après les mots : leur signification et au delà le savoir médical. *Innovation Stratégique en Information de Santé (ISIS)*, (2-3), 48–62.

ROGERS J., ROBERTS A., SALOMON D., VAN DER HARING E., WROE C., ZANSTRA P. & RECTOR A. L. (2001). GALEN ten years on : Tasks and supporting tools. In R. HAUX, R. ROGERS & V. PATEL, Coordinateurs, *Proc 10 th World Congress on Medical Informatics*, Studies in Health Technology and Informatics, p. 256–260, Amsterdam. Available at `http://adams.mgh.harvard.edu/pdf_repository/286_ROGERS.PDF`.

ROSSE C. & MEJINO J. (2003). A reference ontology for bioinformatics : The foundational model of anatomy. *Journal of Biomedical Informatics*.

ROUSSEY C., CALABRETTO S. & PINON J.-M. (2002). Le thésaurus sémantique : contribution à l'ingénierie des connaissances documentaires. In B. BACHIMONT, Coordinateur, *Actes des 6 ^{es} Journées Ingénierie des Connaissances (IC)*, p. 209–220, Rouen, France.

S. RUSSELL & P. NORVIG, Coordinateurs (2002). *Articial Intelligence : A Modern Approach.* Artificial Intelligence. Prentice Hall, 2^{nd} edition edition.

SATTLER U., CALVANESE D. & MOLITOR R. (2003). *The Description Logic Handbook : Theory, Implementation and Applications.*, chapitre Relationships with other formalisms., p. 142–183. Cambridge University Press.

SCHREIBER G., AKKERMANS H., ANJEWIERDEN A., DE HOOG R., SHADBOLT N., VAN DE VELDE W. & WIELINGA B. (2000). *The CommonKADS Methodology*. Knowledge Engineering and Management. MIT Press. ISBN : 0-2621-9300-0.

SÉGUÉLA P. (2001). *Construction de modèles de connaissances par analyse linguistique de relations lexicales dans les documents techniques*. Thèse de doctorat, Université Toulouse III. Disponible à `http://patrick.seguela.free.fr/these_seguela.html`.

SHIFFMAN R., MICHEL G. & ESSAIHI A. (2004). Bridging the guideline implementation gap : a systematic approach to document-centered guideline implementation. *J Am Med Inform Assoc*, **11**, 418–426.

SLODZIAN M. (1999). WordNet et EuroWordNet : questions impertinentes sur leur pertinence linguistique. *Sémiotiques*, (17), 51–70. Numéro spécial *Dépasser les sens iniques dans l'accès automatisé aux textes*, coordonné par B. Habert.

SLODZIAN M. (2000). Wordnet : what about its linguistic relevancy ? In R. DIENG, Coordinateur, *Proc. of the EKAW conference*, Juan-les-Pins, France.

SMITH B. (2004). Beyond concepts : Ontology as reality representation. In A. VARZI & L. VIEU, Coordinateurs, *International Conference on Formal Ontology and Information Systems (FOIS 2004)*, Turin. Available at `http://ontology.buffalo.edu/bfo/BeyondConcepts.pdf`.

SOUALMIA L. & DARMONI S. (2005). Combining different standards and different approaches for health information retrieval in a quality-controlled gateway. *International Journal of Medical Informatics (IJMI)*, p. 141–150.

SOWA J. (1984). *Conceptual Structures : Information Processing in Mind and Machine*. London : Addison-Wesley.

SOWA J. (2000). Ontology, metadata and semiotics. In *8^{th} International Conference on Conceptual Structures (ICCS'2000)*, volume 1867, p. 55–81 : Springer-Verlag LNCS.

SPACKMAN K. & CAMPBELL K. (1998). Compositional concept representation using snomed : Towards further convergence of clinical terminologies. In *Proc AMIA Symp*, p. 740–744.

SPACKMAN K. A., DIONNE R., MAYS E. & WEIS J. (2002). Role grouping as an extension to the description logic of ontylog motivated by concept modeling in snomed. In *Proc AMIA Symp*, p. 712–716.

STAAB S. & STUDER R. (2004). *Handbook on Ontologies*. Handbooks in Information Systems. Germany : Springer, 1^{st} edition edition.

STEICHEN O., BOZEC C. D.-L., JAULENT M.-C. & CHARLET J. (2007). Construction d'une ontologie pour la prise en charge des patients hypertendus en vue d'une analyse de l'individualisation des décisions. In C. O. BAGAYOKO, Coordinateur, *12^{e} Journées Francophones Informatique Médicale (JFIM)*, Bamako, Mali. À paraître.

STEVENSON M. (2002). Combining disambiguation techniques to enrich an ontology. In *Workshop on Machine Learning and Natural Language Processing for Ontology Engineering (ECAI)*, Lyon, France.

SUESSER P. (2000). Le codage des pathologies et les droits des citoyens. In *Development International Institute - Colloque enjeux et perspectives du codage des actes*. Disponible à `http://www.delis.sgdg.org/menu/sante/diicodpat.htm`.

SURE Y., ERDMANN M., ANGELE J., STAAB S., STUDER R. & WENKE D. (2002). Ontoedit : Collaborative ontology development for the semantic web. In *The first International Semantic Web Conference (ISWC)*. Available at `http://www.aifb.uni-karlsruhe.de/WBS/ysu/publications/2002_iswc_ontoedit.pdf`.

SZULMAN S. & BIÉBOW B. (2004). OWL et Terminae. In *14e journées francophones d'Ingénierie des Connaissances (IC)*.

SZULMAN S., BIÉBOW B. & AUSSENAC-GILLES N. (2002). Structuration de terminologies à l'aide d'outils d'analyse de textes avec terminae. *TAL*, **43**(1), 103–128.

TARHUNI M., MEYER R. & BAGAYOKO C. (2005). Mémoire d'ingénierie des connaissances. Master's thesis, Université Paris V.

THIRION B., BAUDIC F., DOUYERE M., LEROY J., PIOT J. & DARMONI S.-J. (1999). Cismef, catalogue et index des sites medicaux francophones : pourquoi, comment. In *Association Européenne pour l'Information et les Bibliothèques de Santé (EAHIL)*, number Newsletter 47. Disponible à `http://www.eahil.net/newsletter/47/CISMef.htm`.

TRONCY R. (2004). *Formalisation des connaissances documentaires et des connaissances conceptuelles à l'aide d'ontologies : application à la description de documents audiovisuels traitant du cyclisme*. Thèse de doctorat, Joseph Fourier - Grenoble 1. Disponible à `http://homepages.cwi.nl/~troncy/Publications/Troncy-PhD04.pdf`.

TRONCY R. & ISAAC A. (2002). DOE : une mise en œuvre d'une méthode de structuration différentielle pour les ontologies. In *13e journées francophones d'Ingénierie des Connaissances (IC)*, p. 63–74, Rouen.

TRONCY R., ISAAC A. & MALAISÉ V. (2003). Using xslt for interoperability : Doe and the travelling domain experiment. In *The 2nd International Workshop Evaluation of Ontology-based Tools (EON)*, volume 87, p. 92–102, Sanibel Island, Florida, United States : CEUR Proceedings. Available at `http://homepages.cwi.nl/~troncy/Publications/Troncy_Isaac_Malaise-eon03.pdf`.

USCHOLD M. & GRÜNINGER M. (1996). Ontologies : Principles, methods and applications. *Knowledge Engineering Review*, **11**(2).

USCHOLD M. & KING M. (1995). Towards a methodology for building ontologies. In *Workshop on Basic Ontological Issues in Knowledge Sharing, The International Joint Conference on Artificial Intelligence (IJCAI)*, Montreal, Canada. Available at `http://www.aiai.ed.ac.uk/project/pub/documents/1995/95-ont-ijcai95-ont-method.ps`.

J. VAN BEMMEL & M. MUSEN, Coordinateurs (1997). *Handbooks of Medical Informatics*. Berlin, Germany : Springer-Verlag.

VAN HEIJST G., SCHREIBER A. & WIELINGA B. (1997). Using explicit ontologies in kbs development. *International Journal of Human and Computer Studies - Knowledge Acquisition*, **46**(2/3), 183–292.

VOROS S., ORVAIN E., LONG J. & CINQUIN P. (2006). Automatic detection of instruments in laparoscopic images : a first step towards high level command of robotized endoscopic holders. In *International Conference on Biomedical Robotics and Biomechatronics*, Pisa, Italy.

WELTY C. A. & GUARINO N. (2001). Supporting ontological analysis of taxonomic relationships. *Data Knowledge Engineering*, **39**(1), 51–74. Available at http://www.cs.toronto.edu/~jm/2507S/Readings/Welty.pdf.

WÜSTER E. (1981). L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, l'informatique et la science des choses. In G. RONDEAU & H. FELBER, Coordinateurs, *Textes choisis de terminologie.*, volume Vol. I : Fondements théoriques de la terminologie., p. 55 – 113. Québec, Canada : Université de Laval - GIRSTERM.

ZIPF G. (1949). *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

ZWEIGENBAUM P. (1999). Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. *Innovation Stratégique en Information de Santé (ISIS)*, (2–3), 27–47. Disponible à http://www-test.biomath.jussieu.fr/~pz/Publications/ZweigenbaumISIS99/isis99.html.

ZWEIGENBAUM P., BACHIMONT B., BOUAUD J., CHARLET J. & BOISVIEUX J.-F. (1995). Issues in the structuring and acquisition of an ontology for medical language understanding. *Methods Inf Med*, **34**(1-2), 15–24.

ZWEIGENBAUM P., BAUD R., BURGUN A., NAMER F., ÉRIC JARROUSSE, GRABAR N., RUCH P., LE DUFF F., THIRION B. & DARMONI S. (2003). Umlf : construction d'un lexique médical francophone unifié. In *Journée Francophone d'Informatique Médicale (JFIM)*, Tunis. Disponible à http://www-test.biomath.jussieu.fr/~pz/FTPapiers/Zweigenbaum:JFIM2003.pdf.

ZWEIGENBAUM P., BOUAUD J., BACHIMONT B., CHARLET J., SÉROUSSI B. & BOISVIEUX J.-F. (1998). From text to knowledge : a unifying document-oriented view of analyzed medical language. *Methods Inf Med*, **37**(4-5), 384–393. Available at http://www.ncbi.nlm.nih.gov/htbin-post/Entrez/query?uid=9865036&form=6&db=m&Dopt=b.

ZWEIGENBAUM P., DARMONI S., GRABAR N., DOUYÈRE M. & BENICHOU J. (2002). An assessment of the visibility of MeSH-indexed medical web catalogs through search engines. *Journal of the American Medical Informatics Association*, **8**, 954–958.

Table des figures

ANNEXE A

Guide de bonnes pratiques méthodologiques

*Ce guide de bonnes pratiques est un résumé de notre méthodologie et de nos réflexions sur les questions d'évaluation, d'évolution et de maintenance des ressources terminologiques et ontologiques. Dans cette annexe, nous allons donc décrire successivement les problèmes que nous nous sommes posés et les solutions que nous avons trouvées. Certains sont des problèmes de recherches, d'autres de mise en forme. Notre but, ici est de donner une vue chronologique et synthétique de notre démarche. Le passage d'une étape méthodologique à une autre se décide en fonction de la qualité et de la quantité des résultats obtenus et de leur intérêt pour répondre au problème. Ce document ne peut être considéré comme définitif car le travail d'Ingénierie ontologique est complexe et demande de résoudre encore de nombreux problèmes (*cf. *Da*FOE*4*A*pp*) avant d'aboutir à une réelle ingénierie mais nous espérons, ici, proposer des indications utiles à un ingénieur des connaissances ayant des compétences en Traitement automatique de la langue et en Informatique.*

Étape 0 : réflexions préliminaires

Cette section présente brièvement quelques points sur lesquels il faut s'interroger *a priori*. Le développement d'une ontologie est un travail qui peut prendre beaucoup de temps aussi mieux vaut anticiper et éviter les écueils.

Définir les limites du domaine de connaissances

Avant même de collecter les ressources textuelles nécessaires à la modélisation des connaissances du domaine, il faut avoir, *a priori*, une idée précise de ce à quoi va servir l'ontologie. Une fois cette question réglée, il faut définir les limites du domaine de connaissances en se demandant

quelles sont les connaissances utiles, quel sera le niveau de langue adéquat, comment trouver les ressources pertinentes, ...

Quelques critères à respecter concernant le modèle

Nous avons tenté de respecter les critères présentés ci-dessous dans la construction de l'ontologie de la pneumologie.

- le modèle doit être structuré avec des principes rigoureux, explicites et propagés dans l'ensemble du modèle ;
- le modèle doit être adapté en fonction des besoins et, tout particulièrement, en fonction des besoins des applications informatiques cibles, il doit donc être cohérent ;
- le modèle doit s'attacher à représenter de manière explicite les connaissances ;
- le modèle construit doit être maintenable et extensible, c'est pourquoi il faut garder constamment à l'esprit, tout au long de l'élaboration, les principes de structuration que l'on décide d'appliquer ;
- le modèle doit être une ressource à part entière et être, en tant que tel, séparé des applications qui l'utilisent ;
- si possible, nous pensons qu'il faut utiliser des langages standardisés, largement diffusés et diffusables, pour exprimer ce modèle.

Le respect de ces critères facilite l'ajout de nouvelles connaissances et la maintenance du modèle.

Étape 1 : Matériel

Sélection des ressources textuelles

Dans un projet de construction d'ontologies à partir de textes, le corpus, son statut et sa collecte sont d'une importance primordiale à la fois comme source de connaissances pour construire le modèle et comme source de référence tout au long du processus d'élaboration. Cette élaboration du corpus est une tâche délicate qui nécessite du temps, aussi il paraît important de ne pas remettre en cause sa qualité et son adéquation une fois la collecte des ressources achevée. Le choix des éléments du corpus se fait donc en fonction de l'application visée par l'ontologie.

- Il faut bien identifier les différents genres de textes disponibles et nécessaires, et sélectionner ceux qui sont pertinents par rapport aux objectifs.
- Les textes constituant le corpus doivent être collectés avec l'aide de spécialistes du domaine à modéliser et les connaissances qu'ils contiennent doivent avoir fait l'objet d'un consensus à l'intérieur de cette communauté.
- Ils doivent ensuite être triés et organiser en groupes homogènes. Cependant, si le corpus doit conserver une certaine homogénéité dans le choix du genre textuel, il doit également conserver une part d'hétérogénéité dans la mesure où les divers textes qui le composent

doivent, si possible, provenir de sources variées. Il est ainsi plus représentatif. C'est le cas de nos deux corpus : l'un rassemble des comptes rendus d'hospitalisation provenant de six hôpitaux et l'autre est constitué d'un ensemble de documents correspondant à un livre de cours sur la pneumologie.

- Les textes sont ensuite analysés en connaissance de cause, en tenant compte de leur nature, de leur origine, ...
- Enfin, la taille du corpus varie d'un travail à l'autre, il semble difficile de donner des indications précises quant à un nombre de mots optimum. Cependant, la loi de G. K. Zipf, permet de déterminer statistiquement le nombre de « mots-clés » qu'il est intéressant d'étudier dans un corpus et représente un moyen, ou en tout cas une indication, pour évaluer la taille nécessaire du corpus (Zipf, 1949).

Outils

Nous utilisons le logiciel SYNTEX-UPERY comme outils d'analyse de traitement du langage. SYNTEX est un module d'analyse syntaxique fondée sur l'hypothèse que les mots qui ont un sens proche se caractérisent par des dépendances similaires (Bourigault & Lame, 2002). Ainsi, ce module permet d'obtenir des relations de dépendances syntaxiques entre mots ou syntagmes (noms *vs* syntagmes nominaux, verbes *vs* syntagmes verbaux et adjectifs *vs* syntagmes adjectivaux). A la fin du traitement nous obtenons un réseau de dépendances syntaxiques – ou réseau terminologique – dont les éléments sont les candidats termes qui vont nous servir pour construire l'ontologie. Le module UPERY met ensuite en œuvre le principe de l'analyse distributionnelle « à la Harris » (Harris, 1968) : il calcule des proximités distributionnelles entre les candidats termes du réseau sur la base des contextes syntaxiques partagés. Nous obtenons un réseau de candidats termes, leurs proximités contextuelles et leurs liens avec le corpus source. Les résultats de l'analyse sont visualisables dans TERMONTO, l'interface d'accès et de traitement des données du logiciel.

L'éditeur DOE [1] permet de construire notre ontologie selon la sémantique différentielle. Ce logiciel permet également de compléter l'ontologie en ajoutant à chaque concept sa traduction en anglais, ses synonymes ainsi qu'une définition encyclopédique. Il en va de même pour les relations.

L'éditeur PROTÉGÉ [2] intervient au moment de formaliser et d'opérationnaliser l'ontologie.

Étape 2 : construction du corpus de référence

Bien souvent les ressources amenées à constituer le corpus de référence sont disponibles dans des formats inexploitables par des outils de TAL. Nous pensons en particulier aux formats propriétaires comme Microsoft Word. Pour construire nos propres corpus [CRH] et [LIVRE], nous avons du enregistrer l'ensemble des documents Word au format « texte » avant de les

[1] The Differential Ontology Editor, `http://opales.ina.fr/public`

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

retravailler à l'aide de programmes Perl *ad hoc* pour les convertir au format XML. Nous avons ensuite adapté ce format XML aux besoins de SYNTEX-UPERY qui prend en entrée du pseudo-XML.

Nous avons également développé plusieurs programmes pour anonymiser les ressources constituant le corpus [CRH]. Dans un soucis de cohérence, nous avons choisi de garder les occurrences des noms propres et des dates en les remplaçant par ceux de notre cru.

Étape 3 : traitement du corpus

Le corpus [CRH] est ensuite traité par le logiciel SYNTEX-UPERY. Le résultat de l'analyse distributionnelle nous permet de construire les éléments de base - *i.e.* primitives - de l'ontologie. Le second corpus [LIVRE] est analysé par le biais de patrons lexico-syntaxiques prédéfinis qui permettent d'extraire des couples d'unités lexicales correspondant au motif de la relation sémantique recherchée (hypéronymie, synonymie...). Les résultats obtenus nous aident à contrôler et enrichir la hiérarchie de l'ontologie. Nous avons choisi les méthodes employées en fonction des genres textuels et des caractéristiques terminologiques de nos corpus : redondance, structure, richesse ...

À titre indicatif, le corpus [CRH] donne 36 881 syntagmes nominaux (SN) et le corpus [LIVRE] en donne 17 666. Sachant que le processus de sélection des candidats termes est manuel, il nous a donc fallu trouver un moyen pour faire le tri dans ces SN.

Étape 4 : sélection des candidats termes

Les résultats fournis sur la base du corpus [CRH] servent de support dans le choix de candidats termes (CT) représentatifs de la pneumologie en tant qu'activité médicale. Nous distinguons deux étapes dans leur sélection. Le travail est manuel et s'appuie sur les fonctionnalités de TERMONTO.

1) Nous parcourons l'ensemble des résultats fournis par l'analyse syntaxique et choisissons d'étudier, en premier lieu, les syntagmes nominaux (SN) dont la fréquence d'apparition en corpus est supérieure à 12 (2 % du corpus). Nous repérons les grands axes conceptuels typiques du corpus et donc du domaine représenté. A chaque CT, nous associons un critère de validité (ainsi nommé dans TERMONTO), compris dans un intervalle allant de 1 à 6, correspondant à l'un de ces axes : 1 (CT non pertinent appartenant à l'axe Autres), 2 (réservé aux CT déjà modellisés dans l'ontologie), 3 (CT appartenant à l'axe Symptômes), 4 (CT appartenant à l'axe Pathologies), 5 (CT appartenant à l'axe Signes) et 6 (CT appartenant à l'axe Traitements/Examens). Par exemple, nous fixons à 6 le critère de validité pour tous les CT de ce dernier axe - *e.g. examen*, *doppler*, *radiographie*, *etc*. Au début de la méthodologie, tous les CT ont un critère de validité égale à 1 et à la fin égale à 2 car ils sont, en principe, tous définis dans l'ontologie. Les critères de validité 3, 4, 5 et 6 sont utilisés temporairement durant la phase de construction. Ces regroupements permettent une première phase de travail sur les rapprochements des CT par contexte.

2) L'analyse distributionnelle rapproche deux à deux les termes partageant les mêmes contextes

(descendants en tête et en expansion). Comme cette analyse est symétrique, elle rapproche également les contextes en fonction des termes qu'ils partagent (voisins en tête et en expansion). Les descendants en tête donnent des informations sur ce qui pourrait être des concepts fils ou des concepts définis. Les descendants en expansion donnent des informations sur la place du concept dans la hiérarchie, sur le concept père. Les voisins en tête et en expansion nous permettent de constituer des regroupements de candidats termes sémantiquement proches du candidat terme étudié. Ces regroupements sont d'une grande aide pour élaborer la structure hiérarchique de l'ontologie, aussi bien l'axe horizontal que vertical.

Étape 5 : structuration de la hiérarchie ontologique

La méthodologie de construction d'ontologies différentielles utilisée est constructive, elle permet de placer de manière précise chaque concept dans la structure hiérarchique (Bachimont, 2000). Les regroupements des candidats termes sous de grands axes conceptuels nous permettent très rapidement de développer des micro-structures hiérarchiques. Donc, pour structurer l'ontologie, il convient d'articuler les candidats termes choisis dans la précédente étape en précisant les principes différentiels qui les définissent : le principe de communauté avec le concept père, le principe de communauté entre les concepts frères, le principe différentiel avec le concept père et le principe différentiel entre les concepts frères. À ce stade, la définition des principes différentiels est entièrement manuelle. Les candidats termes des 4 axes conceptuels (3, 4, 5 et 6) sont définis selon ces principes.

Nous essayons, autant que faire ce peut, de réfléchir à la place de chaque concept dans la hiérarchie et de respecter les choix conceptuels qui sont obligatoirement faits à ce moment là de la construction. L'intérêt de définir ces principes est d'obliger constamment l'ingénieur des connaissances à penser la place de chaque concept par rapport aux concepts environnants.

Les résultats de l'analyse par patrons lexico-syntaxiques sur le corpus [LIVRE] nous aident à définir les principes différentiels. Les patrons lexico-syntaxiques représentent des motifs de relations sémantiques spécifiques. Ils sont construits autour d'un marqueur, également appelé pivot, qui est l'indice d'une relation lexicale, comme le marqueur *entre autres* pour la relation d'hyperonymie. Ainsi, un patron de la forme *DET SN, entre autres SN* permet d'extraire l'unité lexicale *Les méningites, entre autres pathologies ...* et de mettre en relation d'hyperonymie *méningite* et *pathologie*. Cette méthode a été présentée dans (Hearst, 1992) et expérimentée dans plusieurs travaux, notamment dans (Caraballo, 1999). Les patrons lexico-syntaxiques liés à l'hyperonymie mettent en relation des couples père-fils potentiels intéressants pour contrôler et enrichir la structure hiérarchique de l'ontologie. Dans le cadre de la construction d'ontologies différentielles, nous appliquons cette méthode à la recherche d'énoncés définitoires en corpus. Les résultats nous aident à renseigner les principes différentiels et donc à vérifier la cohérence de notre modélisation. Ils nous permettent d'enrichir la hiérarchie de l'ontologie en prenant en compte les informations apportées par la comparaison des analyses terminologiques (la hiérarchie issue de l'analyse distributionnelle du corpus [CRH] et celle issue du repérage par patrons lexico-syntaxiques sur le corpus [LIVRE]) – complémentaires ou divergentes – issues des deux corpus. La comparaison de ces structures est détaillée dans (Baneyx *et al.*, 2005c). Enfin, les

résultats apportent des renseignements intéressants à intégrer à l'ontologie : synonymes, acronymes …. Les patrons que nous employons ont été développés par Malaisé *et al.* (2004). Le corpus [LIVRE], d'un genre pédagogique, à la particularité d'être très structuré et se révèle particulièrement propice à ce type de recherche. Par exemple, les patrons utilisant le marqueur *Il s'agit de* ont pu associer un titre de section (par exemple « Asthme ») à sa description dans la première ligne du texte (« Il s'agit d'une maladie inflammatoire des voies aériennes »). Les unités lexicales extraites sont validées manuellement et les hiérarchies créées sont visualisables sous DOE.

Étape 6 : réutilisation d'une top-ontologie

À ce stade intermédiaire, l'ontologie est bien ce qu'on appelle une ontologie de domaine et les concepts supérieurs sont soit des concepts manifestement du domaine – *e.g. Signe, RoleEnMilieuHospitalier, ModeAdministration, ReflexionMedicale, ConditionExamen, …* – soit des concepts plus généraux, ayant une signification ou des instantiations particulières dans le domaine médical – *e.g. ObjetIndenombrable, EtatObjetPhysique, ObjetNaturel, ….* Toujours à ce stade, les relations, compléments des concepts, ont été définies, sans que leur champ d'application – l'arité en logique de description – n'ait été précisée.

Une ontologie de haut niveau, appelée *top-Ontologie*, vocable que nous conserverons par la suite, est alors nécessaire pour organiser les concepts médicaux les uns par rapport aux autres. Une conséquence indispensable et intéressante de cette démarche est de définir les relations entre les concepts à ce niveau-là (par exemple [abstract-object]-(has-view-point)->[meta-abstract-object]), pour que leur applicabilité – leur *range* – se propage aux niveaux inférieurs. Nous avons choisi de réutiliser la top-ontologie MENELAS. Que ce soit pour les concepts ou les relations, tout n'est pas utile à notre modélisation de la pneumologie car le système MENELAS (*cf.* section 2.10) tentait de prendre en compte une modélisation plus complexe que notre travail. Nous avons donc « coupé » certaines branches de la hiérarchie. D'après notre expérience, il est possible d'adapter ou de reprendre des morceaux de top-ontologies développées pour d'autres projets. Il faut, par contre, qu'elle soit du même domaine et construite selon des principes philosophiques proches.

Étape 7 : formalisation

D'après la méthode ARCHONTE, l'ontologie que nous avons obtenu à ce stade est une ontologie référentielle rassemblant des concepts et des termes du domaine (le « terme préféré » en français et en anglais, les synonymes), des connaissances de type encyclopédique sur le domaine, ainsi que des relations sémantiques bien connues comme *est_un*, *est_une_partie_de*, *est_caracterise_par*, *est_pratique_par* … Cette ontologie peut être traduite en HTML grâce à un export prévu à cet effet dans la nouvelle version de DOE.

L'étape de formalisation est faite à l'aide du logiciel PROTÉGÉ qui offre notamment la possibilité de représenter graphiquement l'ontologie. Concrètement, nous avons utilisé l'export OWL de DOE, compatible avec le format d'import OWL de PROTÉGÉ (Troncy *et al.*, 2003). Cependant,

cette compatibilité est pour l'instant limitée. Les seules informations conservées d'un logiciel à l'autre sont la définition en langage naturel des principes différentiels et le label du concept. Aussi, nous avons développé un certain nombre de programmes annexes qui permettent de récupérer et de réinjecter les informations manquantes.

L'étape de formalisation permet d'introduire des axiomes logiques qui définissent le comportement des individus qui constituent les extensions des concepts formels[3]. Nous rajoutons des définitions formelles aux concepts en précisant les individus des concepts telle que la liste des pneumologues (Dr Dupont, Dr Mozart ...) pour le concept *Pneumologue*, la liste des êtres humains pour le concept *EtreHumain* ... La liste des pneumologues recoupe celle des êtres humains. Ainsi, la formalisation permet de créer un nouveau concept formel *PersonnePneumologue* défini à l'intersection de ses deux concepts formels pères *EtreHumain* et *Pneumologue*. Comme l'explique B. Bachimont (2000), les concepts formels primitifs ont une sémantique référentielle déterminée par les engagements sémantiques et ontologique tandis que les concepts formels définis sont uniquement déterminés en fonction d'un engagement ontologique. L'ajout de nouveaux concepts à ce stade de développement modifie la structure hiérarchique car on passe d'une arborescence fondée sur des relations de similarités et de différences à une arborescence fondée sur une logique d'inclusion ensembliste. Dans l'ontologie formelle, les concepts qui n'entretiennent pas de relation de subsomption (couple père-fils) s'excluent mutuellement. Dans l'ontologie référentielle, les concepts peuvent admettre des extensions qui ont un sous-ensemble commun. L'héritage multiple est donc possible et la structure construite n'est plus un arbre mais un treillis.

Étape 8 : opérationalisation

Il reste à opérationaliser OntoPneumo. Il s'agit de traduire l'ontologie obtenue à l'étape précédente en une ontologie destinée à servir dans un système informatique. Nous ne parlons pas ici de l'éditeur grâce auquel l'ontologie est formalisée et opérationalisée mais bien de l'outil dans lequel elle va servir. Pour cela, elle doit être spécifiée dans un langage de représentation des connaissances doté de capacité d'inférence.

En sortie de PROTÉGÉ, nous avons logiquement choisi d'utiliser le langage OWL qui répond parfaitement à nos besoins en termes d'expressivité et de maniabilité. L'utilisation de ce standard préconisé par le W3C favorisera la réutilisation de l'ontologie dans d'autres projets et/ou sous d'autres environnements de développement. Cela devrait également simplifier la tâche de maintenance de l'ontologie, notamment en permettant à d'autres ingénieurs des connaissances de modifier les fichiers. Ce format de sortie ne préjuge pas du format d'utilisation de l'application qui peut être encore OWL ou autre chose. Dans notre cas, nous avons fait la démarche d'enregistrer l'ontologie dans un base de données relationnelle en raison d'un certain nombre de critères de pérennité et de calculabilité. Mais il faut savoir que cela pose des problèmes de représentation et d'expression des connaissances (en particulier au niveau des inférences) qui nécessitent encore des recherches pour aboutir à une application totalement opérationnelle (*cf.* DAFOE4APP).

[3]Nous conseillons au lecteur intéressé le tutoriel d'A. Rector disponible à l'url suivante : `http://www.cs.man.ac.uk/~rector/modules/cds/OWL\%20Biomedical\%20Ontology\%20Tutorial-v1.pdf`

Étape 9 : validation et évaluation

La validation de l'ontologie doit être un processus itératif sur lequel intervient l'expert du domaine. Nous avons validé notre ontologie au cours d'entretiens avec un médecin pneumologue. La confrontation de nos points de vue sur l'organisation des connaissances du domaine et leur modélisation informatique s'est révélée être très enrichissante. Un regard extérieur aide particulièrement à cette étape. Il est également souhaitable de vérifier la cohérence de la structure ontologique en utilisant des raisonneurs comme RACER. Nous insistons sur la rigueur des principes de modélisation à appliquer tout au long du processus, de la construction de l'ontologie à sa maintenance. Ce sont les garants d'une modélisation valide.

Mais, ne l'oublions pas, une ontologie est avant tout un outil. C'est donc en situation qu'il convient d'évaluer sa qualité et son intérêt. Cette étape sert à juger de la valeur ajoutée (compréhension, utilisabilité...) et de la qualité de l'ontologie du point de vue de l'utilisateur Il devient nécessaire de penser l'évaluation de l'ontologie en testant ses performances sur des tâches spécifiques.

La problématique qui reste encore à approfondir est la possibilité d'expliciter des critères qui permettraient d'affirmer la bonne adéquation de l'ontologie à l'application. Ces critères peuvent être à chercher du côté de la formalisation ou avec des visées plus empiriques.

Une question reste en suspend : s'il paraît nécessaire de valider et d'évaluer l'ontologie, qui doit le faire ? La réponse qui vient en premier est l'utilisateur. Mais l'expérience tend à prouver que les utilisateurs finaux des systèmes à base ontologique, s'ils sont capables de donner leur avis sur tel ou tel concept, sur une partie de l'arborescence, n'ont que très rarement les compétences requises pour avoir une vision d'ensemble et juger la totalité de la construction et de son schéma (Navigli & Velardi, 2004). Il est donc nécessaire de prévoir également une évaluation au niveau des concepteurs (Porzel & Malaka, 2004).

Étape 10 : évolution et maintenance

Dans le cadre de notre application, nous avons évalué les résultats que nous obtenons. Nous distinguons deux phases d'évaluation. Entre ces deux phases, nous avons analysé les erreurs et imprécisions de notre outil et avons fait évoluer l'ontologie pour améliorer sensiblement nos résultats. La plupart des problèmes étaient dus à l'absence, au niveau de l'ontologie, de certains termes dont notre outil avait besoin. Nous avons donc enrichi l'ontologie en créant de nouveaux concepts et en augmentant le nombre des synonymes en rapport avec les termes préférés. Le respect systématique des principes qui guident la construction de l'ontologie et déterminent la position de chaque concept dans la hiérarchie, permet de savoir précisément comment placer un nouveau concept.

Pour l'instant l'évolution de l'ontologie se fait essentiellement par l'ajout manuel de nouveaux concepts. Mais notre méthode, et particulièrement la structuration de la hiérarchie à l'aide des axes différentiels, ainsi que les principes rigoureux que nous nous sommes donnés, facilitent grandement l'évolution et la maintenance du modèle et permettent d'envisager des interfaces pour supporter cette évolution.

ANNEXE B

Extraits d'OntoPneumo

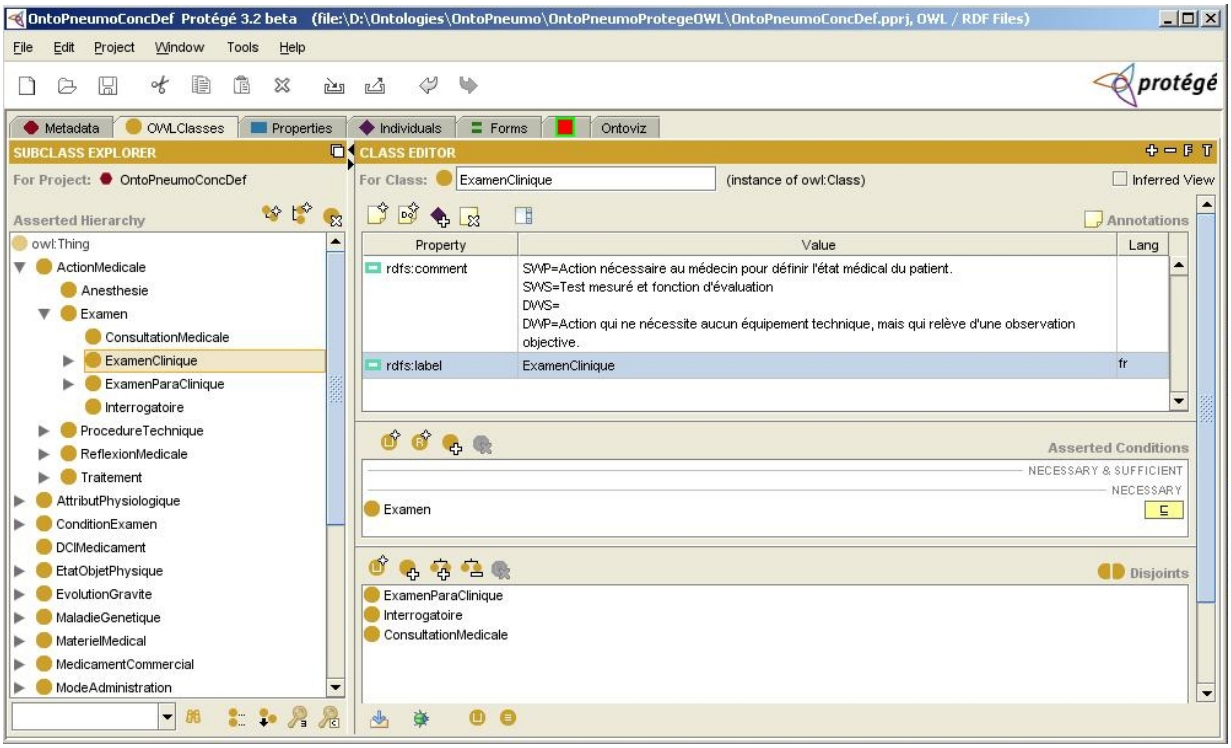


FIG. B.1 – Concept *ExamenClinique* vu avec PROTÉGÉ 3.2.

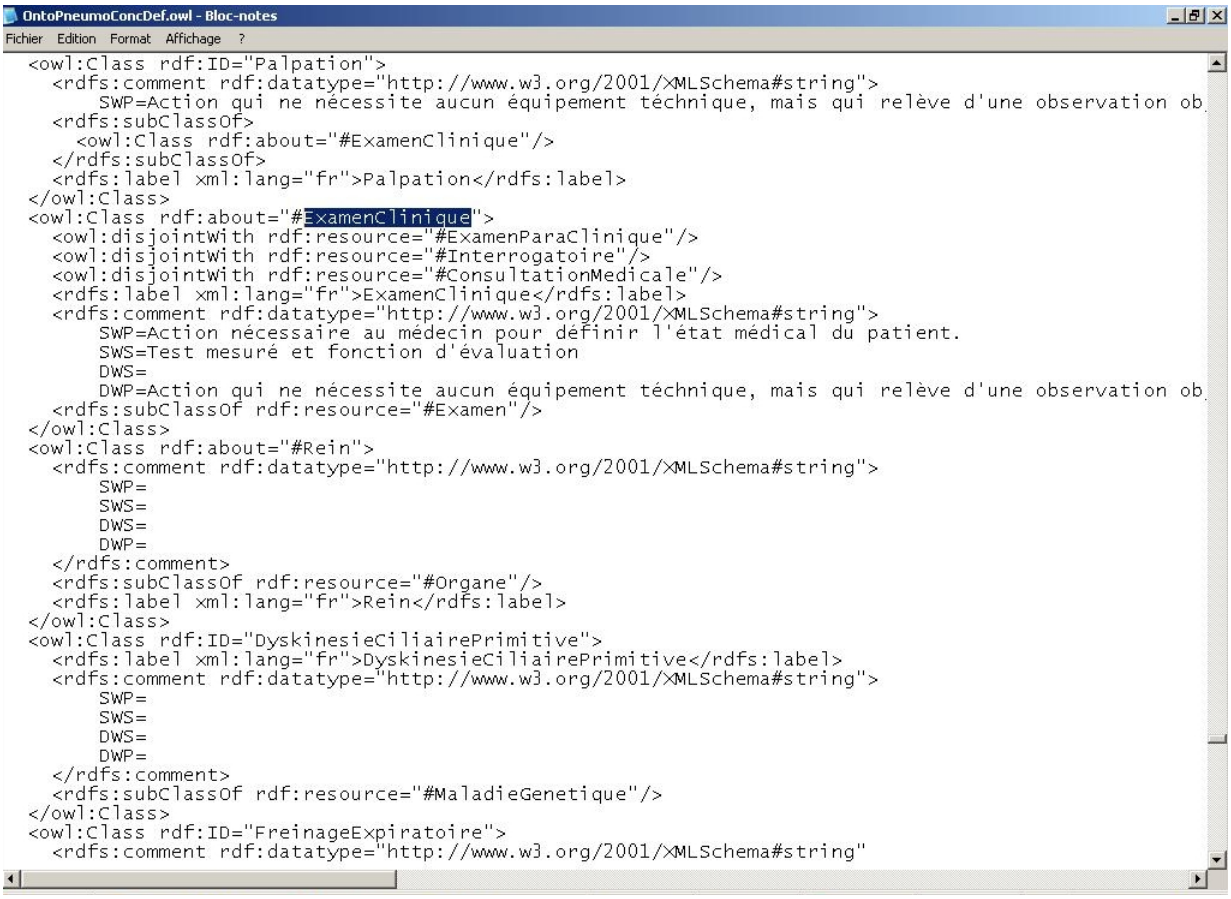


FIG. B.2 – Extrait du fichier OWL d'OntoPneumo centré sur la classe *ExamenClinique*.

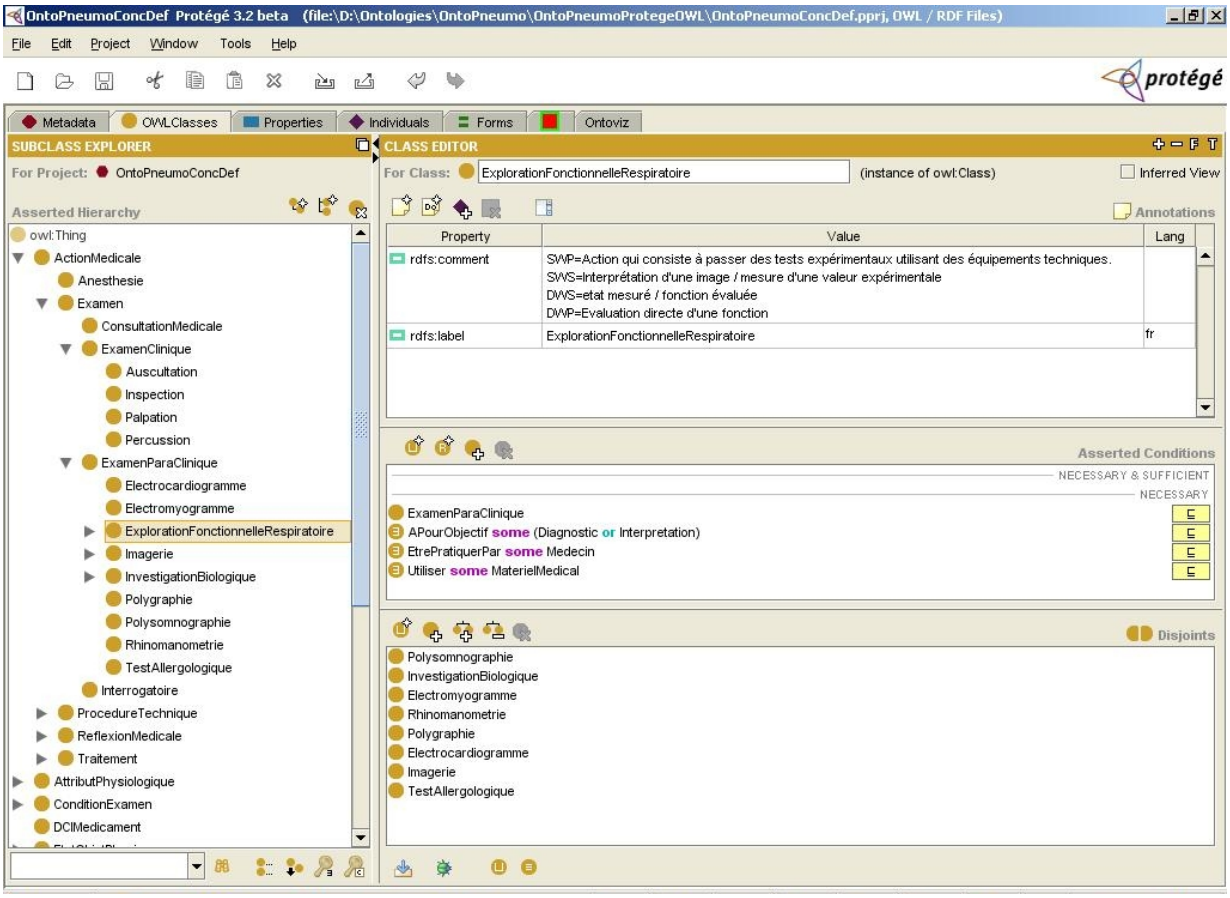


FIG. B.3 – Hiérarchie des examens paracliniques centrée sur le concept *ExplorationFonctionnelleRespiratoire* vue avec PROTÉGÉ 3.2.

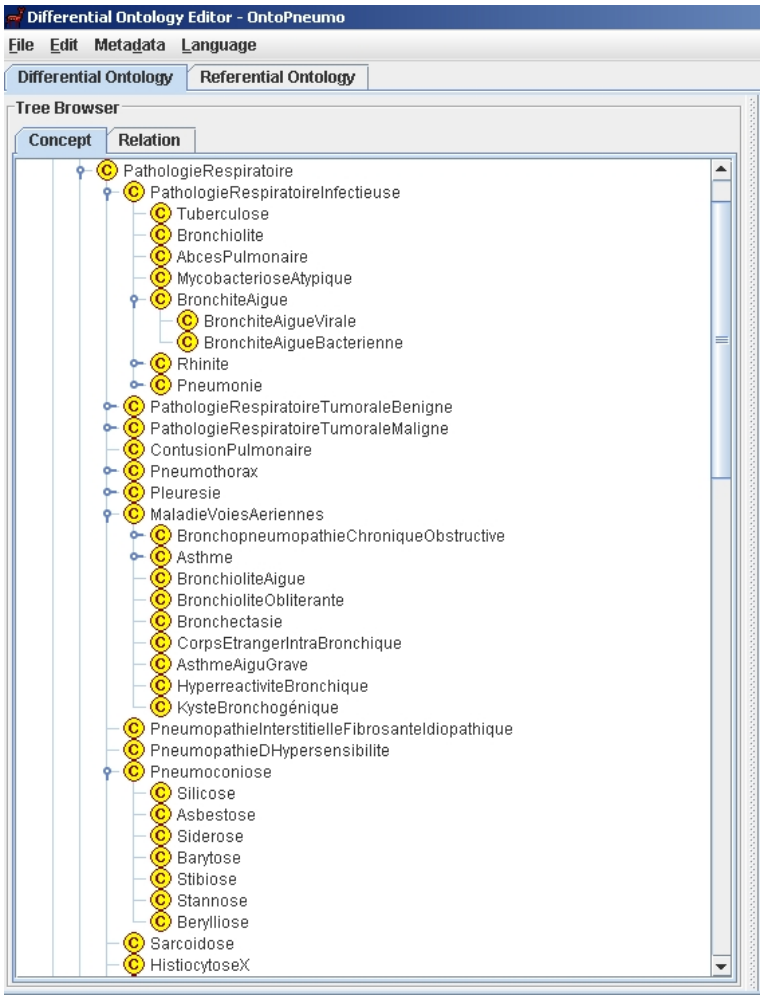


FIG. B.4 – Hiérarchie des pathologies respiratoires vue avec DOE.

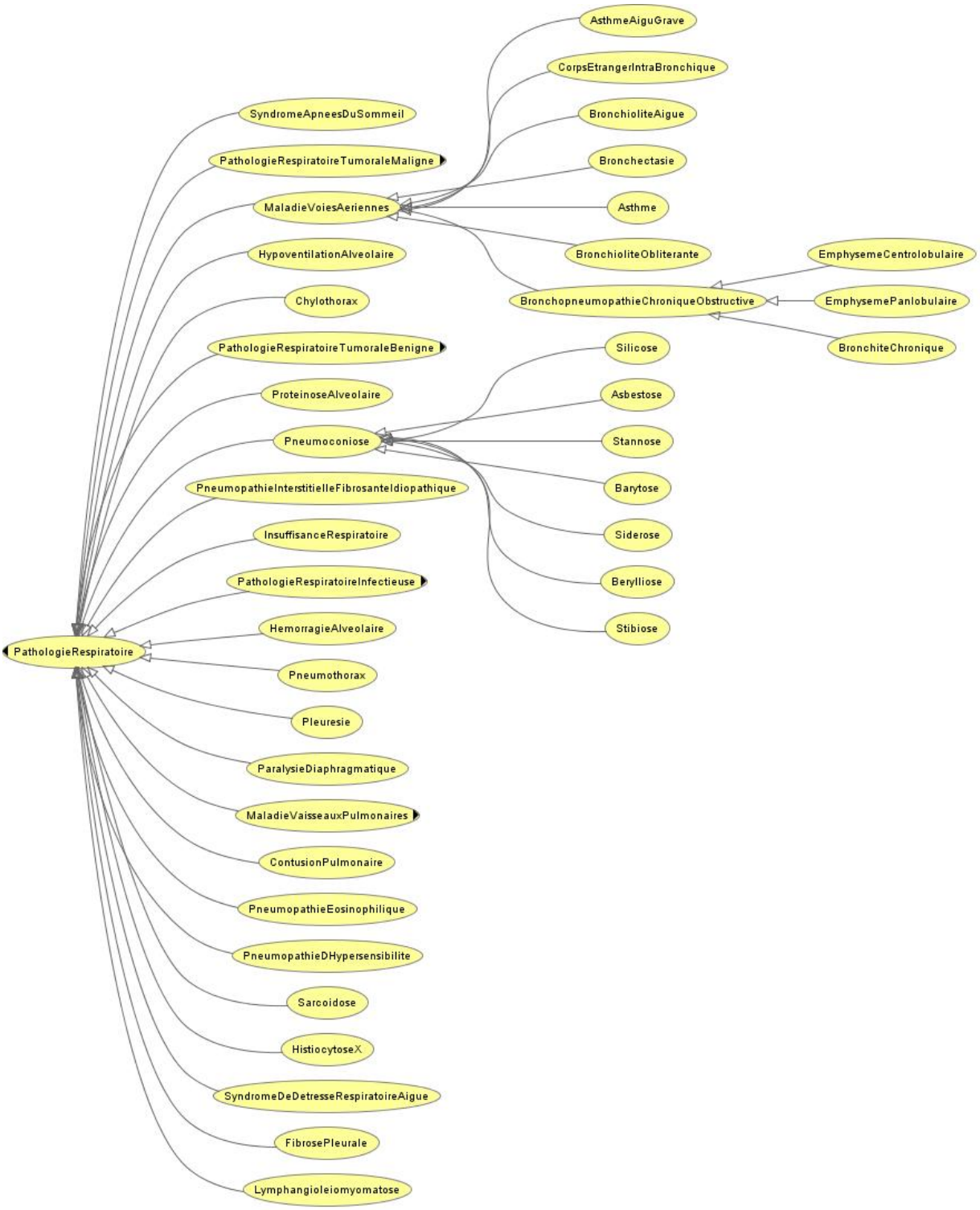


FIG. B.5 – Représentation graphique de la hiérarchie des pathologies avec le pluggin OWLViz de PROTÉGÉ 3.2.

ANNEXE C

Patrons lexico-syntaxiques

Marqueur	patron
Une sorte de	sorte P DET
Par exemple	exemple P[1] par ET P[2] MOT ET S[2]MOT
Etre un	être NS[1] pas ET P[1] « ce » ET P[2] MOT ET S[1-3] (« des »|« les »|« un »|« une »|« le »|« la »)
A savoir	savoir P[1] « à » ET NP[2] CatMS[PCT]
(	(P[1] CatMS[NC] ET S[1] CatMS[NC] ET NS[1] CHIFFRE ET S[2])
C'est-à-dire	dire P[1] à ET P[2] être
C'est-à-dire	(cad|c.a.d.|c'est-à-dire)
C'est-à-dire	d P[1]a ET P[2]d
Au sens de	sens P[1] « au »
En d'autres termes	terme P[1] autre
Soit	soit P[1] CatMS[PCT] ET NP soit
Enfin	enfin P[1] CatMS[PCT]
Il s'agit de	agir P[1] se ET NS[3] CatMS[VB]
Entendre par	entendre S[1-6] par
Par ... entendre	par S[1-6] entendre
Vouloir dire	vouloir S[1] dire
Indiquer	indiquer NS[1] (pas|par|sur)
Comme	comme P définir
Comme	comme S définir
Comme	comme P[1] CatMS[NC] ET S[1] CatMS[DET]1,3
Dire à l'indicatif ou au participe passé	dire[CatMS :(VIND|VPARP)]
Ou	ou P[1] ,
Autrement dit	dire P[1] autrement ET P[2] MOT
Même chose que	même S[1] chose ET S[2] que
De même que	même P[1] de ET S[1] que
Equivaloir à	équivaloir

FIG. C.1 – Patrons lexico-syntaxiques de recherche d'énoncés définitoires 1/2, (Malaisé, 2005).

Employer pour	employer S[1] pour
Action de	action S[1] de ET S[2] CatMS[VINF]
Préciser le sens	préciser S sens
(appeler | nommer | référer | dénommer | désigner | dénoter | signifier | définir)	(appeler | nommer | référer | dénommer | désigner | dénoter | signifier | définir) NP[1] se ET NS[1-4] « à » ET NS[1] .
(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception) NP[1] « au » ET P[1] CatMS[DET] ET S porter NS[1-6] sur
(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception) P(prendre | recevoir | appliquer | employer | réserver | utiliser | donner | renvoyer | référer | définir)
(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	(nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception) S (prendre | recevoir | appliquer | employer | réserver | utiliser | donner | renvoyer | référer | définir)
Etre (le | ce) (nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	être S (le | ce) ET S[2-3] (nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)
Sous le (nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)	sous S[2] (nom | terme | mot | expression |vocable | appellation | désignation | dénomination | concept | notion | acception)

FIG. C.2 – Patrons lexico-syntaxiques de recherche d'énoncés définitoires 2/2, (Malaisé, 2005).

Index

RÉSUMÉ

Depuis une vingtaine d'années, l'accès aux connaissances médicales est un enjeu majeur pour les professions de santé comme pour le grand public. Les limites actuelles des outils de traitement de l'information ne proviennent pas de leurs performances pour stocker et traiter rapidement des gros volumes, mais de leur incapacité à prendre en compte les spécificités des vocabulaires métier des utilisateurs. Le développement de ressources terminologiques et ontologiques pour faciliter l'usage des terminologies nationales et internationales, disponibles notamment dans le domaine de la médecine, revêt par conséquent une importance particulière. Il faut également souligner la pertinence de telles recherches dans la mouvance des Sciences et Technologies de l'Information et de la Communication, dans le cadre de la société de l'information et dans le contexte du Web sémantique.

Dans ce contexte, notre réflexion a porté sur la collecte, l'organisation, la représentation et la formalisation des connaissances en médecine, tout particulièrement, dans le domaine de la pneumologie. Nous avons été amenés à considérer le problème dans son ensemble, afin de comprendre les mécanismes qui sous-tendent la constitution de ressources terminologiques et ontologiques à partir de textes. Nous avons également considéré chaque tâche séparément, afin de proposer, pour chaque étape, si ce n'est une solution, au moins un savoir-faire personnel susceptible d'apporter des éléments de réponse. L'objectif principal de cette thèse consiste à mettre au point une ontologie dans le domaine de la pneumologie pour faciliter, d'une part, l'aide au codage médico-économique des pathologies et, d'autre part, la représentation des connaissances pertinentes relatives au patient, dans ce domaine de spécialité. Nos recherches couvrent l'ensemble du cycle de vie d'une ontologie, de la mise au point d'une méthodologie de construction à partir de textes à son utilisation dans un système opérationnel. Nous contribuons aux recherches dans les domaines de l'Ingénierie des connaissances et de l'Informatique médicale.

La méthode de travail adoptée est une démarche expérimentale ascendante qui consiste à partir des problématiques concrètes rencontrées pour aller vers la résolution des questions scientifiques sous-jacentes. Selon cette démarche, nous avons tout d'abord cerné les besoins des pneumologues en termes de représentation des connaissances. Ensuite, nous avons mis au point une méthodologie, destinée à l'ingénieur des connaissances, fondée sur la méthode ARCHONTE définie par B. Bachimont. L'enchaînement des processus d'extraction, de sélection et de choix des candidates termes du domaine ainsi que l'aide fournie par les patrons lexico-syntaxiques pour renseigner les principes différentiels la rendent relativement facile d'emploi (ou moins difficile qu'une autre) pour un ingénieur des connaissances. L'ontologie construite compte à ce jour 2 260 concepts primitifs. Enfin, nous avons développé un outil de codage semi-automatique proposant deux types de codages : (1) un codage médical qui représente graphiquement les informations pertinentes relatives aux pathologies du patient et qui, à terme, servira de descripteur pour indexer intelligemment les comptes rendus d'hospitalisation ; (2) un codage médico-économique pour lequel nous obtenons un rappel de 80% et une précision de 87%. Nos résultats concernant l'ontologie et l'outil nous encouragent à poursuivre nos recherches et à améliorer les solutions proposées.

Mots-clés : ontologie, thésaurus, aide au codage, sémantique différentielle, Ingénierie des connaissances, Informatique médicale, Pneumologie.

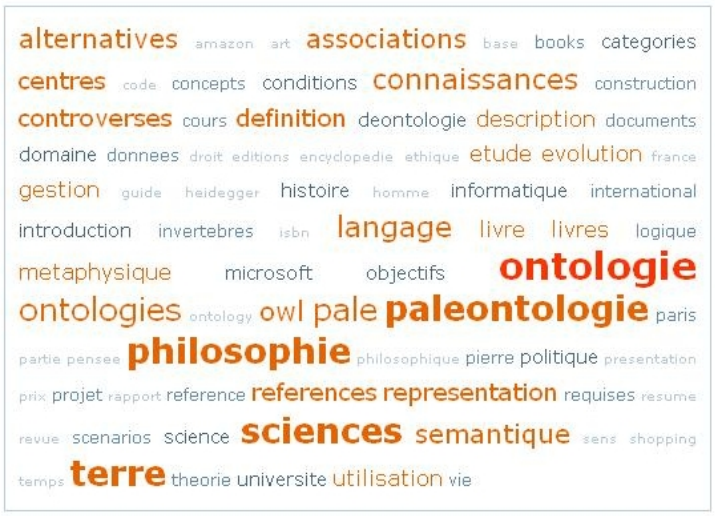


FIG. C.3 – Résultat du Nébuloscope pour le mot « Ontologie ».

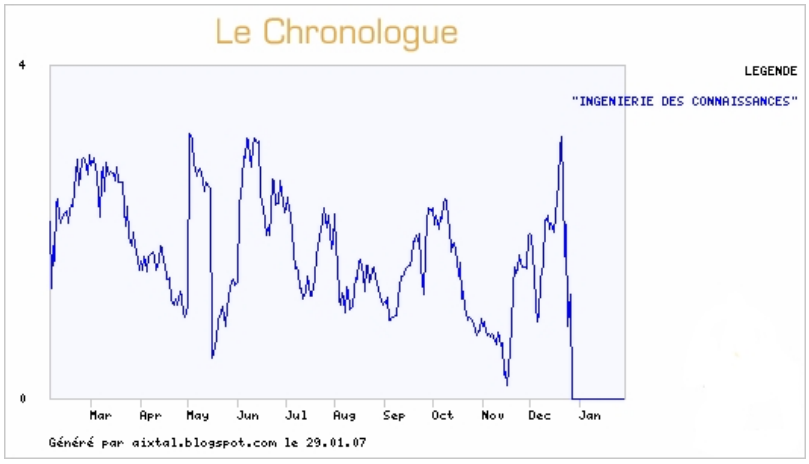


FIG. C.4 – Résultat du Chronologue pour l'expression « Ingénierie des connaissances ».

Le Nébuloscope et le Chronologue sont des outils développés par Jean Véronis[1].

[1] http://aixtal.blogspot.com/