

---

## Utiliser et construire des ontologies en médecine : le primat de la terminologie

Jean Charlet<sup>\*,\*\*</sup> — Audrey Baneyx<sup>\*</sup> — Olivier Steichen<sup>\*,\*\*\*</sup> — Iulian Alecu<sup>\*</sup> — Christel Daniel-Le Bozec<sup>\*,\*\*\*</sup> — Cédric Bousquet<sup>\*</sup> — Marie-Christine Jaulent<sup>\*</sup>

<sup>\*</sup> Centre des Cordeliers, INSERM, U872, Éq. 20, Paris, F-75006 France ; Université Pierre et Marie Curie, Paris ; Université Paris Descartes, Paris.

<sup>\*\*</sup> AP-HP, DSI/STIM, Paris, F-75004 France. Jean.Charlet@spim.jussieu.fr

<sup>\*\*\*</sup> AP-HP, HEGP, Paris, F-75015 France.

---

**RÉSUMÉ.** *Le Web sémantique réclame par définition de construire des ontologies permettant, par leur caractère formel, d'automatiser un certain nombre de tâches liées principalement à la recherche d'information et à la classification. Dans cet article, nous montrons que la médecine a des spécificités qui contraignent l'usage des ontologies et leur construction. Ces spécificités amènent à prendre en compte le caractère linguistique des thésaurus médicaux et à les utiliser en coordination avec les ontologies pour exploiter les informations médicales. En complément, la construction des ontologies peut prendre en compte l'expression linguistique des connaissances en utilisant les textes générés par les professionnels de santé durant leur activité. Nous donnons des exemples avec des applications dans trois spécialités médicales différentes.*

**ABSTRACT.** *Semantic Web requires to build ontologies allowing, by their formal character, to automate certain number of tasks connected mainly to information retrieval and to classification. In this paper, we want to show that the medicine has specificities which constrain the building of ontologies and their use. These specificities bring to take into account the linguistic character of numerous medical thesaurus and to use them in coordination with ontologies to exploit the medical information. Moreover, the building of the ontologies can take into account the linguistic expression of knowledge by using texts generated by the health professionals during their activity. We give examples with medical applications in three different medical specialties.*

**MOTS-CLÉS :** *ontologie, ontologie d'interface, ontologie de référence, thésaurus de spécialité, ingénierie des connaissances, traitement automatique des langues.*

**KEYWORDS:** *Ontology, Interface Ontology, Reference Ontology, Medical Specialty Thesaurus, Knowledge engineering, Natural Language Processing.*

---

## 1. Introduction

La médecine a été un des terrains d'expérimentation privilégiés de l'intelligence artificielle (IA) avec la construction de systèmes experts puis de systèmes à base de connaissances (SBC). Elle a été aussi le terrain des échecs de ces approches. Echecs qu'on peut analyser en s'intéressant à une mauvaise prise en compte de l'interaction homme-machine – *i.e.* répartition des rôles, ergonomie, décision en situation mal informée... – et du milieu organisationnel dans lequel les SBC devaient se développer.

Une autre raison des échecs, est l'accent mis par l'IA sur le raisonnement dans les années 1970-80 sans avoir totalement pris conscience de l'importance des connaissances du domaine. Par la suite, l'IA *via* l'ingénierie des connaissances et l'informatique médicale s'est spécifiquement intéressée aux ontologies qui semblaient pouvoir résoudre un certain nombre de problèmes amonts de la représentation des connaissances. Cet intérêt a été d'autant plus fort en médecine que celle-ci a une forte tradition de nommage des objets médicaux. En effet, les connaissances médicales se structurent depuis le XVII<sup>e</sup> siècle autour de l'anatomie. Ainsi, les médecins nomment les structures anatomiques et tous les concepts qui font la médecine. Ils contruisent pour cela de nombreux thésaurus.

Depuis une dizaine d'années, ces réflexions se sont insérées dans le contexte de renouveau sur la représentation des connaissances et, plus largement, sur la sémantique des objets du web, à savoir le web sémantique. La médecine est alors un terrain d'expérimentation des technologies, langages et standardisations du web sémantique. Terrain d'autant plus intéressant qu'elle conserve ses spécificités en termes de rapports entre l'homme et la machine et de création de thésaurus et classifications en tous genres. Dans ce contexte, on voit se dessiner 2 types de projets. Les premiers mettent l'accent sur la réalisation d'outils logiciels utilisant des représentations munies de sémantique formelle et des mécanismes inférentiels puissants. Les seconds reposent aussi sur une sémantique forte mais la combinent avec des ressources linguistiques pour mettre l'accent sur une meilleure adaptation aux fonctionnements cognitifs de l'utilisateur en n'automatisant que certains raisonnements.

Dans cet article, nous nous plaçons résolument dans le second type de proposition et nous voudrions montrer que le développement des ontologies médicales dans le contexte du web sémantique est l'occasion de penser des applications médicales qui ont pour caractéristiques d'utiliser des ontologies sans nier l'existence des terminologies médicales et qui les utilisent pour leur capacité de représentation (cf. section 2) et de raisonnement (cf. section 4). Notre position ici est :

- de prendre acte que le web sémantique, par la normalisation de ses langages de représentation des connaissances, en particulier sur les ontologies, réactive les recherches sur les ontologies, les thésaurus et leur utilisation conjointe ;
- d'argumenter que l'accent doit être mis sur ce type d'applications qui prennent en compte la dimension linguistique irréductible de la médecine, science ou activité pratique qui s'exprime principalement en langue. En particulier, nous devons, dans la plupart des sous-domaines de spécialité de la médecine, constater que les connais-

sances sont exprimées avec des termes répertoriés dans des thésaurus ou classifications médicaux ;

- de montrer que, dans ce contexte, le recours aux textes générés par les professionnels de santé durant leur activité est un moyen privilégié de construire les ontologies nécessaires, même si ce n'est pas le seul.

Pour cela, nous allons présenter trois applications caractéristiques de ce point de vue et développées à l'INSERM UMR\_S 872, éq20. Ces applications sont développées sur des cas réels et ont déjà apporté un certain nombre de résultats significatifs. La première application décrite, le projet OntoPneumo vise à construire une ontologie de la pneumologie pour l'aide au codage médical. La seconde application, est un projet de recherche sur les déterminants du raisonnement médical qui a abouti à la construction d'une première ontologie et a déjà eu pour effet de proposer une mise à jour des formulaires d'entrées de données cliniques dans le domaine de l'hypertension artérielle. La troisième application est une application de pharmaco-vigilance fondée sur une ontologie construite spécifiquement pour augmenter les signaux de détections de causalité possibles entre un médicament et un effet indésirable.

A la suite de ces présentations, nous discutons des perspectives ouvertes par ses applications d'un point de vue méthodologique en mettant l'accent sur trois points importants de notre point de vue : 1) le primat de la terminologie sur le raisonnement, c'est-à-dire le fait que les systèmes à base de connaissances construits s'attaquent à la question des rapports entre formalismes et expressions textuelles des terminologies, prioritairement à l'exécution d'inférences complexes ; 2) la question des méthodologies de construction des ontologies et les conséquences sur les différents niveaux d'abstraction de ces mêmes ontologies et 3) la nécessité qu'il y a à créer des ponts entre les différentes ontologies et terminologies du domaine dans une démarche de *serveur de terminologie*. Nous étendons cette dernière réflexion au principal outil utilisé par les chercheurs dans ce contexte, à savoir UMLS, et à la SNOMED, ontologie de référence dont le statut et l'usage doivent être discutés.

Dans la section 2, nous présentons le projet OntoPneumo. Dans la section 3, nous présentons le projet sur l'hypertension artérielle. Dans la section 4, nous présentons l'application sur la pharmaco-vigilance. Nous discutons les modes de construction et d'utilisation des ontologies dans ces applications en section 5 et nous concluons en section 6.

## 2. OntoPneumo

Le projet OntoPneumo, décrit ci-après, montre comment le web sémantique médical peut (et devrait) prendre en compte le caractère majoritairement documentaire et linguistique de la médecine. La construction de l'ontologie OntoPneumo utilise fortement les ressources terminologiques du domaine à modéliser pour rendre compte, le plus précisément possible, non seulement des pratiques médicales actuelles en pneumologie mais également des vocabulaires utilisés par les médecins.

## 2.1. Contexte

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 des établissements, le gouvernement souhaite disposer d'informations quantifiées et standardisées. Ainsi, le codage des dossiers patients, qui correspond au repérage des diagnostics et des actes médicaux survenus et pratiqués durant un épisode de soins, est devenu une obligation légale des structures de soins. La procédure de codification couramment appelée codage PMSI<sup>1</sup>, est le plus souvent réalisée manuellement par les praticiens qui s'aident d'un thésaurus de spécialité (Bensadoun, 2001). Les thésaurus proposés par les sociétés savantes et appelés *thésaurus de spécialité* sont construits en associant, quand c'est possible, un code PMSI à un terme du thésaurus pour permettre aux médecins de coder à partir de leur terminologie usuelle<sup>2</sup>. Mais il est aujourd'hui manifeste que les outils 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 d'utilisation, et le maintien de la complétude ainsi que de la cohérence du thésaurus est impossible. On constate que le sens des libellés de ces thésaurus médicaux repose sur les facultés d'interprétation du lecteur humain et qu'ils ne sont donc pas adaptés à une exploitation par l'ordinateur. Enfin, le codage PMSI étant établi pour des motivations médico-économiques, il est inadapté à une indexation purement médical du dossier médical. Dans ce contexte, il nous semble indispensable de décrire la sémantique et l'organisation des objets du domaine médical afin de se doter de modèles conceptuels (non contextuels et non ambigus) dont le sens est inscrit dans la structure même du modèle. Une telle modélisation est appelée ontologie (Staab *et al.*, 2004; Aussenac-Gilles *et al.*, 2000; Charlet, 2002; Rector, 1998).

## 2.2. Objectifs et hypothèse

L'objectif de nos recherches actuelles est de développer un environnement de codage de l'information médicale dépassant le seul cadre médico-économique du PMSI.

---

1. Le Programme de médicalisation du système d'information (PMSI), vise à introduire des concepts de comptabilité analytique dans la gestion administrative des hôpitaux : les diagnostics et actes effectués dans un établissement de santé sont codés et comptabilisés, rapportés à un patient et aux différents coûts de la structure. Cela permet ainsi de bâtir des indices de coûts relatifs par groupe homogène de malades. Le PMSI utilise un système de codage international, la CIM-10, pour les diagnostics, et un système français, développé grâce à une approche ontologique anglo-saxonne, la CCAM, pour les actes. Le PMSI évolue maintenant vers une comptabilité qui vise à analyser le coût de chaque acte : c'est la tarification à l'activité ou T2A.  
 2. *Thésaurus et classifications* sont tous deux des modes d'organisation plus ou moins hiérarchiques de termes mais n'expriment pas cette hiérarchie de la même façon et ne permettent pas le même type d'accès. Ce n'est pas l'objet de cet article. Pour plus de précision, voir par exemple <http://www.uhb.fr/urfist/Supports/Indexation/PlanCoursIndexation.html>. Dans la suite, nous utiliserons le terme thésaurus pour ces deux organisations.

Notre hypothèse de recherche est qu'une telle exigence est atteignable en développant un environnement de codage à base d'ontologies et s'appuyant sur l'analyse de travaux existants. En effet, une fois construite et acceptée par une communauté particulière (ici celle des pneumologues), une ontologie doit traduire un consensus explicite et un certain niveau de partage au sein de la communauté, deux aspects essentiels pour permettre l'exploitation des ressources construites par différentes applications ou agents logiciels. Nous ne souhaitons pas complètement automatiser la procédure de codage. Notre idée est plutôt de proposer un système de représentation des connaissances avec lequel le pneumologue 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 appelons cette représentation des connaissances médicales « codage médical ». A partir d'un tel codage, nous pourrions par exemple proposer au pneumologue de 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. Nous proposons également le codage PMSI, plus classique en médecine mais obligatoire, prenant en compte d'autres critères que les seuls aspects médicaux.

Pour représenter les données propres au patient nous modélisons le thésaurus de spécialité de la pneumologie sous la forme de concepts primitifs et/ou définis dans l'ontologie. Ainsi, nous sommes à même d'identifier et de coder les pathologies listées dans les comptes rendus d'hospitalisation. 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. Mais c'est parce qu'il est réducteur que ce codage permet au médecin une représentation résumée de ce qu'il sait sur le patient et, par là, un rappel rapide. Nous devons donc proposer des outils qui permettent le codage en même temps qu'ils permettent aux médecins d'assumer le caractère réducteur de ce processus. De plus, il faut aussi qu'ils puissent interagir avec le système formel dans les termes de leur domaine de spécialité, répertoriés dans les thésaurus de spécialité.

## **2.3. Matériel et méthode**

### **2.3.1. Matériel : outils et corpus**

Nous utilisons le logiciel SYNTAX-UPERY comme outils d'analyse et de traitement (automatique) du langage (TAL). Un outil comme SYNTAX se nourrit d'un corpus de texte qui doit couvrir avec le plus d'exhaustivité possible l'activité étudiée, ici la pneumologie. Dans ce but, nous avons collecté des comptes rendus d'hospitalisation (corpus intitulé [CRH]) dans six hôpitaux de l'Assistance Publique-Hôpitaux de Paris. Au total, nous disposons de 1 038 CRH. Ce premier corpus [CRH] compte environ 417 000 mots. Sachant qu'il a été établi dans (Le Moigno *et al.*, 2002b) que 350 000 mots est un minimum pour obtenir de bons résultats avec nos outils, le corpus [CRH]

semble être une bonne base d'expérimentation. Le second corpus, intitulé [LIVRE], est construit d'après un ouvrage pédagogique et correspond environ à 823 000 mots.

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 *et al.*, 2002). 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 (cf. figure 1).

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

Figure 1. TERMONTO : interface d'accès aux données de SYNTEX-UPERY

L'éditeur DOE<sup>3</sup> permet de construire notre ontologie selon la sémantique différentielle. L'ontologie est exportée en OWL<sup>4</sup>, un langage de représentation des connaissances préconisé par le consortium W3C<sup>5</sup>. Ce format garantit l'interopérabilité de notre modèle et rend l'ontologie disponible pour l'éditeur PROTÉGÉ.

3. The Differential Ontology Editor, <http://homepages.cwi.nl/~troncy/DOE/>

4. Web Ontology Language, <http://www.w3.org/TR/owl-features/>

5. <http://www.w3.org/>

### 2.3.2. Méthode

La méthodologie mise en œuvre permet de décrire les variations des sens des termes considérés en contexte. C'est pourquoi cette méthode 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. Nous distinguons quatre étapes : 1) la constitution du corpus des connaissances et son analyse par des outils de TAL, 2) la normalisation sémantique des termes du domaine grâce à la mise en œuvre des principes différentiels, 3) l'engagement ontologique qui permet de formaliser les concepts, 4) l'opérationnalisation de l'ontologie dans un langage de représentation des connaissances interprétable par l'ordinateur (Bachimont *et al.*, 2002). Cette méthode est décrite en détails dans (Baneyx, 2007). Les deux premières étapes ayant été abordées dans la section 2.3.1, nous nous intéressons ensuite aux étapes trois et quatre.

#### Formalisation de l'ontologie avec les logiques de description

Les étapes de formalisation et d'opérationnalisation sont faites à l'aide de l'éditeur d'ontologies PROTÉGÉ 3.1<sup>6</sup> qui permet de formaliser l'ontologie en utilisant les logiques de description. L'étape de formalisation permet d'introduire des axiomes logiques qui définissent le comportement des individus qui constituent les extensions des concepts formels. Ainsi, nous formalisons précisément chaque concept de la hiérarchie de l'ontologie et définissons les rapports qu'entretiennent les concepts les uns avec les autres par le biais des relations. Par exemple, l'ingénieur des connaissances décide des relations qui peuvent lier le concept de *Medecin* et celui d'*Asthme* : *Diagnostiquer*, *observer*. Cette étape de formalisation permet, par exemple, de créer un nouveau concept formel *PersonnePneumologue* défini qui hérite de son concept père *EtreHumain* et se spécialise avec la relation *A-Une-Specialite* dont le domaine de valeur est *SpecialiteMedicale* qui prend, ici, la valeur *Pneumologie*.

La liste des *PersonnePneumologue* recoupe celle des *EtreHumain*. 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. La structure de l'ontologie n'est plus un arbre mais un treillis.

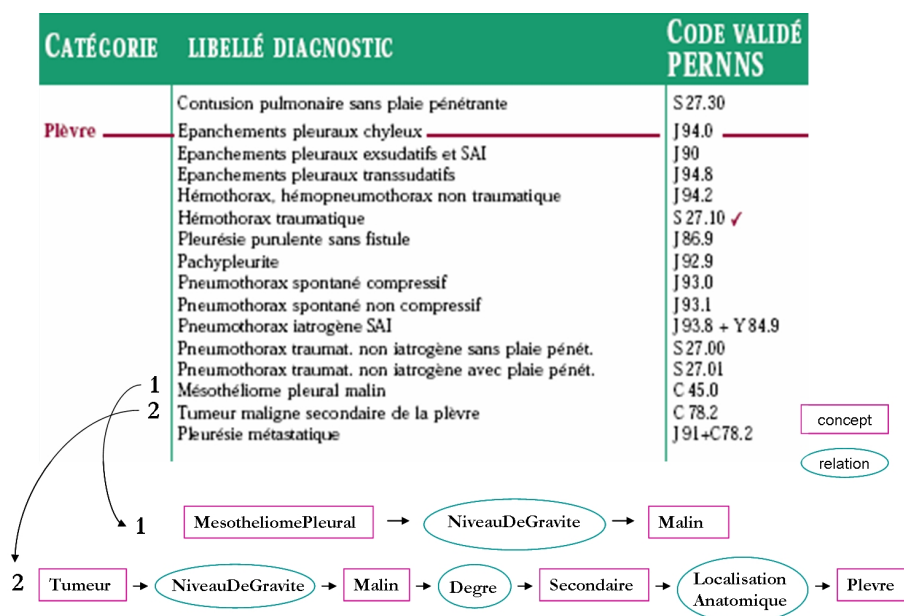
## 2.4. Résultats

L'ontologie du domaine compte à ce stade 1 460 concepts primitifs auxquels viennent s'ajouter les 400 concepts primitifs de la topontologie et de la *core*-ontologie (cf. section 5.2) du projet MENELAS<sup>7</sup> et les concepts définis à l'aide du logiciel PROTÉGÉ. OntoPneumo compte au total 1 860 concepts primitifs.

6. <http://protege.stanford.edu/>

7. <http://estime.spim.jussieu.fr/Menelas/Ontologie/html/>

L'enrichissement manuel de l'ontologie effectué à l'aide des 337 expressions contenues dans le thésaurus de spécialité garantit que presque toutes les expressions du thésaurus sont modélisées (cf. figure 2). En effet, chacune de ces expressions a un lien (de synonymie, de variances lexicales) avec les concepts de l'ontologie. Par exemple, la partie du thésaurus consacrée aux broncho-pneumopathies obstructives chroniques (groupe des maladies qui implique une limitation du passage de l'air dans les poumons) contient 17 expressions, telle que la bronchite chronique simple. Le niveau de couverture de ce chapitre est de 92 % et implique la modélisation de 38 concepts et 16 relations. Les 8 % restant sont dus à des problèmes d'imprécision. Par exemple, nous pouvons représenter la bronchite chronique simple mais pas la surinfection de la bronchite chronique parce qu'OntoPneumo contient le concept *infection* mais pas le concept *surinfection*.

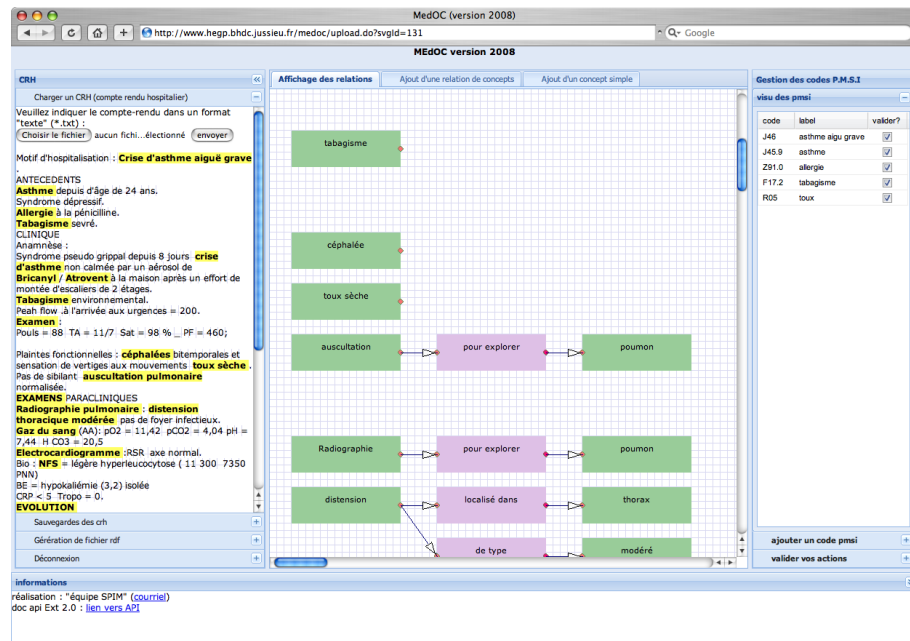


**Figure 2.** Modélisation de deux expressions du thésaurus de spécialité

OntoPneumo est utilisée au cœur d'un outil de codage médical semi-automatique proposant deux types de codages : a) le codage médical représente graphiquement les informations relatives aux pathologies du patient et, à terme, ces informations serviront de descripteur pour indexer intelligemment les comptes rendus d'hospitalisation. Nous voyons ce codage comme une alternative au problème posé par la standardisation du langage médical ; b) le codage PMSI repose au médecin pneumologue une liste de codes médico-économiques générée en fonction des pathologies pertinentes



identifiées. Nous obtenons, pour l'instant, un rappel de 80 %<sup>8</sup> et une précision de 87 %<sup>9</sup>. La figure 3 est une copie d'écran de l'interface de l'outil et en montre les principales fonctionnalités : 1) le chargement du compte rendu et l'identification des informations pertinentes, 2) la représentation graphique du codage médical, 3) les codes médico-économiques proposés et 4) les fonctionnalités d'interaction entre l'utilisateur et l'interface de codage médical.



**Figure 3.** Interface de l'outil de codage. A gauche, le compte rendu analysé avec surlignage des mots repérés comme renvoyant à des concepts. Au milieu, affichage des concepts et de leurs relations. A droite, proposition de codes PMSI

Nous considérons que notre travail a abouti à un démonstrateur valide. Il nous reste à compléter et valider totalement l'ontologie et à la mettre en œuvre au sein d'une application industrielle de codage (médical et médico-économique) qui sera évaluée dans plusieurs unités de soins de pneumologie. Ceci sera fait dans le cadre du projet MedOC qui fait partie du projet DaFOE4App<sup>10</sup>.

8. Le rappel est le rapport du nombre de réponses pertinentes trouvées au nombre total de réponses pertinentes. Un rappel de 100 % signifie que toutes les réponses pertinentes ont été trouvées.

9. La précision étant le rapport du nombre de réponses pertinentes trouvées au nombre total de réponses sélectionnées par l'outil. Une précision de 100 % que toutes les réponses trouvées sont pertinentes.

10. <http://www.dafoe4app.fr>.

### 3. Hypertension artérielle

#### 3.1. Objectifs

Les guides de bonne pratique synthétisent les modalités de prise en charge les mieux étayées sur une question médicale. Ils sont souvent employés comme référence pour l'évaluation des pratiques médicales. Cette utilisation est problématique car les résultats de la recherche clinique ne sont jamais les seuls déterminants des décisions médicales. Les caractéristiques des patients, les spécificités de leurs problèmes et les circonstances particulières de prise en charge doivent également être intégrées dans la décision (Haynes *et al.*, 2002). Ces particularités justifient parfois d'agir en désaccord avec les recommandations des guides de bonne pratique.

Même individualisées à l'encontre des recommandations existantes, les décisions doivent rester justifiées. Leur évaluation fait alors appel à une analyse détaillée et comparative des cas individuels (Green *et al.*, 1998; Friedman *et al.*, 2006). Au vu des caractéristiques particulières d'un patient et d'une situation clinique, il doit être possible de comprendre les décisions non conformes aux recommandations.

Notre objectif clinique, dans le contexte d'une unité hospitalière spécialisée en hypertension artérielle, est d'étudier les déterminants et la logique de l'individualisation des pratiques, en particulier quand elle ne se conforment pas aux recommandations des guides de bonne pratique.

#### 3.2. Hypothèses

Analyser de façon fine la logique d'individualisation des décisions médicales n'est possible à grande échelle qu'avec l'aide d'outils informatiques. Ces derniers peuvent notamment regrouper les patients similaires pour faciliter l'analyse comparative des décisions prises à leur sujet. A cette fin, les cas pris en charge dans le service doivent être représentés *a)* de manière formelle pour leur traitement informatique et *b)* avec suffisamment de détail pour fournir des résultats fins.

Les ontologies sont des systèmes terminologiques adaptés à nos besoins, car elles supportent des représentations sémantiques à la fois formelles et détaillées. Dans la mesure où chaque ontologie possède une couverture, une granularité et une structure qui lui est propre, elle ne peut habituellement pas être réutilisée pour une autre tâche que celle prévue initialement (Coiera, 1995). Notre objectif actuel est donc de construire une ontologie consacrée à la prise en charge de l'hypertension, en vue d'assister l'analyse qualitative des décisions médicales dans un service spécialisé.

La modélisation ontologique se fait à la fois à partir de corpus représentant la production textuelle des praticiens durant leur activité et à partir de conceptualisations préexistantes du domaine. Le choix de ces ressources vise à épuiser la gamme des justifications possibles pour les décisions médicales dans la prise en charge spécialisée de l'hypertension.

### 3.3. Matériel et méthodes

#### 3.3.1. Matériel

Un formulaire informatisé est utilisé depuis 30 ans dans l'unité d'hypertension de l'hôpital européen Georges Pompidou pour recueillir les observations médicales. Environ 5 000 observations semi-structurées sont saisies chaque année. Le modèle d'information clinique initial a été conceptualisé par des experts du domaine. Il a évolué en fonction du progrès des connaissances médicales et du retour d'expérience sur son utilisation (Degoulet *et al.*, 1990).

Le formulaire compte actuellement 176 questions, pour la plupart optionnelles (cf. figure 4). Elles sont à réponses booléennes (diabète ? « oui »), temporelles (date du dernier infarctus du myocarde ? « 21/12/1996 »), prédéfinies dans une liste à choix unique ou multiple (récapitulatif cardiaque ? avec une liste de choix comportant, entre autres, « hypertrophie ventriculaire gauche » et « valvulopathie »), numériques (pression artérielle diastolique en millimètres de mercure ? « 95 ») ou textuelles libres (conclusion ? « prévoir des explorations hormonales en raison d'une hypokaliémie persistant à l'arrêt des diurétiques »).

The screenshot shows a medical observation form for a patient named DENISE (71 years old). The 'COEUR' section includes questions about recent heart failure, chest pain, and heart failure symptoms, with radio button options for 'aucune anomalie', 'souffle diastolique', 'souffle systolique', or 'autres'. The 'ECG' section includes a date field for the most recent ECG and a multi-select list for ECG anomalies. The 'ECHOGRAPHIE' section includes a date field for the echocardiogram and a multi-select list for echocardiogram anomalies. The 'Récapitulatif cardiaque' section includes a multi-select list for various cardiac conditions. Red boxes and labels highlight specific data entry types: 'Choix simple' for a radio button selection, 'Date' for a date field, 'Binaire' for a yes/no selection, 'Valeur numérique' for a numerical input field, 'Texte libre' for a text area, and 'Choix multiples' for a multi-select list.

Figure 4. Onglet « coeur - ECG » de l'observation médicale

La liste des items cliniques trouvés dans le formulaire (en-têtes de sections et de sous-sections, questions, réponses prédéfinies) a été extraite de la base de données

hospitalière à l'aide d'une requête SQL. Les réponses en texte libre des 5 109 observations saisies en 2005 ont été anonymisées : les noms des patients et des médecins ont été remplacés par des identifiants numériques et les dates de naissance par l'âge en années. Leur collection forme un corpus de 350 000 mots en langue française.

Huit guides de bonne pratique, publiés entre 1999 et 2005, ont été colligés pour constituer un corpus de 56 000 mots en langue anglaise, susceptible d'être analysé à l'aide d'outils de traitement automatique des langues.

### 3.3.2. *Méthode*

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

Le formulaire d'observation et les guides de bonne pratique sont des conceptualisations théoriques de la prise en charge de l'hypertension artérielle, élaborées par des experts informés de l'état des connaissances médicales dans leur domaine. Les deux premières étapes de modélisation s'appuient sur ces ressources qui fournissent les concepts-clés intervenant dans la prise en charge réglée des hypertendus et donnent des indices pour structurer l'ontologie.

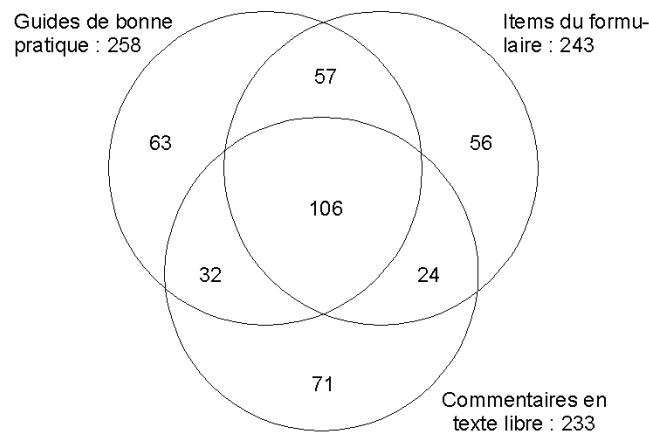
Les commentaires en texte libre saisis dans les observations donnent accès aux problèmes concrets et imprévus rencontrés par les cliniciens dans leur pratique quotidienne. Ils alimentent la dernière étape de notre démarche de modélisation, destinée à enrichir le noyau ontologique précédent avec les concepts utilisés pour la description et la justification des décisions individualisées.

## 3.4. *Résultats*

### 3.4.1. *Etat d'avancement*

Les concepts provenant des items du formulaire et des guides de bonne pratique ont été extraits dans leur intégralité. En ce qui concerne le corpus des commentaires en texte libre, seuls les concepts rencontrés plus de 50 fois dans ont été retenus dans un premier temps. Ce seuil correspond à environ 1 % des questionnaires. Il permet d'isoler les concepts récurrents, qui ont une forte probabilité d'être liés à des aspects génériques de la prise en charge des patients hypertendus vus en milieu spécialisé, à l'instar des concepts issus des items du formulaire et des guides de bonne pratique. Ainsi, nous disposons des concepts fondamentaux utilisés pour la prise en charge réglée de l'hypertension en milieu spécialisé. Chacune des sources de concepts a contribué à part égale à la constitution de ce noyau ontologique (cf. figure 5). Au-delà de la modélisation ontologique, les concepts identifiés dans les guides de bonnes pratiques et les concepts récurrents dans les commentaires en texte libre ont été estimés suffi-

samment pertinents par les cliniciens pour servir de point de départ à une démarche d'actualisation du formulaire d'observation clinique.



**Figure 5.** Nombre de concepts issus de chaque ressource

#### 3.4.2. Concepts issus des items du formulaire d'observation

Les items du formulaire ont directement fourni 243 concepts cliniques utilisés pour la prise en charge des patients hypertendus. Au-delà des concepts cliniques provenant des questions et des réponses prédéfinies, les titres des sections et des sous-sections de l'observation correspondent à des concepts structurants, comme celui d'histoire familiale, et révèlent comment les cliniciens se figurent leur domaine. Ces concepts structurants ont été notés comme indices pour la structuration ontologique.

#### 3.4.3. Concepts issus des guides de bonne pratiques

L'analyse des guides de bonne pratique avec les outils SYNTAX et UPÉRY a révélé 258 concepts cliniques, dont 163 déjà identifiés dans les items du formulaire. Deux exemples de concepts additionnels parmi les 95 trouvés dans les guides de bonne pratique sont celui d'apnées du sommeil comme cause de l'hypertension et celui de démence comme conséquence de l'hypertension. L'importance de ces deux notions est reconnue depuis peu et elles n'ont pas encore été intégrées dans le formulaire d'observation. L'analyse terminologique des guides de bonne pratique a également révélé des concepts structurants, comme celui d'atteinte des organes cibles. L'analyse distributionnelle a fait apparaître quelques groupes sémantiques intéressants. Nombreuses caractéristiques de l'hypertension ont ainsi été rapprochées pour s'organiser selon différents axes : hypertension permanente, épisodique, blouse-blanche ou masquée (selon les moments où l'hypertension est présente) ; hypertension globale ou systolique isolée (selon la ou les composante(s) élevée(s) de la pression artérielle) ; hypertension artérielle légère, modérée ou sévère (selon le niveau de pression artérielle) ; etc.

#### 3.4.4. *Concepts issus des commentaires en texte libre*

L'analyse des commentaires en texte libre, avec les mêmes outils, a identifié 233 concepts ayant plus de 50 occurrences dans le corpus, dont 71 sont complètement originaux (trouvés ni dans les items du questionnaire ni dans les guides de bonne pratique). La plupart d'entre eux s'intègrent dans des protocoles de prise en charge spécialisée, adaptés au recrutement de l'unité d'hypertension (forte prévalence des hypertensions compliquées et/ou avec une cause spécifique). Le concept de dissociation rénine-aldostérone, par exemple, compte 80 occurrences dans les commentaires en texte libre et constitue un déterminant important pour la prise en charge de certains patients, selon des règles de décision ou des habitudes locales. Ces concepts ne relèvent donc pas de l'individualisation des pratiques mais d'un autre niveau de prise en charge réglée, au-delà des recommandations trouvées dans les guides de bonne pratique.

Néanmoins, parmi ces concepts fréquemment rencontrés dans les commentaires en texte libre, on en trouve déjà qui se rapportent aux particularités de certains cas réclamant une individualisation des décisions. Ainsi, la notion de refus du patient (d'une hospitalisation, d'un traitement, du sevrage tabagique, de l'inclusion dans un protocole, etc.) n'est pas explicitée dans les guides de bonne pratique et ne relève d'aucune règle de prise en charge. Cependant, elle apparaît régulièrement dans les commentaires en texte libre et influence significativement les décisions médicales.

Enfin, ce corpus de commentaires en texte libre indique lui aussi des principes d'organisation conceptuelle, cette fois-ci plus par l'analyse distributionnelle que par la présence de concepts structurants. Par exemple, de nombreux concepts liés aux décisions thérapeutiques médicamenteuses ont été regroupés : ajuster ou simplifier le traitement, débiter un nouveau médicament, continuer ou interrompre un ancien médicament, augmenter ou diminuer une posologie, etc.

### 3.5. *Perspectives*

Nous organisons actuellement le noyau ontologique en une monohiérarchie strictement taxinomique, dans le respect des principes de la sémantique différentielle. Nous commençons par structurer les concepts provenant des items de l'observation et des guides de bonne pratique, en nous inspirant des indices obtenus lors de l'analyse des ressources terminologiques (concepts structurants). Nous ajouterons ensuite les concepts récurrents dans les commentaires en texte libre, en réorganisant la structure si nécessaire. Une fois ce noyau ontologique consolidé, nous poursuivrons l'analyse terminologique du corpus de commentaires en texte libre, de manière à intégrer dans l'ontologie les concepts moins fréquents, qui reflètent les pratiques individualisées.

Une fois l'ontologie construite, nous l'utiliserons pour représenter les cas cliniques pris en charge dans l'unité d'hypertension. La représentation des cas avec l'ontologie, c'est-à-dire l'association à chaque observation des concepts qui lui correspondent, n'est complètement automatisable que pour la partie structurée de l'observation. En ef-

fet, les items du formulaire d'observation sont liés de façon univoque avec les concepts qu'ils ont révélés. Par opposition, le contenu des commentaires en texte libre ne peut pas être automatiquement représenté avec les concepts qui en sont issus.

Une fois réalisée cette représentation automatique mais partielle des cas, il sera possible de regrouper les cas similaires grâce à une mesure de similitude sémantique. L'analyse qualitative des cas similaires conduira à identifier des caractéristiques singularisantes importantes ayant justifié une prise en charge différente. L'ontologie et la représentation des cas seront alors complétées manuellement avec ces nouveaux concepts. A l'issue de cette première boucle, un nouveau calcul de similitude sémantique entre les représentations enrichies pourra définir des groupes de cas plus proches afin de raffiner l'analyse qualitative. Le processus pourra être répété autant de fois que nécessaire, jusqu'à ce que la représentation de chaque cas comporte l'ensemble des déterminants de sa prise en charge, que ce soit sur un mode générique ou sur un mode individualisé.

## 4. Pharmacovigilance

### 4.1. Objectifs

#### 4.1.1. *Présentation rapide de la pharmacovigilance*

La pharmacovigilance a pour objet la détection, l'évaluation et la prévention des effets indésirables liés aux médicaments (EIM). Il s'agit d'un véritable enjeu de santé publique car les EIM causent aujourd'hui plus de 3 % des hospitalisations en France et sont une des dix principales causes de décès aux Etats-Unis. Ces effets indésirables sont signalés par les professionnels de santé qui envoient les déclarations spontanées aux centres régionaux de pharmacovigilance ou à l'unité de pharmacovigilance du laboratoire pharmaceutique qui commercialise le produit suspect. L'OMS (Organisation Mondiale de la Santé) définit un signal en pharmacovigilance comme toute information sur une relation causale possible entre un médicament et un effet indésirable, la relation étant préalablement inconnue ou mal documentée (Edwards *et al.*, 1994).

#### 4.1.2. *La détection du signal dépend fortement d'une classification terminologique*

La détection d'un signal nécessite le codage reproductible et fiable des effets indésirables ; codage sur lequel peuvent ensuite se reposer des regroupements permettant de signaler des EIM de même type. Les terminologies WHO-ART (*World Health Organization – Adverse Reaction Terminology*) et MedDRA (*Medical Dictionary for Drug Regulatory Activities*) sont utilisées en pharmacovigilance pour le codage et pour l'analyse statistique des données (Brown, 2002). Des conditions cliniques similaires peuvent être décrites au moyen de termes différents. Par exemple, on dispose dans MedDRA de 25 termes, dispersés dans la structure hiérarchique, pour coder une neuropathie en fonction de sa localisation anatomique ou de son évolution. La structure hiérarchique actuelle de la terminologie (et c'est également vrai pour WHO-ART) ne suffit pas pour regrouper par raisonnement ces termes. Plus généralement, en l'ab-

sence d'une sémantique formelle liée aux termes, il n'est pas possible de réaliser de façon efficace un regroupement algorithmique de termes désignant des conditions cliniques similaires. La principale conséquence en est que des signaux d'EIM potentiels qui pourraient être regroupés et proposer un signal d'EIM plus plausible ne le sont pas. Pour lutter contre cela, des experts de l'OMS ont travaillé à un regroupement manuel de termes préférés MedDRA dans des grands groupes de pathologies pouvant correspondre à un EIM.

#### 4.1.3. *Objectif*

Notre objectif est de construire une ontologie des effets indésirables dans laquelle chaque terme est défini formellement. La conséquence de ces définitions formelles est de structurer les terminologies pour faciliter le regroupement sémantique des concepts qui sont en relation selon certains critères médicaux. L'idée sous-jacente est de pouvoir construire des groupes de termes en fonction de la demande de l'utilisateur en s'appuyant sur la classification automatique des concepts à partir de leur définition formelle. L'étude qui est présentée ici est restreinte à la terminologie WHO-ART.

Cette section consacrée à la pharmacovigilance décrit l'application de plusieurs méthodes du web sémantique :

- l'alignement d'une terminologie (WHO-ART) avec une ontologie,
- la modélisation ontologique d'une liste de termes au moyen d'une méthode d'extraction des connaissances à partir d'un réseau sémantique,
- la structuration de l'ontologie résultante au moyen de concepts primitifs et de concepts définis avec le langage OWL,
- le raisonnement sur l'ontologie afin de construire des groupes de termes qui désignent une condition clinique similaire au moyen du moteur d'inférence RACER (Haarslev *et al.*, 2001).

L'étude qui est présentée ici est restreinte à la terminologie WHO-ART (WHO-ART compte environ 1 900 termes préférés et est entièrement incluse dans MedDRA qui compte environ 15 000 termes préférés).

## 4.2. *Hypothèses*

L'investissement humain, nécessaire pour la construction et la maintenance d'une ontologie comportant plusieurs milliers de concepts, est très lourd si l'on souhaite produire une solution opérationnelle. Plusieurs travaux de modélisation de terminologies médicales ont déjà été mis en oeuvre par d'autres équipes. Il est donc plus économique de réutiliser une partie des concepts formels déjà proposés plutôt que de construire une solution sur mesure. Nous avons formulé l'hypothèse que les définitions formelles des concepts de WHO-ART peuvent être obtenues en ajoutant de nouvelles propriétés extraites d'un système terminologique plus sophistiqué tel que la SNOMED-CT. Il s'agit d'une ontologie généraliste destinée à la description des informations dans



le dossier patient informatisé qui fait l'objet d'une commercialisation au niveau international et dispose par conséquent de moyens financiers et humains conséquents. Par ailleurs, nous considérons que L'UMLS (*Unified Medical Language System*), qui propose une fusion de nombreux produits terminologiques médicaux (dont SNOMED-CT, MedDRA et WHO-ART) est le cadre approprié pour extraire ces définitions formelles.

### 4.3. Matériel et méthode

#### 4.3.1. Matériel

UMLS est un projet de la « National Library of Medicine (NLM) ». Il contient un métathésaurus peuplé de concepts médicaux et un réseau sémantique qui explicite les relations sémantiques entre les concepts. La raison d'être d'UMLS est de regrouper les différents noms/termes et les différentes vues d'un même concept qui proviennent de terminologies médicales diverses ainsi que d'identifier les relations utiles entre ces différents concepts.

WHO-ART est développé par l'OMS et s'organise sur trois niveaux hiérarchiques. Seul le niveau des termes préférés nous intéresse puisque c'est celui qui est utilisé pour coder les effets indésirables. Il existe 1857 termes préférés (comme par exemple : « Thrombophlébitis leg ») qui sont répartis dans 32 classes organes.

SNOMED-CT, est une ontologie dont l'objectif est le codage de l'information médicale. Elle n'est pas dédiée à la représentation des effets indésirables mais elle a le mérite de fournir une description formelle des concepts qu'il inclut (Spackman *et al.*, 1998). SNOMED-CT descend de la SNOMED internationale qui est un système terminologique poly-hiérarchique. Son utilisation augmente ainsi la possibilité de regroupements<sup>11</sup>.

Les versions des ressources que nous utilisons sont : UMLS 2005AA ; WHO-ART 2004- third quarter, RACER PRO 1.8, PROTÉGÉ 3.1 et PROTÉGÉ OWL Java API.

#### 4.3.2. Méthodes

Nous proposons une méthode générique pour structurer WHO-ART en s'appuyant sur la hiérarchie de SNOMED-CT. Notre méthode est organisée en plusieurs étapes. Tout d'abord, les termes WHO-ART sont mis en correspondance avec les termes SNOMED-CT en utilisant le métathésaurus d'UMLS. Deuxièmement, nous extrayons la sous-hiérarchie de SNOMED-CT qui correspondra à la hiérarchie des concepts primitifs de notre ontologie. Troisièmement, les définitions des termes WHO-ART (qui

---

11. La SNOMED a ainsi 2 versions différentes. C'est à l'origine une classification d'environ 250 000 concepts, la SNOMED-I v3.5. Elle a été ensuite enrichie et « ontologisée », c'est la SNOMED-CT de 366 170 concepts. Le passage de l'une à l'autre se fait non sans mal en raison de la volonté des acteurs de conserver tous l'existant, en particulier tous les termes du thésaurus (Spackman *et al.*, 2002; Charlet, 2002). Enfin, le principal écueil de la SNOMED-CT est qu'elle n'est pas disponible en français, ce qui est rédhibitoire pour les applications médicales.

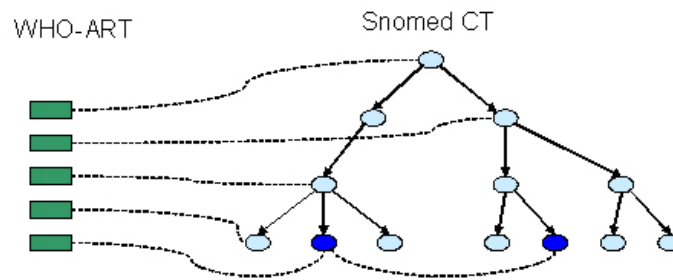
désignent les concepts définis de notre ontologie) sont écrites en OWL et automatiquement classées pour produire l'ontologie des EIM. La quatrième étape est la validation.

#### 4.3.2.1. Alignement de WHO-ART avec SNOMED-CT

Au cours de la 1<sup>re</sup> étape, nous créons une table de correspondance entre les termes préférés WHO-ART et leurs synonymes dans la SNOMED-CT. Cette étape s'appuie sur le fait qu'un concept UMLS regroupe des termes synonymes de plusieurs terminologies médicales dont WHO-ART et la SNOMED-CT.

#### 4.3.2.2. Extraction de la hiérarchie de concepts primitifs

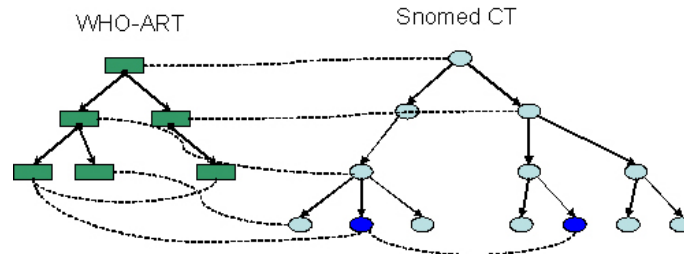
A partir des termes SNOMED-CT obtenus dans l'étape précédente, nous reconstruisons la partie de la structure de SNOMED-CT qui inclut tous les termes de WHO-ART et uniquement ceux-là (cf. figure 6).



**Figure 6.** Correspondance entre les termes WHO-ART et les termes SNOMED-CT

Les concepts de SNOMED-CT qui sont liés aux termes WHO-ART deviennent les concepts primitifs de notre ontologie. Nous considérons une seule propriété, la synonymie, pour extraire les définitions formelles. La définition d'un concept WHO-ART, notée *ConceptW* est l'union des restrictions existentielles de tous les concepts SNOMED-CT « synonymes » du concept WHO-ART. Le choix de l'opérateur union provient du fait qu'il permet d'obtenir plus de regroupements que l'intersection (mais il ramène aussi plus de bruit). La condition nécessaire « partial » exprime que le concept défini appartient à la hiérarchie WHO-ART. Ainsi, la hiérarchie des concepts primitifs (SNOMED-CT) est clairement séparée des concepts définis (WHO-ART) (cf. figure 7).

Les définitions établies de cette manière sont traitées par un moteur d'inférence afin d'être classées. Le résultat de cette procédure est une nouvelle structure WHO-ART liée à la SNOMED-CT qui indique, par exemple, que le concept WHO-ART « Thrombosis arterial », est synonyme du concept SNOMED-CT « Arterial thrombosis ».



**Figure 7.** *Hierarchie des concepts primitifs et définis*

#### 4.3.2.3. Validation

L'idée principale pour valider la nouvelle structure de WHO-ART est de comparer des groupes de termes élaborés manuellement par les experts et disponibles dans MedDRA avec les groupes de termes obtenus par raisonnement à partir de notre ontologie. Cette stratégie est possible en conservant uniquement les termes MedDRA qui sont également des termes WHO-ART (rappelons que WHO-ART est inclus dans MedDRA) et qui ont un synonyme dans la SNOMED-CT.

Deux situations sont ensuite comparées : la correspondance et la composition. Dans le cas le plus fréquent, la correspondance, le groupe de termes peut être directement créé en sélectionnant un concept SNOMED-CT et en récupérant tous les concepts qu'il subsume. C'est le cas lorsque le terme qui désigne le groupe de termes a directement un synonyme dans la SNOMED-CT. Dans le cas où un seul concept SNOMED-CT n'est pas suffisant, la composition, nous considérons l'union de plusieurs concepts SNOMED-CT.

#### 4.4. Résultats

Au cours de la première étape de la méthode, 85.9 % (1 597) des termes WHO-ART ont été mis en correspondance avec les termes SNOMED-CT. L'ontologie contient 7 357 concepts dont 1 596 concepts définis. La hiérarchie de concepts primitifs extraite pour les effets indésirables représente 1.6 % de la SNOMED-CT.

Nous illustrons notre propos avec l'exemple d'un médicament X qui semble associé à des épisodes d'hémorragie ou de perforation de la partie supérieure du tube digestif. Afin d'extraire de notre base de pharmacovigilance l'ensemble des déclarations spontanées décrivant cette condition clinique, nous avons besoin de constituer le groupe de termes correspondant. Nous recherchons alors les concepts SNOMED-CT qui couvrent le sens de ce groupe de termes et réalisons donc l'union des concepts SNOMED-CT décrivant d'une part les hémorragies digestives hautes et d'autre part les

perforations digestives hautes. Ainsi, la définition correspondant au groupe de termes « saignement ou perforation haute gastrointestinale » est :

$WSC_1 = (is\_synonym\_with\ some\ "upper\ gastrointestinal\ hemorrhage" \ or\ (is\_synonym\_with\ some\ "gastrointestinal\ perforation"))^{12}$

La constitution du groupement par les deux concepts SNOMED-CT, *upper gastrointestinal hemorrhage* et *gastrointestinal perforation* et les concepts subsumés par ceux-là, nous permet de retrouver 13 concepts. Ce résultat est à comparer aux 18 retrouvés manuellement par les experts. Il correspond à un taux de 76 %. Sur 8 groupes de termes évalués, le taux est en moyenne de 71 % avec des écarts entre 20 et 94 %.

Ces résultats sont développés dans (Alecú *et al.*, 2006). L'amélioration de la couverture des groupes de termes passe par l'enrichissement des définitions formelles *via* la définition d'autres relations sémantiques comme des relations d'évocations entre signes et pathologies permettant de regrouper des déclarations d'EIM portant sur les pathologies ou les seuls signes biologiques. Ces relations peuvent être extraites de SNOMED-CT dans un premier temps. C'est le travail dans lequel nous sommes actuellement engagés.

## 5. Discussion

### 5.1. Le primat de la terminologie sur le raisonnement

Les 3 projets décrits ici visent tous à construire des ontologies mais prennent en compte la dimension linguistique de la médecine : en particulier, dans 2 des 3 projets, les liens sont faits entre les ontologies construites et les thésaurus incontournables car correspondant au mode d'expression des connaissances et à un historique ne se démentant pas et extrêmement important en médecine. Dans la même proportion de 2 sur 3, la construction d'ontologies se fait avec des outils de TAL appliqués aux textes générés par les professionnels de santé durant leur activité de soin (comptes rendus d'hospitalisation, texte libre) ou pédagogique (livre de cours, guide de bonne pratique). Sans nier l'utilité d'applications mettant en œuvre des couches d'inférences sur des ontologies, il est clair que, quand on veut automatiser des actions de traitement des connaissances dans le domaine médical, donc mettre en œuvre le web sémantique médical (Léger *et al.*, 2005), il est indispensable de tenir compte du caractère majoritairement documentaire et linguistique de la médecine (Charlet, 2007).

Ce constat rejoint des questionnements et des propositions faites dans (Bourigault *et al.*, 2004) pour d'autres domaines que la médecine : les domaines d'expertise s'exprimant principalement en langue, leur formalisation à tout avantage à utiliser les traces linguistiques des activités expertes.

12. WHO-ART *special category*, un regroupement fait manuellement par les experts.

## 5.2. Niveaux et méthodes de construction d'ontologies

Les projets développés ici, en même temps que le travail d'autres chercheurs<sup>13</sup> nous ont montré que les ontologies de domaines que nous développons dans nos applications n'étaient pas du tout aussi universelles que l'ingénierie des connaissances l'espérait au début des années 1990. Au contraire, il est maintenant clair qu'analyser le monde pour le catégoriser est extrêmement dépendant de domaines d'activités, et donc des activités supportées par le système à base des connaissances que l'on fabrique.

Une ontologie est maintenant classiquement découpée en 3 niveaux :

1) la « top-ontologie », que l'on devrait plus précisément appeler « ontologie formelle » pour reprendre l'appellation des philosophes (Bachimont, 2004). C'est le niveau le plus abstrait structurant les connaissances de haut niveau avec des catégories dont l'organisation dépend de réflexions philosophiques. La question de l'unicité ou non de cette ontologie est un débat dans lequel nous n'entrerons pas ici<sup>14</sup> ;

2) la *core*-ontologie, fournissant les concepts structurant du domaine et décrivant les relations entre ces concepts – en médecine, on y trouve des concepts de *diagnostic*, *signe*, *structure anatomique* et des relations comme celles liées à la localisation d'une pathologie sur une structure anatomique ;

3) l'ontologie du domaine, c'est-à-dire les concepts du domaine tels qu'ils sont manipulés par les professionnels – ici de santé. Le troisième et dernier niveau est celui que l'on construit avec les outils de TAL puisque l'on analyse les documents produits en activité avec ceux-ci.

Le second niveau correspond ici à la façon dont est décrite la médecine. Les différentes expériences que nous avons menées jusque-là – voir les 3 projets dans cette article mais aussi le projet MENELAS dans le domaine de la chirurgie cardiaque (Zweigenbaum *et al.*, 1995; Charlet, 2002) ou le projet RÉA dans le domaine de la réanimation chirurgicale (Le Moigno *et al.*, 2002a) – nous ont montré que les ontologies construites traduisent un point de vue médical sur la spécialité médicale modélisée mais aussi sur le reste du patient. Ainsi, si un chirurgien cardiaque s'intéresse d'abord au cœur et à tous les vaisseaux autour, il a un point de vue sur les nerfs, conduits qu'il ne faut pas couper durant une intervention. Chacune des modélisation de spécialité prend en compte tout le corps humain à des niveaux de granularité et de spécialisation différents. C'est ce qui rend ces modélisations irréductibles les unes aux autres et impossibles à combiner. À l'inverse, on peut penser que la *core*-ontologie qui modélise la médecine reste valable pour toutes les spécialités, dans le contexte de la médecine occidentale. Nous effectuons donc en ce moment un travail où nous essayons de normaliser et de réutiliser la top-ontologie et la *core*-ontologie du projet MENELAS dans les autres ontologies, en particulier la pneumologie.

13. Voir en particulier le groupe TIA <<http://tia.loria.fr>> et (Bourigault *et al.*, 2004).

14. cf. (Bachimont, 2004).

### 5.3. Créer des correspondances entre ressources terminologiques et ontologiques

Dans la présentation de la pharmaco-vigilance, nous avons abordé UMLS qui regroupe les différents noms/termes et les différentes vues d'un même concept qui proviennent de terminologies médicales diverses et qui identifie les relations utiles entre ces différents concepts (cf. section 4.1). Dans la description de notre travail, nous avons montré l'utilité et l'utilisation d'UMLS pour arriver à nos fins. Ainsi UMLS concourt à résoudre la problématique de la présence de nombreuses classifications en médecine et le passage de l'une à l'autre (Zweigenbaum, 2004; McCray *et al.*, 1996). Le problème est que sa couverture est tellement grande et le nombre de ressources répertoriées tellement disparates que UMLS ne peut pas être utilisé dans une application à caractère industriel pour faire des correspondances entre ontologies. Pour organiser le lien entre plusieurs ontologies dans le cadre d'un système d'information de santé, il faut alors construire un *serveur de terminologie*.

Ainsi, un grand nombre de terminologies, développées dans des contextes et pour des objectifs bien précis, sont utilisées pour le codage des informations cliniques du dossier patient informatisé (DPI). Les concepteurs de systèmes terminologiques ont concentré leurs efforts sur la conceptualisation formelle de domaines afin d'améliorer les capacités d'exploitation (recherche d'information, classification, etc.) et de partage des données. Des travaux sont en cours afin de rendre ces systèmes terminologiques utilisables pour le codage de l'information clinique en pratique quotidienne. Ces efforts consistent à définir des liens entre les systèmes terminologiques et les modèles d'information des DPI.

A ce titre, Rector et d'autres soulignent que les systèmes terminologiques doivent répondre à une double exigence qui est d'être utilisables et compréhensibles à la fois par des humains (représentation flexible tenant compte des termes et des expressions pré ou post-coordonnées) et par des machines (représentation formelle autorisant le raisonnement terminologique) (Rector, 1999; Charlet, 2007). Par rapport à ce critère, il apparaît que les systèmes terminologiques de référence, tels que la SNOMED par exemple, ne sont pas facilement utilisables pour le codage en routine car peu lisibles pour des humains.

D'autres types de terminologies peuvent compléter utilement les terminologies de référence. Spackman distingue trois types de terminologies : les terminologies de référence pour l'archivage et l'exploitation des données, les terminologies d'interface pour la saisie des données et les terminologies permettant l'optimisation du traitement automatique des langues. Les deux derniers types de terminologies peuvent être utilisés dans le cadre de solutions de codage pour les informations du DPI, qu'elles soient structurées ou en langue naturelle (Spackman *et al.*, 1997). Dans une revue dédiée aux terminologies d'interface, Rosenbloom définit les principales caractéristiques de ces systèmes terminologiques dont le rôle est de représenter les concepts médicaux utiles pour le codage en gérant la synonymie, en fournissant un juste compromis entre pré et post-coordination et en établissant les correspondances avec les terminologies de référence plus formelles (Rosenbloom *et al.*, 2006). Le projet OntoPneumo, décrit

précédemment (cf. section 2) comme celui de la pharmaco-vigilance (cf. section 4), cherchent à aligner et restructurer des terminologies d'interface (thésaurus de spécialité pour le premier, WHO-ART pour le second) avec des ontologies de référence (ontologie de la pneumologie pour le premier, SNOMED-CT pour le second). Le projet sur l'hypertension ne vise pas cette problématique d'alignement directement mais un travail a été effectué pour comparer l'ontologie construite avec la SNOMED-CT et a permis de constater que l'alignement est possible avec certains concepts mais que, pour d'autres, la SNOMED-CT n'est pas assez granulaire (Steichen *et al.*, 2007).

Ces travaux montrent de plus *a)* qu'une terminologie de référence doit être formelle et doit donc être une ontologie et *b)* que la caractéristique de terminologie de référence est relative au domaine. Ainsi, si on s'en tient à la pneumologie, l'ontologie OntoPneumo développée peut servir de terminologie de référence pour les différents thésaurus du domaine alors que si l'on élargit le domaine à toute la médecine, la SNOMED est un candidat à être une terminologie de référence par rapport à OntoPneumo.

## 6. Conclusion

Le web sémantique réclame par définition de construire des ontologies permettant, par leur caractère formel, d'automatiser un certain nombre de tâches. Dans cet article, nous avons voulu montrer que la médecine a des spécificités qui contraignent l'usage des ontologies et leur construction :

- jusqu'à récemment, les thésaurus ont permis de satisfaire les buts de conceptualisation liée à l'exploitation – recherche d'information, classification, etc. – et le partage des données. Une meilleure automatisation et prise en compte de la dimension des données médicales passe par l'usage des ontologies dans le contexte du web sémantique. Un certain nombre de projets nous ont montré la nécessité de prendre spécifiquement en compte le caractère linguistique des thésaurus et de les utiliser en coordination avec les ontologies pour exploiter les données en leur conservant leur accessibilité cognitive ;

- dans ce contexte, la construction des ontologies peut prendre en compte l'expression linguistique des connaissances de 2 façons : *a)* en tenant compte des thésaurus dont la dimension terminologique a été historiquement prise en compte à la création ou *b)* en utilisant les textes générés par les professionnels de santé durant leur activité qui fournissent une ressource traçant les conceptualisations d'un domaine ;

- enfin, il reste encore du travail pour préciser les conditions d'usage et donc de construction des différentes terminologies impliquées dans une application réelle, en particulier les rapports des différentes terminologies entre elles, des moins formelles proches de la langue aux plus formelles loin de la langue. C'est à ce problème que s'attaquent maintenant les recherches sur les serveurs de terminologie.

## 7. Bibliographie

- Alecu I., Bousquet C., Mougin F., Jaulent M.-C., « Mapping of the WHO-ART terminology on Snomed CT to improve grouping of related adverse drug reactions », in A. Hasman, R. Haux, J. van der Lei, E. D. Clercq, F. H. R. France (eds), *Actes Medical Informatics Europe*, vol. 124 of *Studies in Health Technology and Informatics*, IOS Press, Maastricht, p. 833-8, 2006.
- Aussenac-Gilles N., Bourigault D., « The Th(IC)2 Initiative: Corpus-Based Thesaurus Construction for Indexing WWW Documents », in R. Dieng (ed.), *Proc. of the EKAW conference*, Juan-les-Pins, France, 2000.
- Bachimont B., Arts et sciences du numérique : Ingénierie des connaissances et critique de la raison computationnelle, PhD thesis, Université Technologique de Compiègne, janvier, 2004.
- Bachimont B., Isaac A., Troncy R., « Semantic Commitment for Designing Ontologies: A Proposal », in A. Gomez-Pérez, V. Benjamins (eds), *13<sup>th</sup> International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, vol. (2473) of *Lecture Notes in Artificial Intelligence*, Springer Verlag, Sigüenza, Espagne, p. 114-121, 1-4 Octobre, 2002.
- Baneyx A., Construire une ontologie de la pneumologie : aspects théoriques, modèles et expérimentations., PhD thesis, Université Pierre et Marie Curie (Paris VI), février, 2007. Disponible à <http://tel.archives-ouvertes.fr/tel-00136937/fr/>.
- Bensadoun H., « PMSI et chirurgiens : pourquoi les chirurgiens doivent-ils coder, comment bien coder? », *Journal de Chirurgie*, Masson, février, 2001.
- Bourigault D., Aussenac-Gilles N., Charlet J., « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas », in Pierrel *et al.* (2004), 2004.
- Bourigault D., Lame G., « Analyse distributionnelle et structuration de terminologie. Application à la construction d'une ontologie documentaire du Droit », *Traitement automatique des langues*, 2002.
- Brown E. G., « Effects of coding dictionary on signal generation: a consideration of use of MedDRA compared with WHO-ART », *Drug Safety*, vol. 25, n° 6, p. 445-52, 2002.
- Charlet J., L'Ingénierie des connaissances : développements, résultats et perspectives pour la gestion des connaissances médicales, Habilitation à diriger des recherches, Université Paris 6, décembre, 2002.
- Charlet J., « The Management of Medical Knowledge : between non-structured documents and Ontologies », *Annales des télécommunications*, vol. 62, n° 7-8, p. 808-26, 2007.
- Coiera E., « Medical Informatics », *BMJ*, vol. 310, n° 6991, p. 1381-7, 1995.
- Degoulet P., Chatellier G., Devriès C., Lavril M., Ménard J., « Computer-Assisted Techniques for Evaluation and Treatment of Hypertensive Patients », *Am J Hypertens*, vol. 2, n° 3, p. 156-63, 1990.
- Edwards I., Biriell C., « Harmonisation in pharmacovigilance », *Drug Safety*, vol. 10, n° 2, p. 93-102, 1994.
- Friedman C., Shagina L., Lussier Y., Hripcsak G., « Automated encoding of clinical document based on natural language processing », *Journal of the American Medical Informatics Association*, vol. 11, p. 392-402, 2004.



- Friedman C., Wyatt J. C., « Subjectivist Approaches to Evaluation », in C. Friedman (ed.), *Evaluation Methods in Biomedical Informatics*, 2 edn, Springer, New-York, chapter 9, p. 248-66, 2006.
- Green J., Britten N., « Qualitative Research and Evidence-Based Medicine », *BMJ*, vol. 316, n° 7139, p. 1230-2, 1998.
- Haarslev V., Möller R., « Description of the RACER System and its Applications », *International Workshop on Description Logics proceedings*, 2001.
- Harris Z., *Mathematical Structures of Language*, John Wiley and Sons, New-York, USA, 1968.
- Haynes R. B., Devereaux P. J., Guyatt G. H., « Clinical Expertise in the Era of Evidence-Based Medicine and PatientChoice », *Evid Based Med*, vol. 7, n° 2, p. 36-8, 2002.
- Le Moigno S., Charlet J., Bourigault D., Degoulet P., Jaulent M.-C., « Terminology extraction from text to build an ontology in surgical intensive care », in I. S. Kohane (ed.), *Actes AMIA Annual Fall Symposium 2002*, AMIA, San Antonio, TX, p. 233-52, novembre, 2002a.
- Le Moigno S., Charlet J., Bourigault D., Jaulent M.-C., « Construction d'une ontologie à partir de corpus : expérimentation et validation dans le domaine de la réanimation chirurgicale », in B. Bachimont (ed.), *Actes des 6<sup>es</sup> Journées Ingénierie des Connaissances*, Rouen, France, p. 229-38, 28-30 mai, 2002b.
- Léger A., Nixon L. J., Shvaiko P., Charlet J., « Semantic Web applications: Fields and Business cases. The Industry challenges the research », in M. Bramer, V. Terziyan (eds), *Industrial Applications of Semantic Web*, vol. Vol. 188 of *IFIP International Federation for Information Processing*, Springer, chapter Part 1. Invited Key-note Talks, p. 27-47, 2005. Proceedings of the 1st International IFIP/WG12.5 Working Conference on Industrial Applications of Semantic Web.
- McCray A., Razi A., Bangalore A., Browne A., Stavri P., « The UMLS knowledge source server: a versatile internet-based research tool », *Journal of the American Medical Informatics Association*, vol. 3, n° suppl., p. 164-168, 1996.
- Pierrel J.-M., Slodzian M. (eds), *Techniques informatiques et structuration de terminologies*, Hermès, 2004.
- Rector A. L., « Thesauri and Formal Classifications: Terminologies for People and Machines », *Methods of Information in Medicine*, vol. 37, n° 4-5, p. 501-509, 1998.
- Rector A. L., « Clinical terminology: why is it so hard? », *MIM*, vol. 38, n° 4-5, p. 239-52, 1999.
- Rosenbloom S. T., Miller R. A., Johnson K. B., « Interface terminologies: facilitating direct entry of clinical data into electronic health record systems », *J Am Med Inform Assoc*, vol. 13, n° 3, p. 277-88, 2006.
- Spackman K. A., Campbell K. E., Cote R. A., « SNOMED RT: a reference terminology for health care », vol. 4, p. 640-4, 1997.
- Spackman K. A., Dionne R., Mays E., Weis J., « Role grouping as an Extension to the Description Logic of Ontolog Motivated by Concept Modeling in SNOMED », p. 712-6, 2002.
- Spackman K., Campbell K., « Compositional Concept Representation using SNOMED: Towards Further Convergence of Clinical Terminologies », *Journal of the American Medical Informatics Association*, vol. 5, n° suppl., p. 740-744, 1998.
- Staab S., Steider R. (eds), *Handbook on Ontologies*, Springer-Verlag, 2004.

- Steichen O., Daniel-Le Bozec C., Jaulent M.-C., Charlet J., « Construction d'une ontologie pour la prise en charge de l'hypertension artérielle », in F. Trichet (ed.), *Actes des 18<sup>es</sup> Journées Ingénierie des Connaissances*, Cépaduès, Grenoble, France, p. 241-52, 4-6 juillet, 2007.
- Zweigenbaum P., « L'UMLS entre langue et ontologie : une approche pragmatique dans le domaine médical », in Pierrel *et al.* (2004), p. 111-37, 2004.
- Zweigenbaum P., Bachimont B., Bouaud J., Charlet J., Boisvieux J.-F., « Issues in the Structuring and Acquisition of an Ontology for Medical Language Understanding », *Methods of Information in Medicine*, 1995.

Article reçu le 15 juin 2007  
Version révisée le 19 juin 2008

**Jean Charlet** est docteur en informatique et habilité à diriger des recherches depuis 2002. Il est chargé de mission à l'Assistance Publique-Hôpitaux de Paris et chercheur associé à l'INSERM UMR\_S 872, éq20. Ses principaux centre d'intérêt sont la représentation des connaissances médicales à travers les ontologies et les documents semi-structurés. Il s'intéresse spécialement à la construction des ontologies. Il est responsable de la communauté de recherche Ingénierie des connaissances en France.

**Audrey Baneyx** est docteur en informatique de l'université Pierre et Marie Curie, Paris VI, depuis 2007. Son domaine de recherche est l'ingénierie des connaissances et, principalement, la représentation et l'indexation des connaissances à l'aide d'ontologies. Il s'étend de l'élaboration de ces ontologies à leur exploitation dans des systèmes informatiques. Elle est actuellement post-doctorante dans l'INSERM UMR\_S 872, éq20.

**Olivier Steichen** est médecin, Chef de Clinique - Assistant dans le service de médecine interne de l'hôpital Tenon à Paris est doctorant à l'INSERM UMR\_S 872, éq20. Son domaine de recherche est l'individualisation des décisions médicales en pratique clinique, tout particulièrement dans le domaine de l'hypertension artérielle.

**Iulian Alecu** est médecin. Il est actuellement pharmacovigilant dans le service de Pharmacovigilance des Laboratoires Bayer Schering Pharma et doctorant à l'INSERM UMR\_S 872, éq20. Son domaine de recherche est l'extraction de connaissances à partir des terminologies normalisées utilisées dans le domaine médical afin de rendre plus facile l'accès aux informations sur la tolérance des médicaments, notamment les rapports de pharmacovigilance, et de rendre possible l'aperçu, en temps réel, des variations du rapport bénéfice/risque des médicaments.

**Christel Daniel-Le Bozec** est anatomo-pathologiste, MCU-PH en informatique médicale à l'université Paris Descartes et à l'hôpital Européen Georges Pompidou. Elle est chercheur à l'INSERM UMR\_S 872, éq20. Son domaine de recherche est la représentation et l'exploitation des informations cliniques du Dossier Patient Informatisé (DPI) et particulièrement l'utilisation de systèmes terminologiques pour le codage des informations cliniques.

**Cedric Bousquet** est pharmacien, praticien hospitalier dans le Service de santé publique et d'information médicale au CHU de Saint-Etienne, et fait partie de l'INSERM UMR\_S 872, éq20. Son domaine de recherche est l'utilisation des méthodes d'ingénierie des connaissances pour améliorer la détection des signaux en pharmacovigilance, en particulier la représentation ontologique et le raisonnement sur les termes désignant des effets indésirables liés aux médicaments.

**Marie-Christine Jaulent** est directeur de recherche à l'INSERM et dirige, depuis 2002, l'INSERM UMR\_S 872, éq20 spécialisée dans le domaine de l'ingénierie des connaissances en santé. Ses thématiques de recherche se rattachent à l'ingénierie des connaissances : la modélisation et la représentation des connaissances, les systèmes informatiques d'aide à la décision (raisonnement à partir de cas, reconnaissance des formes et classification, méthodes d'apprentissage symbolique), la prise en compte de l'imprécis et de l'incertain dans la modélisation des connaissances et de la décision, la logique floue.

**ANNEXE POUR LE SERVICE FABRICATION**  
A FOURNIR PAR LES AUTEURS AVEC UN EXEMPLAIRE PAPIER  
DE LEUR ARTICLE ET LE COPYRIGHT SIGNÉ PAR COURRIER  
LE FICHIER PDF CORRESPONDANT SERA ENVOYÉ PAR E-MAIL

1. ARTICLE POUR LA REVUE :  
*RSTI - TSI – 28/2009. Web sémantique*
2. AUTEURS :  
*Jean Charlet<sup>\*,\*\*</sup> — Audrey Baneyx<sup>\*</sup> — Olivier Steichen<sup>\*,\*\*\*</sup> — Iulian Alecu<sup>\*</sup> — Christel Daniel-Le Bozec<sup>\*,\*\*\*</sup> — Cédric Bousquet<sup>\*</sup> — Marie-Christine Jaulent<sup>\*</sup>*
3. TITRE DE L'ARTICLE :  
*Utiliser et construire des ontologies en médecine : le primat de la terminologie*
4. TITRE ABRÉGÉ POUR LE HAUT DE PAGE MOINS DE 40 SIGNES :  
*Des ontologies en médecine*
5. DATE DE CETTE VERSION :  
*13 février 2009*
6. COORDONNÉES DES AUTEURS :
  - adresse postale :
    - <sup>\*</sup> Centre des Cordeliers, INSERM, U872, Éq. 20, Paris, F-75006 France ;  
Université Pierre et Marie Curie, Paris ; Université Paris Descartes, Paris.
    - <sup>\*\*</sup> AP-HP, DSI/STIM, Paris, F-75004 France.  
Jean.Charlet@spim.jussieu.fr
    - <sup>\*\*\*</sup> AP-HP, HEGP, Paris, F-75015 France.
  - téléphone : 00 00 00 00 00
  - télécopie : 00 00 00 00 00
  - e-mail : guillaume.laurent@ens2m.fr
7. LOGICIEL UTILISÉ POUR LA PRÉPARATION DE CET ARTICLE :  
L<sup>A</sup>T<sub>E</sub>X, avec le fichier de style article-hermes2.cls,  
version 1.23 du 17/11/2005.
8. FORMULAIRE DE COPYRIGHT :  
Retourner le formulaire de copyright signé par les auteurs, téléchargé sur :  
<http://www.revuesonline.com>

SERVICE ÉDITORIAL – HERMES-LAVOISIER  
14 rue de Provigny, F-94236 Cachan cedex  
Tél. : 01-47-40-67-67  
E-mail : [revues@lavoisier.fr](mailto:revues@lavoisier.fr)  
Serveur web : <http://www.revuesonline.com>