Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation

Jean Charlet, Gunnar Declerck, Ferdinand Dhombres, Pierre Gayet, Patrick Miroux, Pierre-Yves Vandenbussche

Construire une ontologie médicale pour la recherche d'information : problématiques terminologiques et de modélisation

Jean Charlet[1,2], Gunnar Declerck[1], Ferdinand Dhombres[1,3], Pierre Gayet[4], Patrick Miroux[4], Pierre-Yves Vandenbussche[1,5]

[1] INSERM UMRS 872 ÉQ.20, Ingénierie des connaissances en santé, Paris, France.
Jean.Charlet@upmc.fr

[2] Assistance Publique – Hôpitaux de Paris, Paris, France.

[3] Hôpital Armand Trousseau, AP-HP, Paris, France.

[4] Centre hospitalier de Compiègne, France.

[5] Société MONDECA, Paris, France.

Résumé : Pour un projet de Recherche d'information, nous avons développé ONTOLURGENCES, une ressource termino-ontologique qui assure *a*) le rôle de modèle du domaine répertoriant tous les concepts pertinents et *b*) le lien entre les concepts et la façon dont ils sont nommés dans les documents du Dossier patient informatisé. Cette double fonction permet l'annotation et l'indexation de dossiers patients et la recherche d'informations dans les dossiers indexés. Le développement d'ONTOLURGENCES a été réalisé en 6 étapes dont nous décrivons les principales ici. Ce projet montre (1) que la viabilité d'une telle ressource présuppose une articulation précise des concepts et des termes, et (2) qu'un tel prérequis peut être atteint par la mise en œuvre de procédures « industrielles » fondées sur une architecture de métamodélisation permettant de modéliser l'ensemble des Systèmes d'Organisation des Connaissances et structures de connaissances nécessaires.

Mots-clés : Ontologie, métamodélisation, terminologie

1 Introduction

L'utilisation de systèmes terminologiques pour la création d'ontologies ne va pas sans poser d'importants problèmes. Bien que les ontologies aient un rôle normatif analogue aux terminologies – mettre en place un vocabu-

laire commun et faire usage de représentations et concepts partagés, afin de permettre l'interopérabilité des documents et faciliter l'élaboration de connaissances – leur approche formelle de la sémantique les distingue également de manière nette de ces dernières. Les ontologies sont des architectures de *concepts*, non des listes organisées de *termes*. Les concepts, à la différence des termes, se caractérisent par des définitions *formelles*. C'est ce caractère formel qui permet à l'information d'être manipulée par les machines.

Pourtant, le recours à des terminologies, ou plus radicalement encore à des corpus de textes, pour la création d'ontologies est parfois inévitable. Si l'ontologie a pour vocation à être intégrée à un système de traitement automatique de documents, par exemple de recherche d'informations (RI), les concepts doivent pouvoir être appariés avec les termes apparaissant dans les documents pour que le traitement de l'information soit possible. L'ontologie doit assurer la *couverture terminologique* du domaine, c'est-à-dire que les termes utilisés dans le domaine doivent pouvoir permettre de renvoyer aux concepts de l'ontologie (*cf. infra*). Les représentations conceptuelles se trouvent sinon inexploitables.

Le projet LERUDI (LEcture Rapide en Urgence du Dossier patient Informatisé) vise à développer un système d'information (SI) offrant aux professionnels de santé une vision synthétique du dossier patient informatisé (DPI), et la possibilité d'un parcours rapide de celui-ci, pour permettre des prises de décisions médicales soumises à d'importantes contraintes de temps. Le champ d'expérimentation de ce projet est la lecture d'un dossier hospitalier par un médecin urgentiste. LERUDI est ainsi un système de RI fondé sur une Ressource Termino-ontologique (RTO) [1] nommée ONTOLURGENCES. Cette RTO assure a) le rôle de modèle du domaine répertoriant tous les concepts pertinents et b) le lien entre les concepts et la façon dont ils sont nommés dans les documents du Dossier patient. Cette double fonction doit permettre l'annotation et l'indexation de dossiers patients et la RI dans les dossiers indexés [2].

1. Une RTO (ou TOR en anglais, pour *Terminological or Ontological Resource*) se définit comme une ontologie dans laquelle les termes sont rattachés aux concepts de façon systématique et exhaustive. Il y a plusieurs façons de rattacher des termes à des concepts selon ce qu'on veut pouvoir représenter (Reymonet, 2007; Vandenbussche & Charlet, 2009).

2. L'indexation et la RI dans le DPI sont rendus nécessaire par le caractère textuel du DPI. De nombreuses tentatives de constitution de DPI structurées existent mais, pour des raison trop longues à développer ici, il restera toujours des textes dans les DPI (Bringay *et al.*, 2005).

Le développement d'ONTOLURGENCES a été réalisé en 6 étapes : (1) construction du squelette ontologique de la RTO grâce à une méthode à base d'analyse de corpus, (2) utilisation de ressources terminologiques et ontologiques existantes pour compléter manuellement le système de concepts de la RTO, (3) enrichissement automatique et (4) semi-manuel de de la RTO au niveau des termes, (5) enrichissement de la RTO en concepts en rapport avec les médicaments, et enfin (6) mise en œuvre de procédures de validation et assurance qualité.

Les 2 premières étapes de construction de la RTO correspondant à une méthode classique de construction d'ontologie, largement éprouvée dans notre équipe, elles n'ont pas posé de problèmes notables sauf sur les questions de la couverture du corpus. Les 3 étapes suivantes étant en revanche spécifiques au développement de cette RTO et à son usage dans un système de RI, elles ont posé plus de difficultés. L'enrichissement terminologique de la RTO a notamment nécessité des ressources extérieures, des *Systèmes d'organisation des connaissance* (SOC) [3]. La dernière étape, qui correspond à une action d'assurance qualité, est également spécifique à l'organisation de ce projet, et s'est avérée nécessaire au regard du nombre d'intervenants impliqués dans la construction de la RTO.

Nous voudrions ici montrer, à travers une description détaillée du processus ayant présidé à la construction et à la validation de cette RTO, au sein d'une équipe conséquente, (1) que la viabilité d'une telle ressource présuppose une articulation précise des concepts et des termes, et (2) qu'un tel prérequis peut être atteint par la mise en œuvre de procédures « industrielles » fondées sur une architecture de métamodélisation permettant de modéliser l'ensemble des SOC et structures de connaissances nécessaires.

Dans la section 2, nous présentons brièvement l'intérêt des ontologies pour la RI ; dans la section 3, nous décrivons les deux premièrs étapes de la construction de la RTO avec ses spécificités ; dans la section 4 nous présentons les étapes 3 et 4, liées à l'enrichissement de la RTO ; dans la section 5, nous présentons l'étape 6, correspondant aux procédures de validation et assurance qualité [4] ; enfin, dans la section 6, nous concluons et proposons quelques perspectives.

---

3. C'est le terme que nous utiliserons ici pour dénommer tout système terminologique.

4. Nous ne décrirons pas la 5e étape sur les médicaments, trop longue à développer ici.

2 Pourquoi une ontologie pour la recherche d'informations ?

Pour commencer, il faut évidemment se demander quel est l'intérêt d'une ontologie en RI. Dans notre cas, il est de pouvoir faire un certain nombre de raisonnements fondés sur la structure de l'ontologie et les relations entre les notions. Ainsi, en dehors des relations de subsomption, nous avons modélisé le lien entre des signes ou des maladies et des spécialités médicales. Ces liens permettent de remonter dans une interface – *i.e.* un nuage de mots – les spécialités médicales qui caractérisent un DPI.

Ensuite, une ontologie pour la RI a *de facto*, comme toute ontologie, une structure qui dépend de la tâche visée (Charlet *et al.*, 1996; van Heijst *et al.*, 1997). Cette structure n'est pas une qualité en soi pour la RI mais elle présente néanmoins un double intérêt : a) une ontologie bien structurée est plus facile à élaborer puis maintenir qu'une ontologie mal structurée ; b) une ontologie bien structurée permet des raisonnements valides. Ce second point est évidemment attendu de toute ontologie mais force est de constater qu'il n'est pas toujours satisfait. Une autre propriété importante que se doit de posséder une ontologie pour la RI est la couverture des termes du domaine dans lesquels sont exprimées les notions recherchées. Les deux exemples suivants permettront d'illustrer ces différents points :

Exemple de l'importance de la structuration formelle de la RTO.
Prenons l'exemple d'une question importante que se pose l'urgentiste au sujet d'un patient : « Mon patient a-t-il déjà été infecté par une entérobactérie ? ». Considérons que le dossier du patient comporte un document annoté avec le concept « Salmonelle ». Pour que le système puisse déduire que la salmonelle est une entérobactérie, il faut que la RTO déclare que le concept « Salmonelle » entretient une relation de spécialisation avec le concept « Entérobactérie ». Ainsi la réponse à la question de l'urgentiste sera positive *même si le document du patient n'est pas directement annoté avec le concept plus général d'entérobactérie*.

Exemple de l'importance de la couverture terminologique de la RTO.
L'annotation d'un document par un médicament comme le Paracétamol nécessite que la RTO comporte un concept unique lié à ce médicament et que l'on dispose des termes relatifs à la molécule chimique (paracétamol) et aux principales autres façons qu'ont les médecins de nommer le médicament dans le DPI, p.ex, « Dafalgan » et « param. »

On le comprend, la qualité des informations qui seront proposées à l'uti-

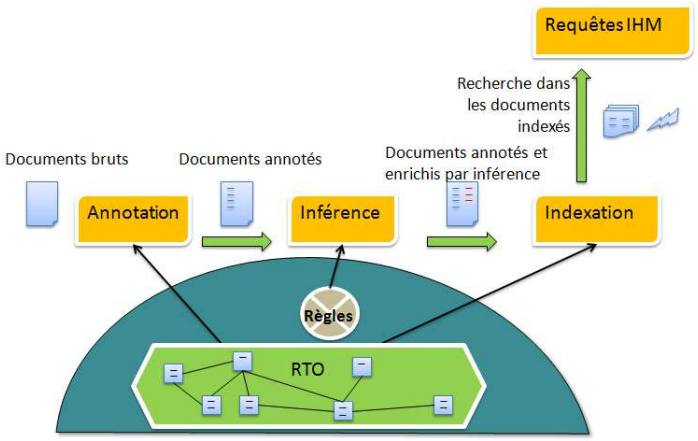

FIGURE 1 – Utilisation de la RTO dans le projet LERUDI. La Ressource Termino-Ontologique supporte les actions d'annotation, d'inférence et d'indexation.

lisateur du système de RI LERUDI dépend donc essentiellement de la qualité et de la richesse de la RTO utilisée. Les processus d'annotation, d'inférence et d'indexation reposent sur sa structuration formelle et sa richesse terminologique (c'est-à-dire sa capacité à couvrir les termes du domaine).

3 Les ressources ontologiques et terminologiques utilisées dans la conception d'ONTOLURGENCES

3.1 Champ conceptuel de l'ontologie

ONTOLURGENCES a été construite en plusieurs étapes et en utilisant différentes ressources. Le domaine cible de la RTO a été précisé graduellement. Dès le début du projet, nous nous sommes rendus compte que le domaine qui avait été initialement prévue pour la RTO devait évoluer. Nous étions partis sur l'idée de construire une ontologie représentant uniquement les concepts constituant la médecine d'urgence. Mais il s'est avéré que, du point de vue de la RI dans un DPI, une telle restriction était une erreur. En effet, le système de RI LERUDI doit permettre à un urgentiste de retrouver rapidement des concepts médicalement pertinents dans ce DPI. Or, ces concepts ne sauraient se réduire aux seuls concepts médicaux spécifiques aux urgences : ils peuvent au contraire relever de n'importe quelle spécialité médicale.

Au lieu de construire une ontologie des concepts des urgences, nous avons donc conçu une ontologie de la médecine dans sa généralité, mais réduite à la seule partie utile aux urgentistes. Cette décision n'a bien entendu pas été sans conséquences sur la taille de l'ontologie qui compte à ce jour environ 11 000 concepts.

Dans la suite de cette section, nous présentons les différentes étapes de l'élaboration d'ONTOLURGENCES et les ressources terminologiques ou ontologiques que nous avons utilisées. Nous n'abordons pas ici le problème de l'utilisation d'une top-ontologie, qui dépasse largement les questions débattues dans cet article [5]. Nous n'abordons pas pas non plus le problème de l'organisation de ces étapes et des cycles d'élaboration. Pour cette question, on pourra se reporter à (Dhombres *et al.*, 2010) et, pour ONTOLURGENCES, à (Charlet *et al.*, 2009). Précisons simplement que, durant le travail d'élaboration d'ONTOLURGENCES, nous avons respecté la méthode ARCHONTE développée par B. Bachimont (2002).

3.2 Traitement des données textuelles

Dans la méthode ARCHONTE, l'ontologie de domaine est construite à partir de l'analyse de documents générés durant l'activité à modéliser. Dans notre cas, nous avons rencontré de grandes difficultés pour obtenir des corpus qui pouvaient remplir cette fonction. En effet, les services d'urgences étant peu informatisés, et les documents écrits par les médecins urgentistes plus courts et moins nombreux que dans d'autres services, il était difficile de trouver des documents en assez grand nombre pour constituer le corpus visé.

Nous nous sommes pour cette raison rabattus sur deux autres types de documents : les actes des congrès *Urgences* de la discipline et les *guides de bonnes pratiques* du domaine. En dehors de la difficulté que nous avons eu à prétraiter ces corpus, le principal problème à été leur couverture par rapport à la cible visée. En effet, le corpus des actes de congrès, celui qui a été totalement traité, a montré ses limites en termes de domaine couvert. Comme on dit de façon familière, les communications de congrès se préoccupent bien souvent de « moutons à 5 pattes », très peu représentatifs des problèmes auxquels les urgentistes se trouvent quotidiennement confrontés. Un travail spécifique a permis de le montrer clairement en comparant

5. Nous utilisons par habitude le haut de l'ontologie de MÉNÉLAS (Charlet *et al.*, 2012). Pour des réflexions sur les top-ontologies, voir (Declerck *et al.*, 2012) dans ce volume.

les termes du domaine les plus fréquemment repérés dans le corpus avec l'incidence réelle des pathologies des urgences (Gayet *et al.*, 2010).

Ce problème de disponibilité de corpus a son importance : dans les domaines où l'on peut fonder la construction de l'ontologie sur des corpus analysés par des outils de traitement automatique des langues (TAL), l'usage des terminologies existantes intervient en validation du travail effectué. Dans le cas qui nous intéresse ici, elles interviennent beaucoup plus tôt dans le processus. Par ailleurs, comme nous l'avons vu en section 3.1, l'évolution du champ conceptuel justifie de s'intéresser à de nouveaux corpus ; point que l'on discutera dans la section 4.4.

3.3 Réutilisation du thésaurus de spécialité

Pour le codage PMSI [6], les urgentistes utilisent un extrait de la CIM-10 qui comporte 1000 termes environ. Ces termes couvrant une bonne partie du répertoire utilisé par les urgentistes pour le codage, il était nécessaire qu'ils soient représentés dans l'ontologie. Nous avons donc créé et défini pour chacun d'eux un concept.

Une des principales limites de ce travail a trait au fait que les termes de la CIM-10 sont adaptés au codage, mais que certains sont difficilement gérables dans une ontologie dans la mesure où ils regroupent plusieurs notions hétérogènes. Par exemple, on peut trouver des termes comme « Sujet attendant d'être admis ailleurs, dans un établissement adéquat » ou « Symptômes et signes relatifs aux fonctions cognitives et à la conscience, autres et non précisés ». Les concepts associés a de tels termes, parce qu'ils articulent de façon complexe une multitude de notions hétérogènes, restent difficilement modélisables.

3.4 Réutilisation de la CCAM

La CCAM (Classification Commune des Actes Médicaux) est une classification qui a l'avantage d'avoir été conçue par des équipes connaissant les ontologies. C'est, *a priori*, une classification dont chacun des concepts

6. 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.

a été validé par une représentation formelle (Rodrigues *et al.*, 1999). La réutilisation de la CCAM nous a ainsi permis de trouver une classification formée selon des principes constants.

Les problèmes sont plutôt venus de la façon dont est organisée la CCAM – pour des visées tarifaires – et des appellations utilisées pour les actes, là aussi construites pour des visées comptables et pas du tout adaptées à leur expression dans des documents médicaux – notre cible. La majeure partie du travail a ainsi consisté à renommer les termes associés aux concepts.

Ces deux premiers exemples de réutilisation confirment, s'il était encore besoin, que les terminologies de toutes natures – jusqu'aux ontologies – étant chaque fois développées en fonction d'un but précis, leur utilisation ou réutilisation est toujours difficile.

3.5 Réutilisation de la SNOMED V3.5

La création de la branche des maladies est toujours une partie majeure de la constitution des ontologies médicales. Comme les corpus visés n'étaient pas disponibles, ou ne couvraient pas tout le domaine, nous avons décidé de compléter le travail en intégrant dans ONTOLURGENCES la branche correspondant aux diagnostics de la SNOMED V3.5 [7]. La procédure a principalement été effectuée par les médecins et a demandé plus de 100 heures de travail : la SNOMED V3.5 était notoirement trop précise – ce à quoi l'on pouvait s'attendre – mais surtout très mal organisée – ce qui nous a passablement surpris. Des 25 000 maladies répertoriées dans la SNOMED V3.5, 6 500 ont été conservées.

3.6 Autres remarques méthodologiques

Pour compléter cette description de la construction d'ONTOLURGENCES, quelques précisions supplémentaires sont nécessaires :

1. Le langage SKOS [8] a été utilisé pour y décrire les termes ratta-

7. La SNOMED V3.5 est une classification multiaxiale initialisée par des anatomopathologistes canadiens. Elle a vocation a représenter toute la médecine et les notions de la société associées. Elle comporte 105 000 concepts. Elle existe en français et a été choisie comme *terminologie de référence* (Rosenbloom *et al.*, 2006) par le gouvernement français. De cette classification a été dérivée une ontologie par intégrations successives d'autres terminologies et par réorganisation, la SNOMED-CT. Cette dernière n'est pas entièrement en français entre autres défauts rédhibitoires (Rector *et al.*, 2011).

8. Le *Simple Knowledge Organisation System* (SKOS) est développé dans le cadre du W3C depuis 2003.

chés aux concepts. Il se définit comme un langage de représentation de systèmes d'organisation de connaissances tels que thésaurus, taxonomies, ou tout autre type de vocabulaire contrôlé ou structuré. Ce standard met à disposition certaines primitives dédiées à la terminologie avec pour chaque langue, un terme préféré `skos:prefLabel`, des synonymes `skos:altLabel` et une définition `skos:definition`. Ces primitives appartenant à un standard couramment utilisé sont donc appropriées pour la représentation des noms et synonymes des concepts de l'ontologie et elles peuvent tout à fait être mobilisées au sein d'une ontologie décrite en OWL. Pour faciliter l'édition de ces labels au sein de l'éditeur Protégé, nous utilisons un *plugin* spécifique [9].

2. Les ressources utilisées pour construire l'ontologie sont diverses. Autant que faire se peut, nous mémorisons l'origine des concepts à l'aide d'une annotation qui précise l'identifiant du concept dans la ressource d'origine, *e.g.SnomedId* pour la SNOMED V3.5 ou *FmaID* pour la FMA (*Foundational Model of Anatomy*).
3. Les concepts de l'ontologie doivent être partagés selon le fait qu'ils sont utilisés pour la RI ou pas. Ces derniers sont ou des concepts structurant de haut niveau – *e.g.* ObjetIntentionnel – ou des concepts médicaux trop généraux pour être discriminants – *e.g.* ExamenClinique. Cette caractéristique est décrite dans l'ontologie par une annotation booléenne – `terminologicalConcept` – qui précise le caractère dit terminologique (si le concept est utile dans la tâche de RI) ou non du concept.

4 Enrichissement terminologique de l'ontologie

4.1 métamodéliser pour supporter l'enrichissement

L'ontologie ONTOLURGENCES met à disposition une conceptualisation du domaine des urgences avec des termes pour désigner les concepts. Cette conceptualisation peut bénéficier en particulier des termes présents dans les SOC du domaine de la santé pour augmenter la détection des concepts dans les documents traités. Pour élaborer cette nouvelle ressource, il faut être capable de représenter ces SOC et l'ontologie à un même niveau de

9. Le *plugin* ARCHONTE a été développé dans notre unité de recherche par L. Mazuel, il correspond à l'intégration d'une partie des fonctionnalités du logiciel DOE dans Protégé associée à une interface d'annotation gérant le multilinguisme et les labels SKOS.

description. C'est ce que nous permet le modèle UniMoKR décrit dans un travail précédent (Vandenbussche & Charlet, 2009).

4.2 Processus

Le système LERUDI fonctionne globalement comme suit : le texte des différents documents composant le DPI est traité par un algorithme qui va chercher à établir une correspondance (le cas échéant, en intégrant des méthodes de TAL) entre les syntagmes (traités comme de pures chaînes de caractères) et le système de concepts de la RTO. Si la chaîne de caractères a pu être appariée avec un concept, le concept servira à indexer le document (phase dite d'interprétation sémantique) (Mazuel & Sabouret, 2007). Or, les dossiers médicaux étant le plus souvent rédigés en langage naturel (ou en tout cas dans le langage semi-normalisé propre à l'activité médicale concrète), pour que l'interprétation sémantique puisse atteindre un niveau satisfaisant (voire optimal : l'optimum étant ici fixé par les performances qu'atteint un urgentiste professionnel moyen), il est bien souvent nécessaire de disposer de l'ensemble des variations lexicales que peut présenter la forme textuelle du concept (synonymes, formes abrégées, etc.). Si une forme rencontrée dans le DPI n'a pas été spécifiée dans l'ontologie, le dossier ne sera pas indexé avec le concept correspondant. Le terme n'apparaîtra pas lorsque l'urgentiste parcourra l'interface. Il devra alors s'accommoder d'une information incomplète, voire erronée [10].

Pour pallier ce problème, deux processus d'enrichissement terminologique de la RTO ont été mis en œuvre : (i) l'enrichissement automatique de la RTO par l'ajout de termes extraits de diverses SOC et (ii) l'enrichissement semi-automatique de la RTO par ajout de syntagmes nominaux apparaissant dans les documents des dossiers médicaux utilisés.

4.3 Enrichissement par alignement de SOC

La RTO enrichie est élaborée à partir de l'ontologie du domaine des urgences développée, alignée à des SOC pertinents pour ce domaine dont

10. Des auteurs comme (Mohammed & Sahroni, 2010) ont avancé l'idée qu'une des principales raisons du rejet des systèmes d'information médicaux avait trait à la qualité de l'information délivrée par ces systèmes. On le comprend bien dans un cadre tel que la prise en charge d'un patient dans un service d'urgences : si le système fournit une information erronée sur le passé médical du patient, les conséquences peuvent s'avérer dramatique. Le cas d'information incomplète est plus difficile à analyser car les médecins urgentistes ayant peu d'information à leur disposition, ils travaillent habituellement dans un contexte de manque d'information.

CIM-10, SNOMED 3.5. Les SOC constituent un vocabulaire contrôlé sur lequel s'appuient les fonctions d'analyse (annotation) des documents des patients. Cet enrichissement se déroule en 3 phases :

Alignement de l'ontologie à la SNOMED V3.5**.** Pour pouvoir aligner ONTOLURGENCES à d'autres SOC, il a été décidé de l'aligner dans un premier temps à une ontologie de référence (Rosenbloom *et al.*, 2006), la SNOMED V3.5. Ce choix à l'avantage de nous ouvrir aux travaux qui ont alignés la SNOMED V3.5 à d'autres SOC. Cet alignement a été effectué en utilisant le logiciel d'alignement ONAGUI (Mazuel & Charlet, 2010) puis en validant manuellement tous les alignements.

Création des mises en correspondance entre la SNOMED V3.5 **et les autres SOC.** Les laboratoires CISMeF et Lertim ont proposé une approche lexicale d'alignement de SOC en utilisant le Metathesaurus UMLS (Merabti *et al.*, 2010). Elle a été implémentée dans le modèle UniMoKR et est utilisée ici (*cf.* § 4.1). Les alignements effectués entre ONTOLURGENCES et les autres SOC permettent d'établir des liens d'équivalence *exactMatch* [11] entre concepts.

Enrichissement lexical des concepts. La troisième phase est l'enrichissement des concepts de l'ontologie des urgences grâce aux termes des concepts des SOC mis en correspondance. La RTO est construite automatiquement à l'export grâce à un pré-traitement qui consiste à ajouter des termes aux concepts de ONTOLURGENCES. Pour cela, nous regardons chaque concept de cette ontologie ; pour ceux qui ont une correspondance avec un autre concept d'un SOC, nous recopions le ou les termes du concept aligné sur le concept de l'ontologie. De cette manière, nous avons des concepts avec des formes lexicales variées de désignation.

4.4 Enrichissement par analyse de syntagmes nominaux

Pour tenir compte de l'extension du champ conceptuel (*cf.* § 3.1), une autre procédure d'enrichissement semi-automatique a été mise en place. Cette procédure reprend les principes de la méthodologie par analyse de corpus (*cf.* § 3.2) avec les étapes suivantes : (1) on commence par analyser, avec des outils de TAL, le contenu de documents produits en activité par les professionnels de santé, ici les DPI, de manière à extraire (en

11. Nous n'avons volontairement utilisé que les liens de stricte équivalence pour éviter d'ajouter des termes trop éloignés et donc qui amènent du bruit dans les détections futures.

mobilisant cette fois des méthodes statistiques) les syntagmes nominaux susceptibles d'être parmi les plus structurants pour le domaine de connaissances considéré ; (2) une fois ces termes identifiés, des professionnels de santé (des urgentistes) : (i) réalisent un filtrage de manière à ne retenir que les termes relevant effectivement du domaine médical, et susceptibles d'être médicalement pertinents lors de la procédure de RI dans le DPI, et (ii) valident les termes synonymes identifiés ; (3) ces termes sont alors : (i) ajoutés comme synonymes (balise `skos:altLabel`) lorsqu'ils correspondent à des concepts médicaux existant déjà dans ONTOLURGENCES ; ou (ii) convertis en concepts, lorsqu'ils réfèrent à des notions ne possédant pas encore de représentation conceptuelle dans ONTOLURGENCES, ce qui nécessite alors de positionner ce concept dans l'ontologie.

5 Procédures de validation

5.1 Pourquoi des procédures de validation ?

Après un an de travail, il est apparu que la mise en place de procédures de contrôle était nécessaire pour maintenir la qualité de l'ontologie, et que ces procédures devaient être rejouées régulièrement. En effet, a) de nombreux intervenants, médecins aussi bien que modélisateurs, travaillent de concert sur l'ontologie et, malgré nos efforts, nous n'avons pas toujours été en mesure de faire passer correctement les consignes ; de plus, b) les consignes sont nombreuses et contraignantes et une même personne peut les appliquer un jour et les oublier un autre.

Ces consignes portent principalement sur les termes associés aux concepts et sur les annotations liées à la pertinence des concepts pour l'indexation des documents.

Dans un premier temps, ces procédures de contrôle ne portent pas sur la structure de l'ontologie. La motivation étant que, à ce niveau du développement de l'ontologie et au vu des compétences de l'équipe, les problèmes rencontrés ont d'abord été des problèmes terminologiques. Mais il est évident que les problèmes de structuration, eux aussi présents, appellent des traitements futurs (*cf.* 6). Nos procédures reposent sur des patrons, ou des anti-patrons lorsqu'elles gèrent des erreurs que l'on veut éviter. Ce travail s'inscrit dans le courant du contrôle de la qualité des ontologies comme on peut le lire, sur des points plus structurels, dans (Roussey *et al.*, 2010) ou (Rector *et al.*, 2004).

5.2 Quel métamodèle ?

Dans leur principe, les procédures de contrôle mises en place visent à assurer que la RTO respecte un métamodèle précis au regard de nos critères. Ce métamodèle peut s'exprimer par la liste de règles suivantes pour la partie qui nous intéresse ici :

- tout concept possède une annotation `terminologicalConcept` en format booléen ;
- tout concept terminologique (*cf.* point précédent) possède un et un seul `skos:prefLabel` en français ;
- tout concept terminologique possède zéro ou un `skos:prefLabel` dans une autre langue ; les langues qui nous intéressent sont l'anglais, pour la communication, et le latin, largement représenté dans l'étymologie des concepts médicaux ;
- en raison du fonctionnement des algorithmes de RI, deux concepts différents ne peuvent pas avoir un même `skos:prefLabel` ou un même `skos:altLabel` (chaîne de caractères identique) ;
- le `skos:hiddenLabel`, proposé par la norme SKOS, est utilisé pour stocker la partie de l'identifiant du concept qui apparaît dans les arborescences (le frag-URI) en fonction de la langue ; celui-ci est signifiant ;

5.3 Les procédures

La mise en œuvre des procédures se fait par l'intermédiaire du chargement de l'ontologie dans un entrepôt et à travers des requêtes SPARQL.

Ainsi, les critères de qualités énoncés précédemment sont vérifiés dans le *triplestore*. S'y rajoutent principalement deux critères :

- *Repérage des classes ayant plusieurs parents.* Qu'un concept possède deux concepts parents n'est pas un problème en soi en OWL, mais cette polyparentalité peut être le symptôme d'une mauvaise modélisation. Dans notre méthodologie, l'ontologie est d'abord formée d'un arbre différentiel et les doubles héritages viennent de la mise en place de concepts définis. Il se peut que, dans une étape intermédiaire, nous mettions deux parents à un concept en attendant de parfaire la modélisation. Mais que ce double héritage soit assumé ou non, il est intéressant de le lister.
- *Requêtes supplémentaires.* Un certain nombre de requêtes sont faites pour a) normaliser les frag-URI par rapport aux minuscules/majuscules et b) pour normaliser les labels et leur enlever, autant que faire se

peut, des caractères qui empêcheraient tout appariement durant la RI (parenthèses, crochets, etc.)

5.4 L'ontologie en chiffres

A ce jour, avant les évolutions discutées en conclusion (*cf. infra*), les chiffres caractéristiques de l'ontologie sont les suivants :

	Nombre	Pourcentage
Concepts	10 610	
Concepts terminologiques	9 805	92,41%
Annotations `skos:prefLabel`	10 610	
Annotations `skos:altLabel`	8 727	
Concepts issus de la CIM-10	861	8.11%
Concepts issus de la SNOMED V3.5 VF	3514	33.11%
Concepts issus de CCAM	2561	24.13%
Concepts issus du FMA	306	2.88%

6 Conclusion et perspectives

Le projet LERUDI a été pour nous l'occasion de mettre au point un processus de travail à même de construire et d'enrichir une RTO complexe. A travers la description du processus de construction puis de validation d'une RTO au sein d'une équipe conséquente, nous avons montré : (1) la nécessité d'articuler précisément les concepts et les termes au sein d'une telle ressource ; (2) la nécessité de développer une architecture de méta-modélisation permettant de modéliser l'ensemble des SOC et structures de connaissances nécessaires ; (3) la possibilité de mettre en œuvre, pour ce faire, des procédures « industrielles » fondées sur cette architecture.

Les principales difficultés rencontrées ont été : (1) la définition du champ conceptuel de la RTO, (2) le manque de représentativité du corpus d'origine qui nous a obligé à relancer une procédure d'explicitation de concepts en partie manuelle, (3) la difficulté à réutiliser des SOC élaborés pour d'autres usages. Finalement, l'intégration des SOC sous un même format et le service de transformation RDF (capable d'effectuer des pré-traitements) permet de générer une RTO qui comporte une lexicalisation suffisante pour supporter les actions d'annotation, d'inférence et d'indexation des dossiers patients. Ce projet a démontré la possibilité de prendre en compte de multiples SOC et la capacité à mettre à disposition une ressource construite à partir de différents traitements de requêtes et de transformations.

Nos perspectives de travail sont : (1) une réorganisation complète des concepts de diagnostics basée sur une vue physiopathologique et une vue anatomique, dans le but d'améliorer la maintenabilité d'ONTOLURGENCES et les raisonnements. En parallèle, (2) un travail sur la complétude des termes avec test de l'ontologie dans l'application est en cours.

Références

BACHIMONT B., ISAAC A. & TRONCY R. (2002). Semantic Commitment for Designing Ontologies : A Proposal. In A. GOMEZ-PÉREZ & V. BENJAMINS, Eds., *13th International Conference on Knowledge Engineering and Knowledge Management (EKAW'02)*, volume 2473 of *Lecture Notes in Artificial Intelligence*, p. 114–121, Sigüenza, Espagne : Springer Verlag.

BRINGAY S., BARRY C. & CHARLET J. (2005). Les annotations pour gérer les connaissances du dossier patient. In M.-C. JAULENT, Ed., *Actes des 16es Journées Ingénierie des Connaissances*, Nice, France : Presses universitaires de Grenoble.

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

CHARLET J., BACHIMONT B., MAZUEL L., DHOMBRES F., JAULENT M. & BOUAUD J. (2012). OntoMenelas : motivation et retour d'expérience sur l'élaboration d'une ontologie noyau de la médecine. *Technique et Science Informatiques*. [in press].

CHARLET J., BACHIMONT B., MAZUEL L., DHOMBRES F., JAULENT M.-C. & BOUAUD J. (2009). Ontomenelas : motivation et retour d'expérience sur l'élaboration d'une ontologie noyau de la médecine. In *3^{e} Journées Francophones sur les Ontologies*, p. 1–11, Poitier, France : ACM.

DECLERCK G., BANEYX A., AIMÉ X. & CHARLET J. (2012). A quoi servent les ontologies fondationnelles ? In S. SZULMAN, Ed., *Actes des 23es Journées Ingénierie des Connaissances*, Paris, France. *À paraître*.

DHOMBRES F., JOUANNIC J., JAULENT M. & CHARLET J. (2010). Choix méthodologiques pour la construction d'une ontologie de domaine en médecine prénatale. In S. DESPRÉS & M. CRAMPE, Eds., *Actes des 21es Journées Ingénierie des Connaissances*, Nîmes, France : Presse des Mines.

GAYET P., CHARLET J., JOSSERAN L., MAZUEL L. & MIROUX P. (2010). Représentation de la médecine d'urgence dans le corpus des abstracts du congrès urgence. In *Actes du congrès URGENCES 2010*. Poster.

MAZUEL L. & CHARLET J. (2010). Alignment between domain ontologies and snomed : three case studies. In C. SAFRAN, H. F. MARIN & S. R. RETI, Eds., *MEDINFO 2010 - Proceedings of the 13 th World Congress on Medical and*

Health Informatics - Partnerships for effective e-Health solutions, volume 160, Cape Town, South Africa : IOS Press. Poster.

MAZUEL L. & SABOURET N. (2007). Degré de relation sémantique dans une ontologie pour la commande en langue naturelle. In F. TRICHET, Ed., *Actes des 18es Journées Ingénierie des Connaissances*, p. 73–85, Grenoble, France : Cépaduès. ISBN 978.2.85428.790.5.

MERABTI T., MASSARI P., JOUBERT M., SADOU E., LECROQ T., ABDOUNE H., RODRIGUES J. & DARMONI S. J. (2010). An automated approach to map a french terminology to UMLS. *Studies in Health Technology and Informatics*, **160**(Pt 2), 1040–1044.

MOHAMMED S. A. & SAHRONI M. N. (2010). Information quality as success determinant for health information systems. In *Proceedings of of the 2010 Regional Conference on Knowledge Integration (ICT 2010)*, p. 674–679.

RECTOR A., BRANDT S. & SCHNEIDER T. (2011). Getting the foot out of the pelvis : modeling problems affecting use of SNOMED CT hierarchies in practical applications. *J Am Med Inform Assoc*, p. 432–40.

RECTOR A., DRUMMOND N., HORRIDGE M., ROGERS J., KNUBLAUCH H., STEVENS R., WANG H. & WROE C. (2004). Owl pizzas : Practical experience of teaching owl-dl : Common errors & common patterns. In *In Proc. of EKAW 2004*, p. 63–81 : Springer.

REYMONET A. (2007). Modélisation de ressources termino-ontologiques en OWL. In F. TRICHET, Ed., *Actes des 18es Journées Ingénierie des Connaissances*, p. 169–180, Grenoble, France : Cépaduès. ISBN 978.2.85428.790.5.

RODRIGUES J.-M., TROMBERT-PAVIOT B., RECTOR A., BAUD R., CLAVEL L., ABRIAL V., IDIR H. & VERY J.-M. (1999). GALEN, il existe quelque chose après les mots : leur signification et au delà le savoir médical. *Innovation Stratégique en Information de Santé*, (2–3), 48–62.

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

ROUSSEY C., SCHARFFE F., CORCHO O. & ZAMAZAL O. (2010). Une méthode de débogage d'ontologies OWL basées sur la détection d'anti-patrons. In S. DESPRÉS & M. CRAMPE, Eds., *Actes des 21es Journées Ingénierie des Connaissances*, p. 43–54, Nîmes, France : Presse des Mines.

VAN HEIJST G., SCHREIBER A. T. & WIELINGA B. J. (1997). Using explicit ontologies in KBS development. *International Journal of Human-Computer Studies*, **45**(2/3), 183–292.

VANDENBUSSCHE P.-Y. & CHARLET J. (2009). Méta-modèle général de description de ressources terminologiques et ontologiques. In F. GANDON, Ed., *Actes des 20es Journées Ingénierie des Connaissances*, p. 193–204, Hammamet, Tunisie.