Méta-modèle général de description de ressources terminologiques et ontologiques

Pierre-Yves Vandenbussche, Jean Charlet

Méta-modèle général de description de ressources terminologiques et ontologiques

Pierre-Yves Vandenbussche[1,2], Jean Charlet[1,3]

[1] INSERM UMRS 872, éq.20, 15, rue de l'école de médecine, 75006 Paris, France

[2] MONDECA, 3, cité Nollez, 75018 Paris, France
pierre-yves.vandenbussche@etu.jussieu.fr

[3] DSI AP-HP, Paris, FRANCE
jean.charlet@spim.jussieu.fr

Résumé : L'intégration des ressources terminologiques et ontologiques d'un domaine est un enjeu majeur en vue de leur pleine exploitation par des organisations. Cette intégration est rendue difficile par l'hétérogénéité des ressources et de leur formalisme de représentation (SKOS, BS 8723, etc.). Ces formalismes se différencient principalement par leur richesse d'expressivité. Dans cet article, nous proposons un nouveau méta-modèle de représentation de terminologies et d'ontologies. Celui-ci a une double particularité. Il propose un formalisme de représentation plus général car il fait l'union de chacune des spécificités des formalismes existants tout en définissant de nouveaux constructeurs qui apportent un pouvoir d'expressivité supplémentaire aux ressources terminologiques. Il se base sur les technologies d'Ingénierie Dirigée par les Modèles, en vue de permettre une intégration automatique de ressources terminologiques provenant d'un formalisme. **Mots-clés** : méta-modèle, terminologie, ontologie, interopérabilité, opérationnalisation.

1 Introduction

Depuis l'apparition de l'informatique et du Web, les ressources terminologiques et ontologiques, des plus linguistiques aux plus formelles, ont pris une position centrale qui rend possible le partage d'informations normalisées. Ces ressources complémentaires doivent souvent être gérées de façon cohérente, les unes par rapport aux autres et dans le temps afin de répondre aux demandes croissantes des utilisateurs. Mais la diversité et le nombre des ressources rendent leur mise en œuvre difficile.

Nous proposons un méta-modèle général de description des ressources terminologiques et ontologiques. Notre approche se positionne sur un même plan que les standards et les normes de représentation spécifiques à un type de ressources mais apporte le moyen de gérer de manière générale une ou plusieurs ressources en conservant l'expressivité de chacune.

Dans la suite de notre article nous rappelons dans la section 2 les motivations de

notre recherche. Après avoir retenu certains critères d'analyse, nous étudions dans la section 3 quelques ressources existantes. De cette étude nous identifions les éléments de modélisation nécessaires à notre méta-modèle et les limites d'utilisations que nous résolvons en partie dans notre méta-modèle. Dans la section 4, nous mettons en exergue à partir des normes et des standards de représentation adaptés à des types de ressources particuliers, les modélisation utiles pour notre méta-modèle que nous présentons dans la section 5. Les discussions et conclusions apportent pour finir des réflexions pour la suite de ce travail.

2 Le contexte de recherche

Depuis plus d'une dizaine d'années, la discipline de l'ingénierie de la connaissance travaille à l'élaboration de ressources terminologiques et ontologiques et à la mise à disposition de connaissances. Notre approche de manière générale et les travaux exposés dans cet article en particulier se situent au cœur de ces enjeux en proposant d'une part, le moyen de représenter et de faire coexister l'ensemble de ces ressources et d'autre part, des services satisfaisant les besoins des utilisateurs grâce à leur modélisation. Dans cette section nous revenons sur certaines définitions des ressources auxquelles nous sommes confrontés puis nous étudions quelles attentes nous avons de ces ressources.

2.1 La diversité des ressources

Il existe tout un panel de systèmes de structuration pour représenter les connaissances : liste contrôlée, classification, thésaurus, terminologie, ontologie... Commençons par replacer l'éventail des ressources auxquelles nous sommes confrontés et dont nous empruntons certaines définitions aux écrits existants sur le domaine.

Une *classification* est « la répartition systématique en classes, en catégories d'êtres, de choses ou de notions ayant des caractères communs notamment afin d'en faciliter l'étude ; c'est aussi le résultat de cette opération » (Bourigault *et al.*, 2004). Notons que les classes peuvent être organisées entre elles hiérarchiquement selon un principe générique-spécifique. Un exemple de classification est la CIM-10 [1].

Un *thésaurus* est un ensemble structuré de termes nécessaires à son utilisation au sein d'une hiérarchie de concepts liés par des relations sémantiques. Eurovoc (cf. 3.2.1) et le MeSH (cf. 3.2.2) sont des thésaurus.

Une *terminologie* « est une liste de termes d'un domaine ou sujet donné représentant les concepts ou notions les plus fréquemment utilisés ou les plus caractéristiques, cette liste étant ou non structurée » (Lefèvre, 2000). Contrairement à un thésaurus, dans une terminologie, l'accent est mis sur l'exhaustivité des termes (synonymes, abréviations...). Un exemple de terminologie est la SNOMED 3.5.

Les attendus des *ontologies* ont beaucoup changé depuis le début de leur utilisation en informatique dans les années 1990. En effet leurs objectifs sont devenus plus modestes et plus réalistes. On peut les définir comme un ensemble de concepts et de relations

1. 10e édition de la classification internationale des maladies. Voir : `http://taurus.unine.ch/icd10/`

pour une utilisation particulière d'un domaine déterminé ; cette structure repose sur une formalisation avouée afin d'effectuer des inférences dans un système informatique. Un exemple d'ontologie est la SNOMED-CT (cf. 3.2.3).

Dans la suite de cet article, nous désignons par *Ressource Terminologique et Ontologique* (RTO) l'ensemble de ces artefacts (Bourigault *et al.*, 2004).

Face à ce pluralisme, un constat peut être fait : la complexité au sein de chaque structure de ressources rend difficile leur catégorisation. Il existe ainsi des terminologies plus ou moins formelles, des thésaurus avec une structure plus ou moins complexe, des ontologies avec ou sans contraintes sur leurs relations... Le classement d'une ressource dans un type particulier de structuration est une tâche ardue en témoignent les critiques de certaines ressources qui prétendent être ce qu'elles ne sont pas.

2.2 Les attentes d'utilisation

Comme nous venons de le voir, il existe une diversité dans la représentation et l'organisation des connaissances qui s'explique par des utilisations différentes. Néanmoins ces structurations ont toutes pour vocation d'appréhender de l'information, de la partager et de permettre un traitement humain et computationnel. Identifions les souhaits d'utilisation de ces ressources :

2.2.1 La recherche d'information

Ce premier point est le plus critique : le but même d'une RTO est de faciliter l'accès à des informations normalisées et plus ou moins formalisées. La recherche d'une information par une personne peut se faire par trois approches différentes : par recherche textuelle ; par recherche arborescente ; par recherche sur le réseau.

- La recherche textuelle est fortement dépendante de la **richesse linguistique** de la ressource. La prise en compte de l'aspect terminologique et la finesse de l'expressivité linguistique vont être des facteurs importants. Par exemple, la représentation des relations de synonymie et de méronymie entre termes accroît considérablement les recherches relatives à un terme ou concept donné.
- La recherche arborescente est liée à la **structuration** même de la RTO. L'organisation d'une terminologie sous forme arborescente (par des relations génériques spécifiques ou partitives) devient nécessaire quand le **volume** est trop important et qu'une simple liste ne suffit plus. Si nous prenons l'exemple de la SNOMED-CT (cf. 3.2.3), on peut entrer dans la terminologie par 19 hiérarchies de haut niveau (p. ex. « Clinical finding/disorder » ou encore « Body structure »). Ces arborescences posent toutefois des problèmes, tels que l'adhésion aux principes de structuration sur lesquels nous reviendrons (cf. 3.3).
- La recherche sur le réseau de connaissances repose sur les **relations sémantiques définies** dans ce réseau. Il s'agit d'obtenir l'ensemble des éléments ou parties du réseau sémantique qui vérifient une requête donnée sur le système. On pourrait ainsi rechercher dans une ontologie de médecine « Toutes les maladies ayant pour localisation l'abdomen ». Cette recherche implique pour être possible, que cette ontologie possède par exemple : les entités *Maladie*, *Abdomen* ; une relation *localisation* d'une maladie dans une région du corps.

2.2.2 L'interopérabilité

Avec les développements du Web et la mondialisation, le partage d'information est au centre des problématiques actuelles. Dans le milieu médical par exemple, les attentes tendent vers des systèmes d'information partagés entre les services et entre les établissements de santé. Parallèlement, les pratiques et les utilisations des ressources ont évolué et il devient caduc de penser qu'il existe *une* terminologie qui capture l'ensemble des connaissances d'un domaine (Aussenac-Gilles, 2005). Ce besoin d'échange d'informations et la prise en compte de plusieurs ressources traduisent la notion d'interopérabilité que P. Miller définit par : « process of ensuring that the systems, procedures and culture of an organisation are managed in such a way as to maximise opportunities for exchange and re-use of information, whether internally or externally. » (Miller, 2000).

Considérons l'interopérabilité sous 3 niveaux (Ferreira da Silva *et al.*, 2006) :

- L'interopérabilité syntaxique constitue le premier niveau et concerne le format de représentation des connaissances. Le fait qu'une ressource soit exprimée ou puisse être traduite sous un format **standardisé ou normé** conduit vers l'interopérabilité syntaxique.
- L'interopérabilité structurelle, à un second niveau, fait référence à l'organisation de l'information au sein d'une ressource qui est régie par son **modèle** sous-jacent.
- L'interopérabilité sémantique est l'interprétation que l'on a d'une représentation d'un domaine fondée sur un consensus. La sémantique est donnée par les **symboles** que l'on a définis sur les primitives du modèle.

La sémantique se définit sur les primitives du modèle exprimées dans une syntaxe donnée. Ces trois niveaux sont donc interdépendants.

3 Expressivité et utilisation des RTOs : trois études de cas

Des attentes d'utilisation que nous venons d'énoncer, nous pouvons maintenant extraire des indicateurs pour juger de l'expressivité et de l'utilisation des RTOs qui seront repris dans une synthèse :

3.1 L'identification de critères d'analyse

- **Le périmètre du domaine décrit** : la ressource doit avoir une définition claire de ses prétentions. Tout d'abord, la finalité de la ressource doit être connue, c'est-à-dire l'application pour laquelle elle a été construite. Si une ressource a un usage donné, alors elle décrit un domaine particulier avec une granularité de l'information représentée.
- **La volumétrie** : dépendant de son périmètre et de la granularité voulue, le volume d'une RTO peut fortement varier.
- **La richesse linguistique** : la linguistique est un point d'entrée pour la recherche par une personne. L'expressivité de la linguistique va dépendre des primitives dé-

finies dans le modèle telles que la gestion de la langue sur un terme, les relations de synonymie et de traduction...
- **L'expressivité formelle** : le caractère formel sous forme d'une logique mathématique permet d'opérer des traitements automatiques sur une ressource, par exemple grâce à la définition formelle d'une relation de subsomption.
- **La conformité aux normes, aux standards et aux recommandations** : l'utilisation de standards pour la construction d'une ressource ou pour l'échange de cette ressource favorise l'interopérabilité.

3.2 Étude de trois RTOs

Pour comprendre la diversité d'expressivité des RTOs actuelles, nous avons étudié les ressources suivantes en nous basant sur les critères que nous venons d'énoncer : Eurovoc, SNOMED-CT, MeSH.

3.2.1 Eurovoc

Eurovoc « est un thésaurus multilingue (21 langues) couvrant tous les domaines de l'activité de l'Union Européenne, Il permet d'indexer les documents dans les systèmes documentaires des institutions Européennes et de leurs utilisateurs »[2]. La construction de ce thésaurus est conforme aux normes ISO 2788-1986 et ISO 5964-1985 et se compose de *descripteurs* ou termes préférentiels, de *non-descripteurs* ou termes non-préférentiels organisés au sein d'une classification hiérarchique à deux niveaux (domaines et microthésaurus). Les relations sémantiques utilisées sont au nombre de quatre : relation d'appartenance au microthésaurus (MT) ; relation d'équivalence de synonymie entre un terme préférentiel et un terme non-préférentiel (UF,USE pour Used For et Use) ; relation hiérarchique (BT,NT pour Broader Term et Narrower Term) ; relation associative (RT pour Related Term). En terme de volumétrie, il existe 6645 concepts reflétés par autant de termes préférés et 10 000 relations dans chaque langue pour plus de 260 000 termes (préférés ou non) toutes langues confondues en 2007.

Le périmètre d'Eurovoc a été établi : le but est de répondre aux besoins de systèmes documentaires généraux sur les activités de l'Union Européenne ; il ne convient toutefois pas à l'indexation et à la recherche de documents spécialisés. Le thésaurus indique qu'il ne peut pas prétendre couvrir les différentes réalités nationales à un niveau suffisamment spécifique (p.ex. existence en Belgique du *Conseil supérieur de la Justice*). Il est toutefois possible au travers d'une fiche de maintenance, de spécifier un nouveau besoin.

3.2.2 MeSH

Le MeSH (Medical Subject Heading) est le thésaurus de référence dans le domaine biomédical produit par le NLM (U.S. National Library of Medicine)[3]. La structure du MeSH est à trois niveaux : un ensemble de termes dont un préférentiel, désigne

2. Voir `http://europa.eu/eurovoc/`
3. Voir `http://ist.inserm.fr:3201/inserm08/index.html`

un *concept* qui fait partie d'une classe de concepts appelée *descripteur*. La navigation dans le thésaurus se fait par recherche ou en entrant par les *Main Headings*. Le MeSH contient près de 25 000 descripteurs et plus de 455 000 termes en janvier 2009. Trois types de relations, ne reposant pas sur la logique formelle, sont utilisées : hiérarchique, synonymique et de proximité sémantique.

Produit depuis 1960, le MeSH est utilisé pour l'indexation par de nombreuses bibliothèques et institutions à travers le monde. le thésaurus a été traduit dans de nombreuses langues mais il est à déplorer qu'il n'existe aucune possibilité de navigation à travers le multilinguisme [4]. Le MeSH n'est conforme à aucune norme, il permet seulement des échanges via le format XML.

3.2.3 SNOMED-CT

La SNOMED-CT (Systematized Nomenclature of MEDicine-Clinical Terms) est une ontologie multilingue (pour l'instant en deux langues et un dialecte) de la santé clinique [5]. Il s'agit d'une structure hiérarchique de concepts désignés par des *descriptions* (termes) sur plus de 31 niveaux de subsomption. la SNOMED-CT est conforme au modèle HL7 version 3 et repose sur une sémantique formelle (Logique de description). Elle contient plus de 311 000 concepts, près de 800 000 termes et 1 360 000 relations en janvier 2008.

Cette terminologie a pour vocation d'être utilisée pour tous documents cliniques allant du dossier patient électronique et des systèmes informatiques des hôpitaux jusqu'à la télé-médecine.

3.3 Synthèse des RTOs étudiées

RTO / Caractéristiques	Eurovoc	MeSH	SNOMED-CT
Périmètre	L'activité de l'UE. Couverture assez large. Faible granularité (7 niveaux de prof.)	Le domaine biomédicale. Couverture assez large. Granularité normale (11 niveaux de prof.)	La santé clinique. Couverture très large. Granularité très fine (31 niveaux de prof.)
Volumétrie	6645 concepts 260 000 termes 210 000 relations	25 000 concepts 455 000 termes	311 000 concepts 800 000 termes 1 360 000 relations
Richesse linguistique	synonymie, associative.	synonymie, associative	synonymie
expressivité formelle	relation hiérarchique	relation hiérarchique	Les concepts et la relation IS_A définis par la logique de description
conformité aux normes	ISO 2788-1986 ISO 5964-1985		HL7 version 3

4. Seul l'INSERM ayant la responsabilité de sa traduction en Français met à disposition une version bilingue Anglais-Français.

5. Voir `http://www.ihtsdo.org/snomed-ct/`

De l'étude de ces RTOs (dont certains résultats ont été présentés dans la partie précédente), on peut dégager certains enseignements :

- On peut considérer qu'un volume trop important peut nuire à l'utilisation effective d'une ressource par une personne. Si nous prenons la SNOMED-CT, la quantité d'information à disposition est très importante en raison d'un périmètre très large et une granularité très fine. La simple recherche arborescente d'un concept dans cette terminologie est d'autant plus ardue. En pratique, l'utilisation qui pourrait en être faite par exemple au sein d'un service hospitalier pour coder un acte, se restreindrait à un sous-ensemble défini des concepts de la RTO de référence. *Comment peut-on permettre l'utilisation d'un sous-ensemble d'une RTO de référence ?*
- Les besoins en recherche textuelle imposent une représentation minimale de primitives linguistiques, mettant à minima en avant un terme préférentiel lié à plusieurs termes synonymes comme c'est le cas dans les RTOs étudiées. La représentation de synonymes permet d'élargir le nombre de réponses à une recherche textuelle, les termes non préférentiels amenant le résultat sur le terme préférentiel. Néanmoins d'autres relations existent (p.ex. *RelatedTerm* dans Eurovoc) et permettent de présenter plus d'information à l'utilisateur. *Quelle expressivité linguistique est utile pour l'exploitation ?*
- L'un des premiers rôles d'une terminologie est de réduire l'ambiguïté. Le sens donné à un concept vient aussi bien de sa position dans le réseau de connaissances que de la définition logique qu'on lui a donnée. Seule la SNOMED-CT décrit de manière formelle ses concepts.
- L'interopérabilité syntaxique et structurelle, quant à elle, a besoin de standardisation. Nous le voyons dans notre étude : mis à part le MeSH qui ne repose sur aucune norme, chaque ressource est conforme à des normes différentes même si ces ressources sont complémentaires. Les enjeux autour de cette interopérabilité inter-référentiels sont multiples : Différencier une RTO d'interface et une RTO de référence (Rosenbloom *et al.*, 2006) ; permettre la communication entre deux utilisateurs faisant référence à deux RTOs différentes... *Quels sont les pré-requis à une interconnexion entre des référentiels différents et comment dès lors garantir l'interopérabilité ?*
- En plus des points soulevés par cette analyse, une autre question nous semble importante : celle de l'adhésion aux principes de modélisation. Toute élaboration d'une ressource terminologique ou ontologique repose d'une part sur une méthode de conceptualisation propre au modélisateur de cette ressource (ou à un consensus) et d'autre part sur la finalité dans laquelle la ressource s'inscrit. *Comment rendre la recherche d'un concept plus efficace en ne connaissant pas à priori le mode de catégorisation utilisé dans l'élaboration de cette RTO ?*

4 Les standards et normes de terminologies et d'ontologies

Dans le domaine des RTOs certaines normes existent et facilitent ainsi l'interopérabilité (cf. 2.2.2). Dans cette section nous présentons les normes et les standards principaux plus ou moins spécialisés pour un type de ressources particulier. Notre méta-modèle s'inspire de la modélisation de ces normes avec lesquelles il doit être conforme (plus général) pour prétendre ne pas réduire l'interopérabilité.

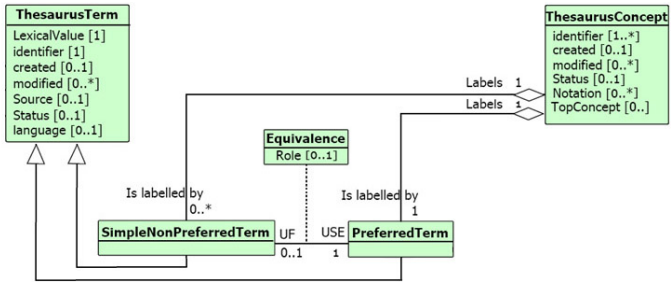

FIGURE 1 – BS8723 : partie du modèle centrée sur la terminologie.

4.1 SKOS [6]

SKOS se définit comme un langage qui permet la 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 linguistique : on a d'une part le *Concept* qui représente une notion et d'autre part la terminologie avec pour chaque langue, un terme préféré *prefLabel*, des synonymes *altLabel* et des chaînes de caractères servant à la recherche (p.ex. un code d'un concept) *hiddenLabel*. SKOS est une famille de langages extensibles. L'extension XL (pour eXtended Labels) considère les *Label*s comme des ressources, ce qui pourrait nous permettre de redéfinir des relations (p.ex. une relation de traduction) entre ces *Labels* pour augmenter la richesse linguistique. SKOS définit certaines relations comme transitives, ce qui permet d'effectuer des inférences logiques.

4.2 BS 8723 [7]

Les normes ISO concernant les terminologies sont en train d'évoluer [8] grâce aux travaux de la British Standard et de son projet BS 8723. La gestion de la linguistique est ici plus fine que dans le standard SKOS. Comme nous le montre la figure 1, un terme préféré (terme qui est utilisé dans une langue pour désigner un concept) est ici exprimé sous la forme d'une primitive, de même que le terme non préféré. Ceci leur permet d'exprimer une relation de synonymie directe entre ces élément du modèle.

4.3 OWL

Le standard OWL permet d'exprimer une connaissance en utilisant une sémantique formelle basée sur la logique des prédicats. A cet égard, OWL convient parfaitement à l'expressivité formelle nécessaire à l'interopérabilité et sert souvent de format de définition de méta-modèle dans la représentation de connaissances. Son formalisme et son expressivité permettent de représenter des ontologies. Cependant, son expressivité et son caractère généraliste sont mal adaptés à

6. Simple Knowledge Organisation System (SKOS) développé dans le cadre du W3C depuis 2003. Voir : `http://www.w3.org/2004/02/skos/`

7. Voir : `http://schemas.bs8723.org`

8. le projet ISO 25964 va remplacer les normes ISO 2788 concernant l'élaboration et le développement de thésaurus monolingue et ISO 5964 concernant l'élaboration et le développement de thésaurus multilingue en se basant sur les travaux de la BS8723.

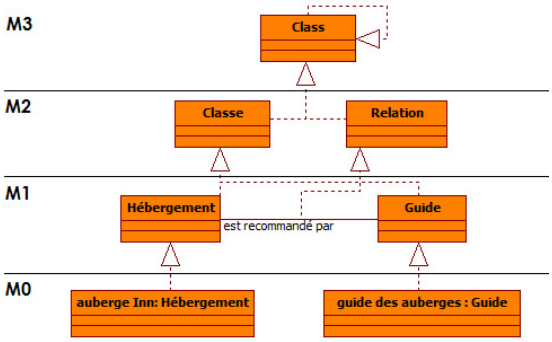

FIGURE 2 – Exemple de méta-modélisation en 4 couches

la description de terminologies ou de thésaurii contrairement au standard SKOS ou à la norme BS 8723. OWL a toutefois les éléments pour décrire ces normes mais ne permet pas nativement de représenter des terminologies.

5 Présentation d'un méta-modèle général de RTO

Certaines normes et standards existent pour décrire un type particulier de RTO. Cependant l'expressivité de leurs primitives ne permet pas de représenter l'ensemble des problèmes conceptuels posés par l'ensemble des ressources terminologiques et ontologiques (El Hachani, 2005). Les motivations qui nous poussent à définir un nouveau méta-modèle général sont doubles : supporter les modèles décrivant les RTOs existantes (p.ex. SKOS, BS 8723 etc.) et proposer grâce à la modélisation une opérationnalisation effective des ressources. L'objectif pour nous est d'abord de définir un méta-modèle en prenant en considération les meilleures pratiques dans les modèles et normes existantes, ensuite d'améliorer cette modélisation par des apports qui rendent possible une meilleure utilisation des ressources. D'autres projets se basent sur une démarche similaire, nous citerons le projet DAFOE. Dans ce projet, un méta-modèle permet également de représenter les RTOs mais à des étapes différentes partant des corpus et menant à l'élaboration d'une ontologie (Charlet *et al.*, 2008). Le méta-modèle du projet DAFOE a pour but d'appréhender ces étapes successives et de faciliter ce processus menant à la formalisation de la connaissance. Notre approche se différencie de par son objectif de mise en relation de ces ressources sur un même plan.

5.1 Méta-modélisation

Face au besoin croissant d'intégration logicielle et d'interopérabilité des systèmes d'information, l'ingénierie logicielle a mis en place depuis 2000, l'Ingénierie Dirigée par les Modèles (IDM) (Bernstein, 2003). Toute base de connaissance afin d'être opérationnalisée est représentée à l'aide d'un modèle exprimé dans un langage particulier. Ceci définit l'activité de méta-modélisation, c'est-à-dire suggérer un formalisme de modélisation, chaque modèle se conformant à un méta-modèle prédéfini. De même, à un niveau d'abstraction supérieur, le méta-modèle a besoin d'être clairement défini par un méta-méta-modèle. Afin d'éviter une décomposition infinie de niveaux d'abstraction, un patron d'architecture en 4 couches, illustré à la figure 2, sert

maintenant de référence. Le modèle M0 étant le monde réel que nous conceptualisons, il serait dans notre exemple sur le domaine de l'hôtellerie : *auberge Inn*, le modèle M1 contiendrait la classe *Hébergement*. Le modèle M2 définirait l'ensemble des éléments pour exprimer le niveau M1 c'est à dire la classe *Classe*, la classe *Relation*. un dernier niveau M3 réflexif (qui s'auto-définit) contiendrait une classe *Class* qui permet de définir les deux classes du niveau 2. Pour bien comprendre cette architecture nous vous renvoyons sur la spécification du MOF [9].

5.2 L'expressivité linguistique

Comme nous avons pu le voir, l'expressivité linguistique est très importante pour l'utilisation finale et pour la maintenance d'une ressource. La séparation entre un niveau terminologique et un niveau conceptuel permet notamment de maintenir la partie terminologique indépendamment des problèmes conceptuels comme l'expose Reymonet *et al.* (2007). C'est dans ce sens que la norme BS 8723 modélise un terme par une classe *Terme* séparée de la modélisation des concepts. Mais de quelles relations et de quelles propriétés avons-nous besoin ? D'après notre expérience, une représentation exhaustive de tous les cas possibles n'est pas rationnelle. La plupart des cas ne nécessitent qu'une représentation simple, ainsi notre méta-modèle doit permettre nativement de représenter les relations et attributs les plus communément utilisés tout en permettant de redéfinir plus précisément certains éléments, ce qui permet aux modèles d'évoluer (cf. figure3).

Ces apports linguistiques permettent de résoudre certaines problématiques précédemment identifiées dans notre synthèse (cf. 3.3) : le méta-modèle distinguant la gestion linguistique et conceptuelle, la maintenance de la RTO s'en voit simplifiée ; le lien d'un terme vers son texte source est capturé par le méta-modèle assurant ainsi une traçabilité très utile particulièrement lors de la construction ou l'enrichissement de connaissances à partir de corpus : la souplesse d'un méta-modèle extensible autorisant l'ajout par exemple de la fréquence d'un terme dans le corpus d'origine.

5.3 Sous-ensemble d'une ressource de référence

La structuration au sein d'un système de connaissances a toujours eu pour but d'organiser les connaissances et de faciliter la recherche d'information. Dès lors que la taille d'une telle ressource est trop importante il devient difficile de retrouver une information. Face à cette problématique soulevée dans notre synthèse, certaines ressources ont adopté des mécanismes de représentation spécifiques. Par exemple : le thésaurus GEMET [10] a mis en place une navigation thématique orthogonale à la hiérarchie verticale habituelle ; la SNOMED-CT a introduit la notion de *RefSet*. C'est un groupement de références de concepts spécialisés pour six utilisations différentes allant d'une simple liste (index) à plat en passant par un groupement par langue jusqu'à une hiérarchie de navigation (taxinomie).

Certains langages permettent la définition de collections [11] n'offrant toutefois pas autant de possibilités d'utilisation que nos deux exemples présentés ci-dessus.

En se basant sur ces fonctionnements, nous avons défini une entité dans notre méta-modèle, nommée ***Concept Group***, qui représente un groupement de concepts. La finalité de cette primitive est de pouvoir définir un sous-ensemble de concepts d'une RTO de référence. Ce sous-ensemble

9. MOF (Meta Object Facility) est un modèle de niveau M3 réflexif, il définit la grammaire d'UML (Unified Modeling Language) au niveau M2. `http://www.omg.org/mof/`

10. GEneral Multilingual Environmental Thesaurus. Voir : http ://www.eionet.europa.eu/gemet

11. Il existe deux primitives RDF permettant de manipuler une collection de concepts : *RDFList* et *RDFSContainer*. SKOS possède la notion de *ConceptScheme* (collection de concepts avec ou sans relation) dont la définition est volontairement permissive.

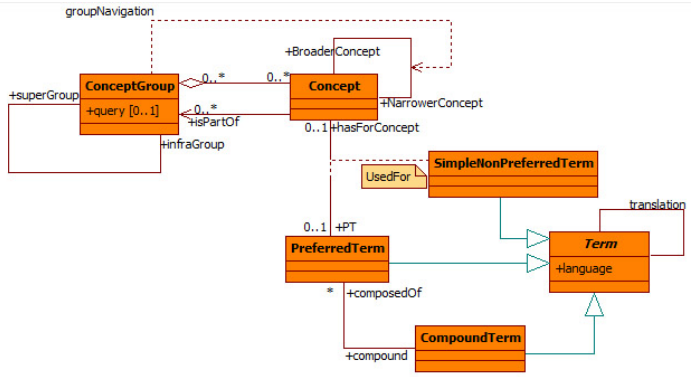

FIGURE 3 – Présentation UML simplifiée d'une partie de notre méta-modèle concernant la linguistique et les groupements de concepts. Cette figure montre l'utilisation de sous-classes de *Term* pour capturer les notions de termes préférés et non préférés reliées à un concept. La notion de *ConceptGroup* faisant référence à un ensemble de concepts définis en intension ou en extension, peut être organisée hiérarchiquement.

correspondant à une utilisation ou à une vue sur la ressource, aurait la possibilité d'être réutilisé ou partagé (pour cela, notre groupement doit avoir un identifiant unique). Nous distinguons deux définitions des concepts dans un *Concept Group*. Premièrement, nous avons ceux définis par **intension** : ensemble des concepts vérifiant la requête d'appartenance au groupement. Deuxièmement, ceux définis par **extension** : ensemble des concepts pointant sur le groupement. Les *Concept Group*s sont également hiérarchisés par une relation.

Ce travail que nous sommes en train de mettre en œuvre a été soumis au groupe de recherche pour l'élaboration de la nouvelle norme ISO 25964. Ils ont ajouté cette primitive mais sans la définition d'appartenance par intension. La définition de la primitive *Concept Group* dans notre méta-modèle répond à la problématique d'utilisation d'un sous-ensemble d'une RTO. Nous avons ainsi introduit par cet élément, de nouveaux points d'entrée dans la RTO : à la recherche arborescente vient s'ajouter une recherche orthogonale par groupement.

6 Discussions et conclusions

L'opérationnalisation de ressources terminologiques et ontologiques gérées de façon cohérente est un enjeu majeur dans l'ingénierie de la connaissance. Les apports d'un méta-modèle général permettent de faciliter l'interopérabilité au sein d'une même application ou entre plusieurs systèmes d'information et d'améliorer l'utilisation de ces ressources. Sur ce dernier point, le choix des primitives définies dans le méta-modèle va déterminer l'exploitation directe des ressources qu'il sera possible de faire. Que ce soit en linguistique ou pour les groupements de concepts, il est possible grâce à l'expressivité d'un méta-modèle, d'améliorer l'utilisation, le partage et la collaboration autour de ces ressources.

Les apports théoriques présentés dans cet article ont déjà été la source d'enrichissements de normes (cf. 5.3). Leurs mises en œuvre au sein de notre outil ITM nous donne l'assurance des résultats d'une telle approche. La description de notre méta-modèle avec une logique mathématique enrichira les traitements automatiques faits par des ordinateurs. Notre modèle ne couvre

pas à ce stade tous les besoins des ressources terminologiques et ontologiques. La gestion dans le temps des versions n'est pas abordée de même qu'une utilisation plus fine des mappings ou projections entre plusieurs ressources. Ces réflexions guideront nos recherches futures.

Références

AUSSENAC-GILLES N. (2005). *Méthodes ascendantes pour l'ingénierie des connaissances* . Habilitation à diriger des recherches, Université Paul Sabatier, Toulouse, France.

BERNSTEIN P. A. (2003). Applying model management to classical meta data problems. In *CIDR*.

BOURIGAULT D., AUSSENAC-GILLES N. & CHARLET J. (2004). Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas. *Revue d'Intelligence Artificielle*, **18**(4), 24.

CHARLET J., SZULMAN S., PIERRA G., NADAH N., TEGUIAK H. V., AUSSENAC-GILLES N. & NAZARENKO A. (2008). Dafoe : A multimodel and multimethod platform for building domain ontologies. In D. BENSLIMANE, Ed., *2^e Journées Francophones sur les Ontologies*, Lyon, France : ACM.

EL HACHANI M. (2005). *Indexation des documents multilingues d'actualités incluant l'arabe : équivalence interlangues et gestion des connaissances chez les indexeurs*. Thèse de doctorat nr en sciences de l'information et de la communication, Université Lumière Lyon.

FERREIRA DA SILVA C., MÉDINI L., GHAFOUR S. A., HOFFMANN P. & GHODOUS P. (2006). Semantic interoperability of heterogeneous semantic resources. *Electronic Notes in Theoretical Computer Science*, **150**(2), 71–85.

LEFÈVRE P. (2000). *La recherche d'informations (du texte intégral au thèsaurus)*. Hermès Science Publications.

MILLER P. (2000). Interoperability : What is it and why should i want it ? *Ariadne*, **24**.

REYMONET A., THOMAS J. & AUSSENAC-GILLES N. (2007). Modélisation de ressources termino-ontologiques en owl. In F. TRICHET, Ed., *Journées Francophones d'Ingénierie des Connaissances (IC)*, p. 169–180, http ://www.cepadues.com/ : Cépaduès Editions.

ROSENBLOOM S., MILLER R., JOHNSON K., ELKIN P. & BROWN S. (2006). Interface terminologies : Facilitating direct entry of clinical data into electronic health record systems. *Journal of the American Medical Informatics Association*, **13**(3), 277–288.