Dommages collatéraux de la fusion de terminologies

Natalia Grabar
STL, CNRS UMR 8163
Université Lille 3, France
natalia.grabar@univ-lille3.fr

Marie Dupuch
CRC, Université Paris 6
INSERM U872
Paris, France
marie.dupuch@crc.jussieu.fr

Fleur Mougin
LESIM, INSERM U897, ISPED
University Bordeaux Segalen, France
fleur.mougin@isped.u-bordeaux2.fr

Abstract

Grâce à la recherche en acquisition et structuration de termes, il existe un nombre croissant de terminologies. Ces terminologies couvrent souvent des aspects complémentaires de domaines scientifiques et techniques. Leur fusion peut être utile pour décrire et modéliser ces domaines de manière plus exhaustive. Par exemple, dans le domaine biomédical, la ressource UMLS regroupe plus de 150 terminologies. Cependant, des dommages collatéraux peuvent résulter de cette fusion. Sur des exemples concrets, nous analysons des situations d'inconsistance et d'incompatibilité des relations et proposons des méthodes simples pour les détecter. Il apparaît que jusqu'à 25 % de relations de synonymie et hiérarchiques sont inconsistantes dans UMLS.

1 Introduction

Grâce aux méthodes de traitement automatique des langues et d'intelligence artificielle proposées pour l'acquisition et la structuration de termes, un nombre croissant de terminologies existe. Comme elles couvrent souvent des aspects complémentaires des domaines scientifiques et techniques, leur fusion peut être utile et nécessaire pour décrire et modéliser ces domaines de manière plus exhaustive. Les aspects liés à la fusion de terminologies sont aussi importants dans d'autres contextes, où des informations nouvelles sont ajoutées, comme la maintenance et la mise à jour (Qi et al., 2008), l'évolution (Klein and Fensel, 2001; Maedche et al., 2002), le transcodage ou encore l'alignement (Fridman Noy and Musen, 2000; D'aquin et al., 2009). À titre d'exemple, dans le domaine biomédical, il existe la ressource UMLS®, ou Unified Medical Language System® (Lindberg et al., 1993), développée et maintenue aux États-Unis par la NLM (National Library of Medicine), qui intègre à ce jour plus de 150 terminologies biomédicales. UMLS a été le premier à avoir proposé une fusion de terminologies biomédicales et à en faire une structure commune. UMLS est diffusé sous forme d'une base de données relationnelles. Le résultat de cette fusion est très utilisé dans le domaine, et ceci dans différents contextes applicatifs (recherche et extraction d'information, codage des dossiers médicaux, systèmes de questions/réponses, Web sémantique, recherche d'images ...), où la complétude des données terminologiques est recherchée.

Si les avantages de la fusion de terminologies sont indéniables, cette fusion peut aussi apporter des inconsistances. Déjà à l'échelle d'une même terminologie, il n'est pas rare d'en trouver (Campbell et al., 1998; Smith et al., 2003; Wei et al., 2009). La situation devient encore plus complexe lorsqu'une fusion de terminologies est effectuée. En effet, les différentes terminologies sont souvent créées avec des objectifs et des principes fondateurs différents. Cela veut dire entre autres que la fusion de ces terminologies peut mener à la génération d'autres inconsistances, d'incomplétudes et de contradictions. Pour détecter, corriger et outrepasser ces défauts, des méthodes d'audit de ressources sémantiques sont proposées. On peut citer en particulier les travaux sur UMLS (Cimino, 1998; Bodenreider, 2001; Bodenreider, 2003) et WordNet (Smrz, 2004; Liu et al., 2004). Dans une revue récente des méthodes d'audit d'UMLS (Zhu et al., 2009), les auteurs distinguent trois aspects

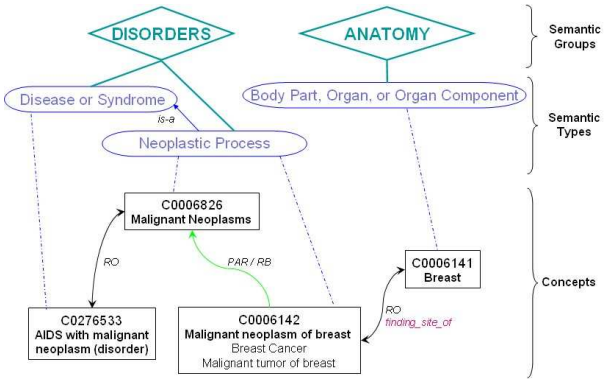

Figure 1: Les deux niveaux d'UMLS : Metathesaurus avec les concepts et les relations, et Semantic Network avec les types sémantiques. Les groupes sémantiques correspondent aux sous-domaines de la biomédicine.

liés au contenu des terminologies et à leur qualité : les termes, la classification sémantique et les relations. Quant aux indicateurs de qualité ciblés, ils sont de cinq types : orientation des concepts, consistance, non-redondance, correction et complétude de la couverture. Les travaux existants se focalisent essentiellement sur l'analyse des termes et de la classification sémantique. Parfois, des méthodes pour optimiser l'utilisation des terminologies, dont UMLS (Mougin and Bodenreider, 2005), sont aussi proposées. Dans ce travail, nous proposons d'étudier la consistance et la compatibilité des relations dans UMLS, et de dégager des règles génériques pour le contrôle de leur qualité.

2 Matériel : UMLS

UMLS est un système incluant deux sources d'information sémantique : le Metathesaurus et le Semantic Network. Le Metathesaurus®, fusionnant plus de 150 terminologies source, constitue un large graphe qui contient plus de deux millions de concepts (ensemble de termes synonymes issus des terminologies source) et plus de 50 millions de liens (également hérités des terminologies source). La quantité importante de liens présents dans le Metathesaurus est due au fait que, par convention, l'ensemble des relations existant dans les terminologies source doit être intégré dans UMLS, pour éviter la perte d'information spécifique à une terminologie source donnée dans son contexte propre. Par ailleurs, certaines relations sont ajoutées par l'équipe de la NLM. Le Semantic Network est un réseau beaucoup plus restreint de 133 types sémantiques *TS* organisés de manière arborescente. Les types sémantiques ont été agrégés en 15 groupes sémantiques ***GS*** (Bodenreider and McCray, 2003). Chaque concept du Metathesaurus est catégorisé par au moins un type sémantique du Semantic Network.

UMLS propose des informations à deux niveaux différents : les utilisateurs peuvent accéder au résultat de la fusion mais aussi aux informations spécifiques à chaque terminologie. Ceci concerne entre autres les relations. Ainsi, au niveau d'UMLS, toute relation est assignée dans une des 11 catégories de relations suivantes : *CHD* (a pour enfant), *PAR* (a pour parent) ; *RB* (a une relation plus large), *RN* (a une relation plus étroite) ; *SIB* (a pour frère) ; *RO* (a une relation autre que de synonymie, plus étroite ou plus large) ; *SY* (synonymie de la terminologie source) ; *RQ* (relation probablement synonyme), *RL* (a une relation de proximité) ; *AQ* (qualifieur autorisé), *QB* (peut être qualifié par). De ces 11 catégories, nous pouvons dégager trois grandes classes : relations hiérarchiques (*PAR, RB, CHD, RN, SIB*), synonymie (SY, RQ, RL) et relations associatives (AQ, QB, RO). Dans les terminologies source, les relations peuvent être plus fines ou plus grossières que ces 11 relations d'UMLS, ou même absentes, mais lors de leur intégration dans UMLS, elles sont homogénéisées. Par exemple, la relation source *isa* est assignée à la catégorie *CHD*, la relation *inverse-isa* à la catégorie *PAR*, la relation *has-component* à la catégorie *RO*, la relation *same-as* à la catégorie *SY*, etc. Ces 11 catégories de relations sont assignées selon la documentation des terminologies source ou selon la compréhension de ces sources par l'équipe de la NLM. Nous exploitons ici la version 2011AA d'UMLS.

3 Méthode d'analyse des relations

Dans ce travail, nous utilisons le terme *concept* dans le sens où il est défini dans UMLS : un ensemble de termes ayant le même sens ou des sens très proches et qui sont regroupés sous un identifiant unique, nommé *CUI*. Nous utilisons le terme *relations multiples* pour nous référer aux situations où il existe plus d'une relation entre deux concepts. Cette situation est très fréquente dans UMLS car rappelons que lorsqu'il intègre une terminologie, UMLS préserve l'ensemble des relations. Cela peut mener à des éventuelles redondances entre les relations issues de terminologies

distinctes mais aussi à des inconsistances de relations simples ou multiples. Nous employons le terme *relation simple* lorsqu'une seule relation existe entre deux concepts. L'objectif de notre méthode consiste à étudier d'une part, les relations multiples entre deux concepts distincts et d'autre part, les paires composées de deux concepts identiques mais qui sont associés par des relations simples ou multiples, autres que de synonymie.

Nous effectuons une analyse quantitative et une analyse qualitative. L'analyse quantitative consiste à étudier les fréquences des relations multiples et les combinaisons fréquentes de ces relations. L'analyse qualitative consiste à étudier la consistance des relations multiples. Nous étudions la compatibilité entre les catégories de relations, en examinant les relations UMLS et les relations telles qu'elles sont définies dans les terminologies source. Nous exploitons aussi la catégorisation sémantique des concepts pour vérifier la consistance des relations existant entre eux. Nous exploitons les *TS* catégorisant les concepts ainsi que leurs ***GS***. Une relation donnée R entre deux concepts *C1* et *C2* est ainsi analysée à deux niveaux (*TS* et ***GS***) selon les principes suivants :

(1) Au niveau des ***GS***. Si la relation *R* est de type hiérarchique ou de synonymie, elle est considérée comme consistante si les deux concepts reliés par cette relation appartiennent au même ***GS*** ou, s'ils appartiennent à plusieurs ***GS***, au moins un des ***GS*** doit être commun.

(2) Au niveau des *TS*. La relation R est considérée consistante si les concepts *C1* et *C2* sont catégorisés respectivement par *TS1* et *TS2* et :

- si la relation R est hiérarchique (à part *SIB*), *TS1* et *TS2* devraient aussi être reliés hiérarchiquement. Plus précisément, si *C1* est associé à *C2* via une relation *RN* ou *CHD* (resp. *RB* ou *PAR*), *TS1* doit être identique à *TS2* ou le descendant (resp. l'ancêtre) de *TS2*. Si les concepts reliés sont catégorisés par plus d'un *TS*, au moins une paire *TS1-TS2* doit respecter la règle pour que la relation soit considérée comme consistante. Notons qu'à ce niveau, la relation *SIB* ne peut pas être évaluée parce que les concepts frères peuvent être catégorisés par des *TS* indépendants;
- si R est une relation de synonymie, *TS1* et *TS2* devraient être identiques.

Ces règles sont implémentées à l'aide de scripts Perl. Telles que définies, ces règles ne peuvent pas être appliquées aux concepts reliés par des relations associatives. En effet, la particularité des relations associatives est de relier les concepts catégorisés par les ***GS*** (et *TS*) différents, comme par exemple ***Disorders*** et ***Procedures*** pour les concepts *Hypotension (C0020649)* et *Blood pressure determination (C0005824)* (le deuxième concept correspond à la procédure qui permet de diagnostiquer le premier concept).

Nb rel	Fréquence	Même concept	%
1	27 682	27 682	100,0
2	436 609	42 412	9,7
3	70 366	12 626	18,0
4	14 546	3 812	26,2
5	6 175	3 701	60,0
6	2 152	1 633	75,9
7	1 370	1 263	92,2
8	781	762	97,6
9	231	231	100,0

Table 1: Nombres de relations multiples dans UMLS, y compris parmi les paires constituées d'un même concept, et leurs fréquences.

4 Résultats

Analyse quantitative. La table 1 indique les fréquences de relations multiples. Sur la première ligne du tableau, nous présentons également le nombre de relations simples entre un même concept, uniquement lorsqu'il s'agit d'une relation autre que de synonymie. Dans UMLS, chaque relation source est représentée dans deux directions : pour tout triplet (C1, PAR, C2) enregistré, le triplet inverse (C2, CHD, C1) est également sauvegardé. Nous avons dédoublé le matériel pour éviter ce biais. Le nombre de relations multiples (première colonne), indépendamment de leur nature, va de deux à neuf. Dans la troisième colonne, nous indiquons les fréquences de ces relations lorsqu'elles existent entre un concept et lui-même. La dernière colonne indique le pourcentage de relations simples et multiples liant un même concept. Nous pouvons observer que ce pourcentage augmente avec le nombre de relations multiples. Il est aussi intéressant de constater qu'à partir de 7 relations différentes, les relations multiples concernent très largement les paires constituées d'un même concept, jusqu'à atteindre 100 % avec 9 re-

	Nombre de relations				Total	
	nb >= 3	%	nb=2	%	nb	%
Relations consistantes	185 602	79,8	535 699	67,9	721 301	70,6
Relations partiellement consistantes	122	0,1	802	0,1	924	0,1
Relations inconsistantes	4 240	1,8	89 451	11,4	93 691	9,2
Relations non évaluées	42 577	18,3	162 442	20,6	205 019	20,1
Total	232 541	100,0	788 394	100,0	1 020 935	100,0

Table 2: Consistance de relations au niveau ***SG*** : relations consistantes, partiellement consistantes, inconsistantes.

	Nombre de relations				Total	
	nb >= 3	%	nb=2	%	nb	%
Synonymie : consistant	25 260	10,9	41 936	5,4	67 196	74,7
Synonymie : inconsistant	6 548	2,8	16 194	2,1	22 742	25,3
Hiérarchique : consistant	98 933	42,5	356 549	45,2	455 482	74,0
Hiérarchique : inconsistant	25 540	11,0	134 059	17,0	159 599	26,0
Relations non évaluées	76 260	32,8	239 656	30,4	315 916	30,9
Total	232 541	100,0	788 394	100,0	1 020 935	100,0

Table 3: Consistance de relations de synonymie et hiérarchiques au niveau *ST*.

lations. De manière générale, avec l'augmentation du nombre de terminologies source, le nombre de relations multiples augmente aussi.

Analyse qualitative. Nous obtenons 456 ensembles possibles de relations multiples. Parmi les plus fréquents, il y a des ensembles homogènes (*CHD RN*, *PAR RB*, *RQ SY*), qui confirment la pertinence des grandes classes que nous avons définies. Les relations de ces ensembles correspondent à des relations très proches, mais plus ou moins précises : dans l'ensemble *PAR RB*, la relation *PAR* est plus précise que *RB*. Mais il y a aussi des ensembles non homogènes comme *PAR RO*, *CHD RO* ou *RL RQ SY*, où une des relations est sous-spécifiée et souvent différente. Il existe par ailleurs des ensembles illustrant la différence de granularité entre les terminologies intégrées dans UMLS : *CHD SIB*, *PAR SIB*, *CHD RN SIB*, *SIB SY*, *RQ SIB SY*, etc. Effectivement, si l'on considère un lien comme *CHD*, cela implique de créer un niveau hiérarchique supplémentaire par rapport à la situation où on le considère comme *SIB*. Dans d'autres situations, les ensembles contiennent des relations contradictoires : *CHD PAR SY*, *CHD PAR RB RN SY*. Les incompatibilités de relations augmentent avec le nombre de relations multiples.

Catégorisation sémantique. Au niveau ***GS*** (table 2), plus de 70 % de relations multiples sont consistantes et seulement 9,2% inconsistantes, tandis qu'au niveau *ST* (table 3), jusqu'à 25 % des relations de synonymie et hiérarchiques sont inconsistantes. Dans ces tables, nous présentons les résultats séparément lorsqu'il existe deux relations multiples (*n=2*) et lorsqu'il existe au moins trois relations multiples (*n>=3*). Notons qu'il y a plus de relations de la première catégorie et ces relations montrent un taux d'inconsistance plus élevé au niveau ***GS*** et *ST*.

5 Analyse détaillée et discussion

En allant au-devant de l'état de l'art, nous appliquons une analyse sémantique fine des relations d'UMLS grâce à l'exploitation des niveaux hiérarchiques supérieurs. Nous étudions aussi la compatibilité des relations et essayons de détecter les raisons de ces incompatibilités et inconsistances.

Catégorisation sémantique. Au niveau ***GS***, les relations de synonymie et hiérarchiques sont correctes avec moins de 10 % de relations inconsistantes. Le pourcentage de relations inconsistantes est plus élevé en présence de deux relations. Un exemple d'inconsistance d'une relation de synonymie est trouvé dans le couple {*Child health care (C0008078)*, *Child Nutritional Physiological Phenomena (C1720755)*} (figure 2(a)). Ces concepts appartiennent à des ***GS*** différents :

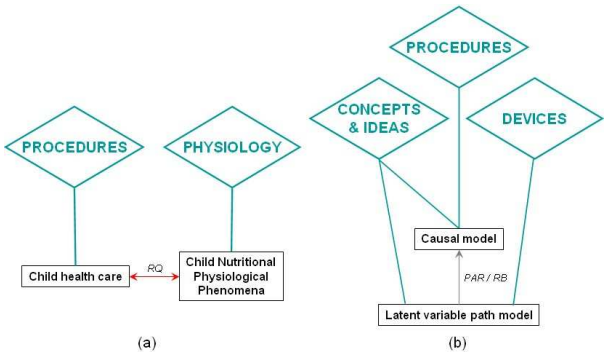

Figure 2: Categorisation sémantique : (a) relation de synonymie inconsistante, (b) relation hiérarchique partiellement consistante.

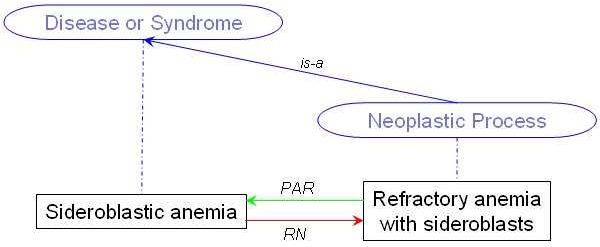

Figure 3: Détermination de la relation consistante (en vert) au sein des relations multiples, selon la categorisation sémantique des concepts.

Procedures et ***Physiology***, respectivement. Dans d'autres situations, un des deux concepts (ou les deux) appartiennent à plus d'un ***GS***. Les relations sont alors jugées partiellement consistantes si les concepts partagent au moins un ***GS***. C'est le cas par exemple de la relation hiérarchique qui existe entre *Causal model (C0681944)* et *Latent variable path model (C0681946)* car ils ont le ***GS Concepts & Ideas*** en commun, en plus de ***Procedures*** et ***Devices*** respectivement (figure 2(b)). Il existe aussi des situations de relations complètement inconsistantes. Ainsi, *Operating room (C0029064)* et *Surgical Procedures, Operative (C0543467)* sont liés par une relation de synonymie inconsistante parce qu'ils appartiennent à des ***GS*** distincts : ***Organizations*** et ***Objects*** pour *Operating room* et ***Procedures*** pour *Surgical Procedures, Operative*.

Au niveau *TS*, environ 25 % des relations de synonymie et hiérarchiques sont inconsistantes. À ce niveau, il est parfois possible de détecter quelles sont les relations inconsistantes. Ainsi, 132 paires sont dans ce cas dans l'ensemble *nb >= 3* et 54 paires dans l'ensemble *nb=2*. Par exemple, *Sideroblastic anemia (C0002896)* est relié hiérarchiquement à *Refractory anemia with sideroblasts (C0334679)* au travers des relations contradictoires *RN* et *PAR* et sont catégorisés respectivement par les *TS Disease or Syndrome* et *Neoplastic Process* (figure 3). Le *TS* de *Sideroblastic anemia* est le père du *TS* de *Refractory anemia with sideroblasts*. La relation *a priori* correcte dans cette situation est donc *CHD* et non *RB*, qui correspond à son contraire. Néanmoins, ce type de raisonnement automatique nécessiterait que la catégorisation d'UMLS soit systématiquement consistante, ce qui n'est pas le cas (Cimino et al., 2003; Gu et al., 2004). Par ailleurs, lorsque les concepts sont catégorisés par le même *TS*, il n'est pas possible de déterminer laquelle des relations multiples est consistante : *Enterovirus meningitis (C0276430)* est lié hiérarchiquement à *Viral meningitis (C0025297)* avec les relations *RB, PAR* et *CHD*. Comme ces deux concepts appartiennent au même *TS Disease or Syndrome*, on ne peut pas déterminer automatiquement quelles sont les relations correctes au sein de cette relation multiple.

Inconsistances dues aux terminologies source. Parmi les terminologies source qui apportent des inconsistances et des incompatibilités, on trouve très souvent MedDRA, RCD, CSP et ICD10. Nous illustrons cette situation avec la terminologie MedDRA dédiée à la représentation des effets indésirables (Brown et al., 1999). Elle contient plus de 80 000 termes, mais seulement cinq niveaux hiérarchiques. Au niveau le plus haut, les termes sont organisés en 26 classes générales liées à des localisations anatomiques, comme *Neoplasms benign, malignant and unspecified*, *Investigations*, *Blood and lymphatic system disorders*, etc. Les termes les plus informatifs se trouvent à des niveaux inférieurs : les termes préférés (*PT*) et les termes de bas niveau (*LLT*). Selon la particularité structurelle de MedDRA : (1) les termes *LLT* sont synonymes de leur *PT* et les libellés des *PT* se retrouvent parmi les *LLT* ; (2) mais les *LLT* peuvent aussi être plus spécifiques par rapport aux *PT* (Merrill, 2008). Cette confusion entre les relations de synonymie et hiérarchiques ainsi qu'une structuration hiérarchique limitée peuvent conduire à des inconsistances, ensuite propagées dans UMLS. Par exemple, la paire {*Prostate cancer stage III (C0278836)*, *Malignant neoplasm of prostate (C0376358)*} est liée via deux relations dans UMLS : *PAR* et *SIB*, toutes les deux issues de MedDRA. Comme ces deux concepts sont caté-

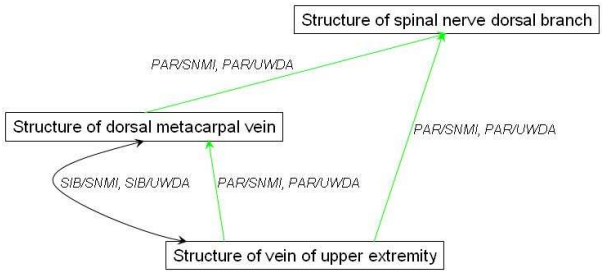

Figure 4: Concurrence de relations *PAR* et *CHD* dans UMLS illustrant une différence de granularité entre les terminologies source.

gorisés par le même *TS*, il n'est pas possible de déterminer quelle est la relation consistante. La cooccurrence de ces deux relations indique une inconsistance dans MedDRA où les termes sont représentés de la manière suivante :

1. *Prostate cancer stage III PAR Prostatic neoplams malignant*
2. *Malignant neoplasms of prostate PAR Prostate cancer PAR Prostatic neoplams malignant*

Dans UMLS, les termes *Prostatic neoplams malignant*, *Malignant neoplasms of prostate* et *Prostate cancer* appartiennent au concept C0376358, tandis que *Prostate cancer stage III* appartient à un autre concept C0278836. Lors de l'intégration des relations, le lien entre C0278836 et C0376358 est catégorisé comme *CHD* à cause de la relation en (1), et en même temps ce lien est aussi catégorisé comme *SIB* à cause des relations en (2). Dans cette situation, la structure de MedDRA et l'interprétation de cette structure par la NLM conduisent à des inconsistances.

Inconsistances dues à la fusion au sein d'UMLS. Plusieurs types d'inconsistances proviennent de la fusion des terminologies. En effet, la conception de chaque terminologie est différente, chacune montre des granularités structurelles différentes et plus ou moins compatibles. Sur la figure 4, la paire {*Structure of dorsal metacarpal vein (C0226799)*, *Structure of vein of upper extremity (C0226792)*} correspond à un cas typique de différence de granularité dans l'organisation des concepts. Deux terminologies décrivent une relation *PAR* et deux autres une relation *SIB*. Notons qu'il est fréquent que l'incompatibilité des relations soit couplée avec des problèmes au niveau de la catégorisation sémantique de leurs concepts.

Un concept a des relations multiples avec lui-même. Comme nous l'avons illustré dans la table 1, il n'est pas rare qu'un concept ait des relations multiples avec lui-même. Dans ce cas, il n'y a pas d'inconsistances au niveau des ***GS*** et des *TS*. Par contre, les relations montrent beaucoup de combinaisons possibles et improbables. Par exemple, le concept *Blepharoptosis (C0005745)* a neuf relations avec lui-même : *CHD PAR RB RL RN RO RQ SIB SY* issues de 15 terminologies source. Si on les fait "voter", 11 terminologies se prononcent pour la relation *SY*, six pour la relation *RQ*, trois pour *PAR* et pour *CHD* et d'autres de manière hapaxique pour les relations qui restent. C'est donc la relation de synonymie qui "l'emporterait". Ces incompatibilités sont en partie héritées des terminologies source et en partie générées lors de la fusion dans UMLS. L'existence de relations entre un concept et lui-même est une situation typique qui mène à des cycles dans les terminologies et qui requiert des contrôles de qualité et de consistance (Bodenreider, 2001).

Détecter et prévenir les inconsistances et les incompatibilités. Les différents tests et analyses menés montrent qu'il existe plusieurs types d'inconsistances et d'incompatibilités aux niveaux des relations et de la catégorisation sémantique. Nous avons aussi vu que ces deux niveaux sont liés entre eux. Les inconsistances peuvent provenir des terminologies source ou bien être induites suite à leur fusion. Indépendamment des terminologies et des contextes de leur utilisation, des tests simples peuvent être effectués pour contrôler leur qualité. En voici quelques exemples, inspirés principalement par le travail présenté :

- Contrôler la consistance de la catégorisation sémantique : (1) pour chaque catégorie de relations, définir les règles de consistance ; (2) exploiter les niveaux hiérarchiques supérieurs et les relations entre les concepts pour vérifier la consistance de relations entre des concepts plus spécifiques.
- Cooccurrence de relations de synonymie et hiérarchiques (différence de granularité) : créer un autre niveau hierarchique.
- Relation hiérarchique entre un concept et lui-même : créer un autre concept.
- Relations autres que de synonymie et hiérarchiques entre un concept et lui-même : ignorer ces relations.

- Contradiction entre deux relations issues de la même terminologie : donner moins de confiance à cette terminologie puisqu'elle semble contenir des inconsistances.
- Contradiction entre deux relations issues de deux terminologies : définir la confiance de chaque terminologie pour déterminer la fiabilité et la consistance de ses relations.

6 Conclusion et perspectives

Il apparaît qu'il est assez commun d'avoir des inconsistances et des incompatibilités dans les terminologies, qui sont ensuite répercutées et amplifiées si une fusion de terminologies est effectuée. Il peut alors être utile de faire des tests pour vérifier leur consistance et qualité. Nous effectuons des tests simples pour vérifier la qualité et assurer la consistance des ressources terminologiques. Notre proposons d'aller au-devant de l'état de l'art et les règles proposées effectuent une analyse fine de la sémantique des relations simples et multiples pour vérifier leurs consistance et compatibilité. Ces règles peuvent être appliquées à UMLS, comme dans ce travail, ou à toute autre terminologie, car nous avons fait en sorte de les généraliser. Parmi les perspectives qui se présentent à ce travail, mentionnons l'évaluation de la consistance des relations associatives en vérifiant que les concepts qu'elles associent appartiennent à des ***GS*** et *TS* différents. Une étude approfondie des relations *SIB* pourrait être faite : celles-ci semblent être aussi proches des relations hiérarchiques que des associatives. Enfin, nous envisageons de mettre à disposition des chercheurs les tests de qualité proposés ici sous forme d'un module Perl.

References

O Bodenreider and AT McCray. 2003. Exploring semantic groups through visual approaches. *J Biomed Inform*, 36(6):414–432.

O Bodenreider. 2001. Circular hierarchical relationships in the UMLS: etiology, disgnosis, treatment, complications and prevention. In *Proc AMIA*, pages 57–61.

O Bodenreider. 2003. Strength in numbers: exploring redundancy in hierarchical relations across biomedical terminologies. In *AMIA*, pages 101–5.

EG Brown, L Wood, and S Wood. 1999. The medical dictionary for regulatory activities (MedDRA). *Drug Saf.*, 20(2):109–17.

KE Campbell, MS Tuttle, and KA Spackman. 1998. A "lexically-suggested logical closure" metric for medical terminology maturity. In *Proc AMIA Symp.*, pages 785–9.

JJ Cimino, H Min, and Y Perl. 2003. Consistency across the hierarchies of the umls semantic network and metathesaurus. *J Biomed Inform*, 36(6):450.

JJ Cimino. 1998. Auditing the unified medical language system with semantic methods. *J Am Med Inform Assoc.*, 5(1):41–51.

M D'aquin, J Euzenat, C Le Duc, and H Lewen. 2009. Sharing and reusing aligned ontologies with cupboard. In *K-CAP 2009*, pages 179–180.

N Fridman Noy and MA Musen. 2000. Prompt: Algorithm and tool for automated ontology merging and alignment. In *AAAI*, pages 450–455.

H Gu, Y Perl, G Elhanan, H Min, L Zhang, and Y Peng. 2004. Auditing concept categorizations in the umls. *Artif Intell Med*, 31(1):29–44.

M Klein and D Fensel. 2001. Ontology versioning on the semantic web. In *Proc Semantic Web Working Symposium*, pages 75–91.

DA Lindberg, BL Humphreys, and AT McCray. 1993. The unified medical language system. *Methods Inf Med*, 32(4):281–291.

Y Liu, J Yu, Z Wen, and S Yu. 2004. Two kinds of hypernymy faults in WordNet: the cases of ring and isolator. In *Proc GWN 2004*, pages 347–351.

A Maedche, B Motik, L Stojanovic, R Studer, and R Volz. 2002. Managing multiple ontologies and ontology evolution in ontologging. In *Symposium IIP*, pages 51–63.

GH Merrill. 2008. The meddra paradox. In *AMIA Annu Symp Proc*, pages 470–4.

F Mougin and O Bodenreider. 2005. Approaches to eliminating cycles in the umls metathesaurus: naïve vs. formal. In *AMIA Annu Symp Proc*, pages 550–4.

G Qi, P Haase, Z Huang, Q Ji, JZ Pan, and J Volker. 2008. A kernel revision operator for terminologies - algorithms and evaluation. In *International Conference on The Semantic Web*, pages 419–34.

B Smith, J Williams, and S Schulze-Kremer. 2003. The ontology of the gene ontology. In *AMIA 2003*, pages 609–613.

P Smrz. 2004. Quality control for Wordnet development. In *Proc GWN 2004*, pages 206–212.

D Wei, M Halper, G Elhanan, Y Chen, Y Perl, J Geller, and KA Spackman. 2009. Auditing snomed relationships using a converse abstraction network. In *AMIA Annu Symp Proc.*, pages 685–9.

X Zhu, JW Fan, DM Baorto, C Weng, and JJ Cimino. 2009. A review of auditing methods applied to the content of controlled biomedical terminologies. *J Biomed Inform*.