Proceedings of the 15th conference on Pattern Recognition and Artificial Intelligence (RFIA 2006)
(electronic proceedings)

# Éliminer les cycles dans les systèmes terminologiques : comparaison de deux approches

# Comparing two approaches to eliminating cycles in terminological systems

F. Mougin[1] O. Bodenreider[2]

[1]EA 3888, IFR 140, Faculté de Médecine, Université de Rennes I
[2]National Library of Medicine, Bethesda, Maryland

Laboratoire d'Informatique Médicale, Avenue du Pr Léon Bernard, 35043 Rennes, France
fleur.mougin@univ-rennes1.fr

## Résumé

*Les ressources terminologiques ont souvent une structure hiérarchique. Les systèmes intégrant plusieurs terminologies ont de ce fait une structure poly-hiérarchique et il n'est pas rare d'y rencontrer des incohérences parmi les relations, sous la forme de cycles dans le graphe terminologique. Notre étude présente les facteurs à l'origine de cycles dans le graphe du Metathesaurus de l'UMLS, un système intégrant plus de 100 vocabulaires biomédicaux. Nous comparons deux approches proposées pour y remédier. L'une, naïve, consiste à empêcher les boucles lors du parcours de graphe en évitant de visiter un nœud plus d'une fois. L'autre, formelle, vise à identifier et éliminer a priori les cycles existant dans le graphe global au moyen de règles et d'heuristiques. Notre comparaison est basée sur une application du graphe terminologique : le calcul de l'ensemble des descendants d'un concept. L'approche naïve est simple à mettre en œuvre mais ne garantit pas la pertinence sémantique des relations supprimées. L'approche formelle améliore la cohérence sémantique des ensembles de descendants, mais elle est complexe à mettre en œuvre et requiert l'intervention d'un expert du domaine. Enfin, il faut noter que la cohérence sémantique des ensembles de descendants reste imparfaite, même avec l'approche formelle, ce qui traduit une limitation de l'intégration terminologique dans l'UMLS.*

## Mots Clef

Terminologies biomédicales, graphe, relations hiérarchiques, cycles

## Abstract

*Terminological resources often have some hierarchical organization. Therefore, systems integrating multiple terminologies have a polyhierarchical organization and may exhibit conflicting hierarchical relations, leading to cycles in the graph. Our study presents some of the causes of cycles in the UMLS Metathesaurus graph, a system integrating more than 100 biomedical vocabularies. We compare two approaches to eliminating these cycles. The naïve approach simply avoids loops during the graph traversal by preventing nodes from being visited more than once. The formal approach eliminates cycles from the graph a priori, using a set of rules and heuristics. An application of terminological graphs is used in this comparison: computing the set of descendants for a given concept. The naïve approach is easier to implement, but does not ensure that inappropriate semantic relations are eliminated. In contrast, the formal approach improves the semantic consistency in sets of descendants but is more complex and requires domain expertise. Finally, although semantic consistency is better with the formal approach, it remains suboptimal due to limitations of the UMLS itself.*

## Keywords

Biomedical terminologies, graph, hierarchical relations, cycles

## 1 Introduction

La plupart des terminologies spécialisées sont créées pour répertorier le vocabulaire (i.e., les termes) utilisé dans un domaine particulier. En médecine, par exemple, les noms de maladies et d'actes médicaux sont collectés par la Classification Internationale des Maladies et la Classification Commune des Actes Médicaux. En pratique, la plupart des terminologies sont plus que de simples listes de termes dans la mesure où leurs termes sont organisés en arbres, afin de faciliter l'utilisation et la maintenance de ces terminologies.

Celles-ci sont utilisées comme source de connaissances dans de nombreuses applications, par exemple pour le recueil de données en médecine [16] ou la classification de ressources documentaires [14]. Les relations hiérarchiques représentées dans les terminologies (*parent/enfant*, *terme plus général/terme plus spécifique*[1]) correspondent généralement à des relations de subsomption (*est-un*, *est-une-sous-classe-de*). En terme de structure, les hiérarchies constituées d'héritage simple sont des arbres et les hiérar-

[1] *Broader than/narrower than*

chies qui autorisent un héritage multiple sont des graphes acycliques orientés (en anglais directed acyclic graph, ou DAG). De telles structures hiérarchiques peuvent être traversées aisément, permettant aux utilisateurs de trouver des chemins parmi les concepts et de réaliser des fermetures transitives.

Certains systèmes terminologiques résultent de l'intégration de terminologies dont la structure hiérarchique correspond à une relation d'ordre partiel. Dans ces systèmes, on s'attend à ce que l'intégration de hiérarchies produise un DAG, c'est-à-dire un graphe dans lequel aucun descendant d'un concept n'est simultanément un ancêtre de ce même concept [1]. En pratique, il est rare de ne pas y rencontrer de cycles.

Les deux causes majeures de la présence de cycles sont les suivantes. D'abord, certains vocabulaires sources ne sont pas des graphes acycliques eux-mêmes. Une fois intégrés dans le système, les cycles existant originellement dans ces vocabulaires deviennent également des cycles dans le graphe global. Par ailleurs, le processus d'intégration peut parfois créer des cycles en autorisant des vues conflictuelles à coexister dans une même structure.

Les problèmes posés par les relations hiérarchiques circulaires dans de tels systèmes peuvent être abordés de différentes manières. Une **approche naïve**, relativement simple, consiste à empêcher les boucles pendant le parcours du graphe composant le système en conservant le suivi des nœuds déjà visités. Mais l'inconvénient de cette approche est qu'elle supprime les liens inappropriés sans établir de distinction sémantique entre bons et mauvais liens. En d'autres termes, cette approche ne garantit pas que les liens ignorés pendant le parcours de manière à éviter de boucler soient réellement les liens qui doivent être éliminés ; le fait que tel lien sera ignoré dépend uniquement de l'ordre dans lequel le graphe est traversé. Une **approche formelle** permet de définir des règles visant à éliminer les liens inappropriés à l'origine des relations hiérarchiques circulaires. Cependant, l'algorithme implémentant une telle méthode est relativement complexe et nécessite également une intervention manuelle, ce qui le rend difficile à appliquer dans des applications spécifiques.

Notre étude porte sur le Metathesaurus® de l'UMLS® (Unified Medical Language System®) [9,17], qui est un système intégrant de nombreuses terminologies biomédicales dans un espace sémantique unique. Il sera l'illustration d'un système global intégrant des structures hiérarchiques distinctes. L'objectif de cette étude est d'identifier les problèmes causés par la constitution de tels systèmes globaux et d'introduire deux approches possibles pouvant les résoudre. Après une présentation de l'UMLS et des deux approches, nous évaluons le bénéfice pratique d'utiliser l'approche formelle pour éliminer les relations hiérarchiques circulaires dans de tels systèmes, en comparaison à l'approche naïve. Pour ce faire, nous comparons la taille et la cohérence sémantique des ensembles de descendants obtenus par chacune des approches pour chaque concept du Metathesaurus de l'UMLS. Nous discutons enfin les avantages et les inconvénients de chaque approche et les limites de cette étude.

## 2 L'UMLS

L'UMLS[2] (Unified Medical Language System) est un système incluant deux sources d'information sémantique : le Metathesaurus et le Semantic Network (Réseau Sémantique).

Le **Metathesaurus®**, intégrant plus de 100 vocabulaires sources, constitue un large graphe comprenant un peu plus d'un million de nœuds (concepts) et de 16 millions de relations entre ces concepts. Chaque concept est constitué de termes synonymes provenant des vocabulaires sources. La quantité importante de relations présentes dans le Metathesaurus est due au fait que, par convention, l'ensemble des relations existant dans les vocabulaires sources doivent être intégrées dans l'UMLS, pour éviter la perte d'information pertinente spécifique à un vocabulaire source donné dans son contexte propre. Environ 5 millions de relations hiérarchiques sont représentées (sous les formes *parent*/*child* et *broader*/*narrower than*). Le graphe ainsi constitué présente des cycles, comme l'ont montré de nombreux travaux [5,8,15]. Comparativement à celui des vocabulaires sources qu'il intègre, le graphe du Metathesaurus est à la fois plus large et plus profond.

Le **Semantic Network** est un réseau beaucoup plus restreint de 135 types sémantiques organisés de manière arborescente. Les types sémantiques ont été agrégés en 15 groupes sémantiques [10]. Chaque concept du Metathesaurus est catégorisé par au moins un type sémantique du Semantic Network, indépendamment de sa position hiérarchique dans le vocabulaire dont il est issu. La version 2004AA de l'UMLS est utilisée dans cette étude.

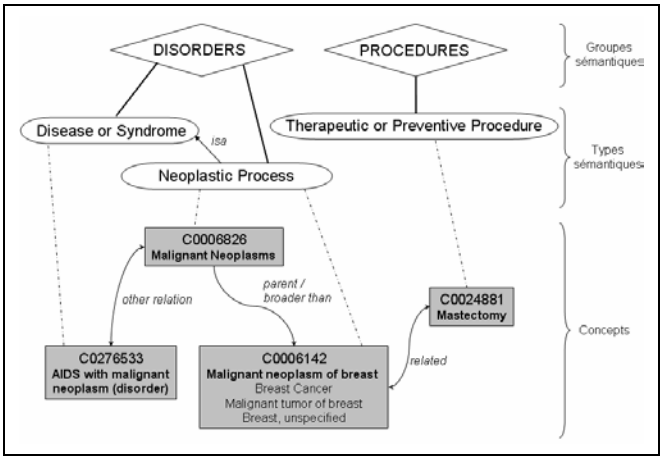


*Figure 1 – Portion de l'UMLS. Les éléments grisés font partie du Metathesaurus, les traits pointillés sont les relations assignant les concepts à un (ou plusieurs) type(s) sémantique(s) et les traits pleins correspondent à l'association types/groupes sémantiques.*

[2] http://umlsks.nlm.nih.gov/

La figure 1 illustre une portion de l'UMLS constituée de quatre concepts, les relations hiérarchiques et associatives existant entre eux, les trois types sémantiques auxquels ils sont associés et les relations hiérarchiques existant entre ces derniers ainsi que le groupe sémantique regroupant ces types sémantiques.

La plupart des relations hiérarchiques présentes dans le Metathesaurus sont valides, comme par exemple, la relation PAR (*est-père-de*) entre *Neoplasms* (C0027651) et *Malignant neoplasm of breast* (C0006142), provenant de plusieurs sources. D'un autre côté, certains vocabulaires utilisent, pour organiser leurs termes hiérarchiquement, des relations qui ne sont pas vraiment hiérarchiques. On a par exemple dans Alcohol and Other Drug Thesaurus[3] (AOD) une relation du type BT (*broader term*) entre les concepts *biological rest* (C0678686) et *Fatigue* (C0015672). Dans AOD Thesaurus, cette relation associative entre fatigue et repos, utile pour la recherche d'information, est implémentée sous la forme d'une relation hiérarchique, et préservée comme telle dans le Metathesaurus. En effet, par convention, lors de l'intégration dans l'UMLS, toutes les relations utilisées pour organiser les vocabulaires sources hiérarchiquement participent à la structure hiérarchique du Metathesaurus.

# 3 Origine des cycles

Une notion fondamentale est nécessaire pour comprendre les mécanismes à l'origine des relations hiérarchiques circulaires dans les systèmes intégrant plusieurs terminologies. Bien qu'enregistrées et utilisées au niveau conceptuel, de nombreuses relations hiérarchiques ont été définies au niveau du terme. Autrement dit, le regroupement de termes synonymes sous un même concept modifie la structure originelle des vocabulaires sources. Tandis que le processus produit un système dont la structure est polyhiérarchique, unifiée et utile, les relations hiérarchiques circulaires peuvent être vues comme son effet secondaire. Dans une étude précédente [4], nous avons identifié différents facteurs à l'origine de cycles dans un système intégrant des terminologies de structure hiérarchique diverse : l'UMLS. Nous rappelons ici brièvement les principales catégories recensées dans cette étude.

## 3.1 Granularité et connaissance implicite

Le niveau de granularité du vocabulaire source peut être différent de celui du système global. Si le premier est plus fin que le second, deux termes très proches (mais cependant distincts dans le vocabulaire source) sont susceptibles d'être regroupés dans un même concept et générer ainsi dans le système d'intégration une relation réflexive, qui était à l'origine une relation hiérarchique (Figure 2). Ce cas de figure peut aussi se présenter dans le cas où le vocabulaire source associe de la connaissance implicite à un terme dans son contexte propre mais qui n'est pas reproductible dans le système global.

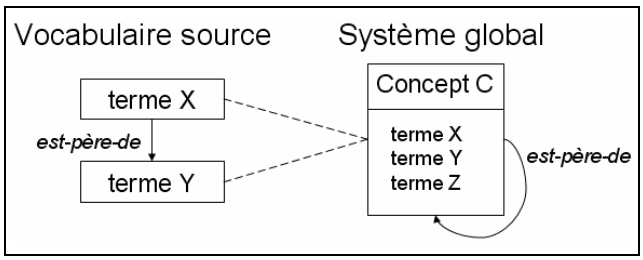


*Figure 2 – Exemple de relation réflexive*

## 3.2 Termes composés

Les termes comprenant des conjonctions de type « et » et « ou » posent des problèmes par leur imprécision [12]. En effet, *C1 et C2* peut être interprété comme *C1 avec C2*. Dans ce cas, le concept *C1 et C2* est fils à la fois de *C1* et du concept *C2* puisque c'est un concept plus précis. L'autre interprétation du concept *C1 et C2* est *C1 ou C2*, rendant ainsi les concepts *C1* et *C2* tous deux fils de *C1 et C2* qui est donc dans ce cas plus général. Cela entraîne l'apparition d'une relation hiérarchique circulaire directe (Figure 3). Par exemple, le concept *Veines de la tête et du cou* peut désigner des structures anatomiques communes aux deux sites (par exemple l'artère carotide) ainsi que l'ensemble des structures appartenant soit à la tête, soit au cou.

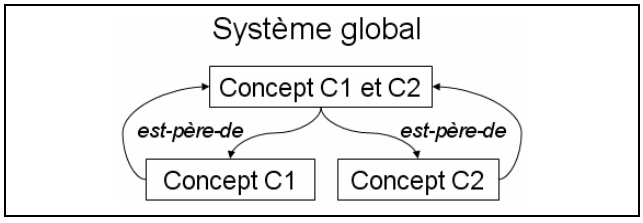


*Figure 3 - Exemple de deux relations hiérarchiques circulaires directes*

## 3.3 Conventions organisationnelles

Certains vocabulaires sources utilisent des relations non hiérarchiques pour organiser leurs termes hiérarchiquement. C'est le cas par exemple de la relation entre acides et sels : un acide n'est pas une sorte de sel, mais, combiné à une base, l'acide produit un sel et de l'eau. Par convention, certaines terminologies comme MeSH[4] représentent l'acide comme le père du sel. Cette représentation est utile pour la structuration des termes et la recherche d'information. Pour autant, il ne s'agit pas d'une relation de subsomption et d'autres terminologies peuvent adopter une autre convention pour représenter les acides et les sels qui en dérivent.

[3] http://etoh.niaaa.nih.gov/AODVol1/Aodthome.htm

[4] Medical Subject Headings, http://www.nlm.nih.gov/mesh/meshhome.html

## 3.4 Termes sous-spécifiés

Dans certains vocabulaires sources, on trouve des termes qui, à dessein, sont sous-spécifiés. Le processus d'intégration ne les différencie pas des termes proches mais plus précis. Par exemple dans l'UMLS, on trouve des termes contenant des expressions telles que « not otherwise specified » ou « NOS ». Les termes de la forme *X, NOS* qui sont souvent classés comme *est-fils-de X* dans les sources sont généralement associés au même concept que le terme *X* dans l'UMLS. Cela provoque donc une relation réflexive. Des cas de figure plus complexes où *X* et *X, NOS* sont reliés dans différents vocabulaires par l'intermédiaire d'un terme *X,xxx* et qui sont regroupés lors du processus d'intégration, créent là aussi une relation hiérarchique circulaire directe (Figure 4).

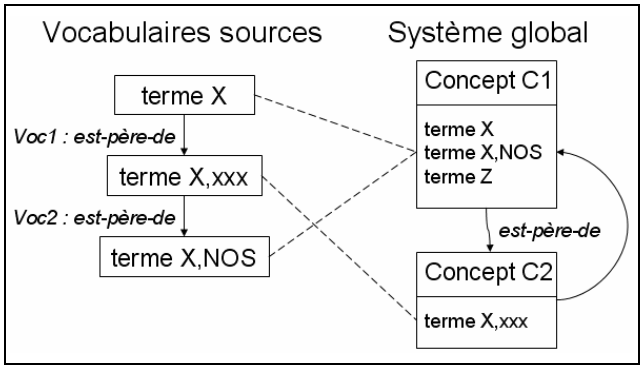


*Figure 4 – Exemple d'un autre facteur causant une relation hiérarchique circulaire directe*

D'autres facteurs plus complexes, non présentés ici, peuvent être à l'origine de relations circulaires hiérarchiques indirectes [4].

# 4 Approches

Deux approches peuvent être implémentées pour supprimer les cycles dans un système contenant des relations hiérarchiques circulaires. L'approche naïve consiste à éviter les boucles lors du parcours du graphe terminologique. L'approche formelle s'attache à définir un certain nombre de règles permettant d'éliminer les cycles *a priori* et de transformer le graphe terminologique en un DAG.

## 4.1 Approche naïve

L'approche naïve s'implémente de manière ponctuelle en fonction de l'application nécessitant un parcours de graphe. Elle consiste simplement à marquer les nœuds visités pendant le parcours du graphe, de manière à éviter de visiter ces mêmes nœuds une deuxième fois. Cette approche est efficace pour empêcher les boucles, mais naïve dans le sens où c'est uniquement l'ordre dans lequel les nœuds sont visités qui détermine quelle relation sera ignorée dans le cas d'un cycle.

Supposons que l'on souhaite obtenir les descendants d'un concept donné dans le Metathesaurus de l'UMLS. La méthode adoptée dans cette étude consiste à parcourir le graphe en profondeur d'abord, en traitant les fils d'un nœud donné dans l'ordre alphabétique des étiquettes (CUIs, i.e. concept unique identifiers) pour des raisons de reproductibilité. Ce choix est arbitraire. Les concepts *Desire for food* (C0003618), *Appetite Regulation* (C0003622) et *Food Intake Regulation* (C0086311) sont choisis pour illustrer le problème posé par l'utilisation de cette approche. La figure 5 montre qu'avec l'approche naïve et en partant du concept C0003618 pour obtenir ses descendants, la relation *C0086311 est-père-de C0003618* est ignorée car elle causerait un cycle dans le graphe (*C0003618 → C0003622 → C0086311 → C0003618*). Au contraire, la même relation est utilisée quand le graphe est parcouru en commençant par C0086311 (Figure 6), tandis que la relation *C0003622 est-plus-général-que C0086311* utilisée dans le graphe précédent est maintenant ignorée en raison de son implication dans le cycle (*C0086311 → C0003618 → C0003622 → C0086311*).

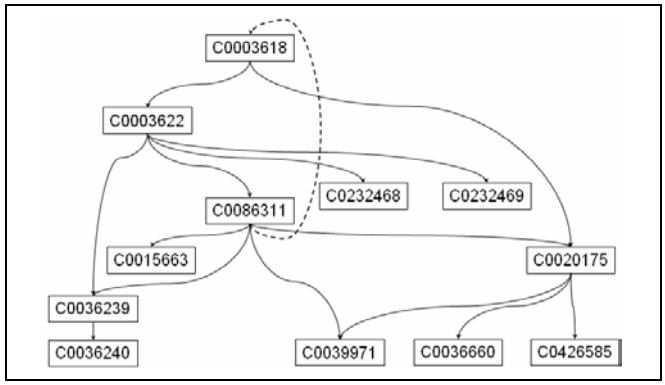


*Figure 5 – Les descendants de C0003618 (la relation C0086311 est-père-de C0003618 est ignorée)*

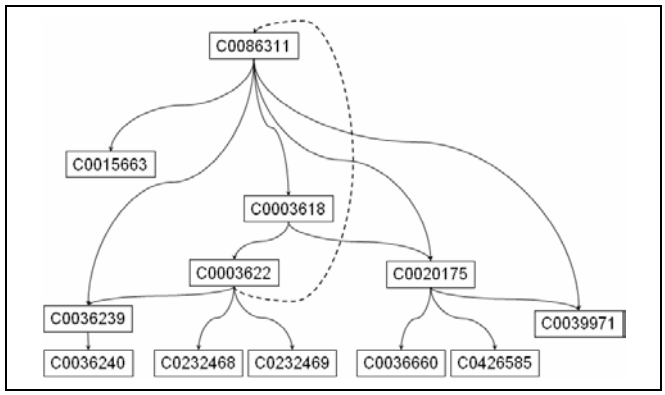


*Figure 6 – Les descendants de C008631 (la relation C0003622 est-plus-général-que C0086311 est ignorée)*

L'approche naïve est relativement simple à implémenter et permet une suppression automatique des cycles, sans recourir à un expert du domaine. Cependant, elle présente des limitations puisqu'elle ne garantit pas que les liens ignorés dans le parcours de graphe soient effectivement ceux qui sont sémantiquement incorrects.

## 4.2 Approche formelle

L'approche formelle, plus théorique, consiste en un ensemble d'heuristiques et de règles définies de manière à identifier et éliminer *a priori* tous les cycles du graphe global. En fonction du type de relation responsable du cycle, le traitement diffère.

**Relation réflexive**

Le traitement dans ce cas est trivial, il suffit de supprimer les relations du type *C est-père/fils-de C* ainsi que *C est-plus-général/spécifique-que C*.

**Relation circulaire directe**

Différents types de traitement peuvent être proposés en fonction des données et informations disponibles dans le système hiérarchique. Par exemple, la redondance des relations présentes entre deux concepts dans différents vocabulaires sources peut être utilisée. En d'autres termes, si la relation *C1 est-père-de C2* existe dans trois vocabulaires différents, tandis que la relation *C2 est-père-de C1* n'apparaît que dans un seul, on préfèrera garder la relation *C1 est-père-de C2* et éliminer l'autre. Des critères de confiance associés à tel vocabulaire source plutôt qu'à un autre peuvent également être pris en compte dans les règles pour déterminer si une relation doit être supprimée, ou au contraire, préservée.

**Relation circulaire indirecte**

Le traitement de ce type de relations nécessite la plupart du temps l'intervention manuelle d'un expert du domaine.

Le détail des règles et heuristiques de la méthode formelle définie pour éliminer les cycles dans le graphe de l'UMLS Metathesaurus est donné dans [4]. Dans l'exemple présenté dans la partie précédente illustrant les limitations de l'approche naïve, nous avons montré que le choix de la relation à supprimer était aléatoire alors que la méthode formelle identifie, elle, de manière consistante la relation *C0086311 est-père-de C0003618* comme étant incorrecte en raison du vocabulaire source dont elle provient.

L'approche formelle permet de construire un DAG à partir du graphe global, ce qui facilite les parcours d'arbres et les calculs de descendants notamment. Une fois terminé le traitement manuel effectué par l'expert, l'algorithme permettant de détecter la présence de cycles est relancé pour vérifier l'acyclicité du graphe global. En pratique, le calcul des graphes d'ascendants nécessitant moins de ressources que celui des descendants, l'acyclicité du graphe des ascendants de chaque concept du Metathesaurus de l'UMLS est testée après chaque modification, jusqu'à ce qu'aucun cycle ne soit identifié.

L'approche formelle reste cependant difficile à implémenter et requiert l'intervention d'un expert du domaine, ce qui présente une limite importante dans le cas de systèmes comprenant un grand nombre de cycles. C'est pour ces différentes raisons que nous avons voulu comparer les résultats obtenus par ces deux approches et évaluer le bénéfice d'éliminer les cycles d'une manière complexe et coûteuse par rapport à une approche plus simple mais moins rigoureuse.

# 5 Un cas d'étude : l'UMLS

## 5.1 Hypothèse

Nous souhaitons comparer les deux approches (naïve et formelle) sur le parcours de graphe du Metathesaurus de l'UMLS en utilisant une application des graphes terminologiques : le calcul de l'ensemble des descendants d'un concept[5]. Pour cela, nous considérons chaque concept du Metathesaurus et calculons ses descendants avec chaque méthode. Notre hypothèse est que l'approche formelle réduit le nombre de descendants obtenus et améliore la cohérence sémantique des ensembles de descendants.

## 5.2 Calcul de descendants

L'ensemble des descendants d'un concept consiste en la première génération des descendants de ce concept (i.e., ses enfants et ses termes plus spécifiques - narrower terms - dans le Metathesaurus) et leurs descendants, récursivement. Dans la théorie des graphes, cette opération s'appelle la fermeture transitive des relations hiérarchiques. Elle est réalisée par un parcours du graphe. Comme le Metathesaurus contient des cycles, des précautions doivent être prises pour éviter les boucles lors du parcours du graphe. L'**approche naïve** consiste simplement à marquer les noeuds visités pendant le parcours de graphe de manière à éviter de visiter le même noeud plusieurs fois. L'**approche formelle** transforme le Metathesaurus en un DAG préalablement au calcul des ensembles de descendants.

## 5.3 Méthodes d'évaluation

L'évaluation porte sur la cohérence sémantique des ensembles de descendants obtenus par les deux approches. Nous définissons d'abord un certain nombre de notions requises pour l'évaluation.

**Définition 1**

*Deux types sémantiques sont compatibles s'ils appartiennent au même groupe sémantique.*

Par exemple, les types sémantiques *Disease or Syndrome* et *Finding*, bien que non liés hiérarchiquement, sont compatibles puisqu'ils appartiennent tous les deux au groupe sémantique ***Disorders.***

**Définition 2**

*Un descendant est sémantiquement cohérent avec son concept source si et seulement si le type sémantique caté-*

[5] Les ensembles de descendants sont utilisés par exemple pour calculer tous les médicaments d'une classe thérapeutique donnée (e.g., tous les antibiotiques, représentés comme les descendants du concept *Antibiotique*).

*gorisant le descendant est le même que le type sémantique du concept source ou un descendant de ce type sémantique. (En cas de catégorisation multiple, c'est-à-dire lorsque les concepts sont catégorisés par plusieurs types sémantiques, la cohérence est requise pour au moins une des paires de types sémantiques seulement).*
Cette définition est similaire en ce qui concerne la cohérence sémantique par rapport aux groupes sémantiques.
Par exemple, le concept *Adrenal cortex diseases* (C0001614) est catégorisé comme *Disease or Syndrome*, un type sémantique du groupe ***Disorders***. Tous ses descendants appartiennent également au groupe sémantique ***Disorders***. Cependant, si la plupart des descendants sont catégorisés comme *Disease or Syndrome* ou *Neoplastic Process*, les types sémantiques suivants qui ne sont pas des descendants de *Disease or Syndrome* dans le Semantic Network sont aussi utilisés pour catégoriser certains descendants de *Adrenal cortex diseases* : *Anatomical Abnormality*, *Congenital Abnormality*, *Finding*, *Injury or Poisoning*, *Pathologic Function* et *Sign or Symptom*. Dans cet exemple, la cohérence sémantique des descendants du concept source est bonne du point de vue groupe sémantique car seul le groupe sémantique du concept source (***Disorders***) est représenté parmi les descendants. Il y a cependant une dispersion importante en terme de types sémantiques puisque six d'entre eux sont représentés en plus des descendants du type sémantique du concept source. Plusieurs descendants sont donc sémantiquement incohérents avec ce dernier.

**Définition 3**
*La cohérence sémantique d'un ensemble de concepts est mesurée par le rapport entre le nombre de concepts sémantiquement cohérents avec le concept source et le nombre total de concepts dans cet ensemble.*

**Comparer les ensembles de descendants**
Les ensembles de descendants obtenus pour un concept donné du Metathesaurus par les approches naïve et formelle ont été comparés respectivement comme suit. D'abord, une simple intersection des deux ensembles est réalisée pour identifier les concepts communs aux deux ensembles et ceux qui sont spécifiques à chacun. Nous recherchons ensuite la cohérence sémantique des deux ensembles en étudiant la distribution des types (et groupes) sémantiques dans ces ensembles. De plus, nous vérifions la compatibilité de chaque type (et groupe) sémantique représenté dans les descendants, par rapport à ceux du concept source. Le nombre de types (et groupes) sémantiques représentés dans les ensembles de descendants constitue l'aspect quantitatif de la cohérence sémantique, tandis que la compatibilité des types (et groupes) sémantiques représentés dans les descendants, par rapport à ceux du concept source, définit la cohérence sémantique de manière qualitative.

## 5.4 Résultats

**Résultats globaux**
En comparant les ensembles de descendants pour un concept source donné du Metathesaurus obtenus par les approches naïve et formelle respectivement, nous avons identifié quatre cas distincts, présentés dans le tableau 1.

1) Les ensembles de descendants sont tous les deux vides (le concept source est une feuille).
2) Les ensembles de descendants sont identiques.
3) Le parcours du graphe a été interrompu après avoir atteint plus de 50 niveaux de profondeur[6] (avec l'approche naïve). L'ensemble de descendants enregistrés est incomplet.
4) Les ensembles de descendants sont complets et différents. L'analyse plus approfondie des différences concerne ce groupe.

*Tableau 1 – Catégories de concepts du Metathesaurus en fonction des différences existant parmi leurs descendants obtenus par les approches naïve et formelle*

| Catégorie | Nombre de concepts sources | |
|---|---|---|
| Aucun descendant | 765 811 | (75,0 %) |
| Mêmes descendants | 221 641 | (21,7 %) |
| Incomplet (interrompu) | 6 830 | (0,7 %) |
| Descendants différents | 26 584 | (2,6 %) |
| *Total* | *1 020 866* | *(100,0 %)* |

**Nombre de descendants**
Nous considérons uniquement les 26 584 concepts sources dont les ensembles de descendants sont complets et présentent des différences. Les caractéristiques statistiques du nombre de descendants obtenu pour chaque approche sont résumées dans le tableau 2. Le nombre de descendants est toujours plus grand avec l'approche naïve. Plus précisément, quel que soit le concept source, l'ensemble de descendants obtenus avec l'approche formelle est inclus dans celui calculé par l'approche naïve. De plus, cette dernière a tendance à identifier environ 75% de descendants en plus par rapport à l'approche formelle. Le concept dont le nombre de descendants est le plus grand est *Cyclic compound* (C0596399); la différence la plus importante entre les deux approches correspond au concept *Chemical bonding* (C0596307).

*Tableau 2 – Nombre de descendants obtenus par les approches naïve et formelle (minimum, maximum, médiane et moyenne)*

| Approche | Min. | Max. | Med. | Moy. |
|---|---|---|---|---|
| **Naïve** | 1 | 102 161 | 58 | 1112,4 |
| **Formelle** | 0 | 100 333 | 22 | 639,0 |
| **Diff. (N-F)** | 1 | 59 416 | 13 | 473,4 |

[6] Pour éviter de construire de larges graphes dus à des relations erronées, nous limitons la profondeur maximum à 50 niveaux, sachant qu'une telle profondeur n'est jamais atteinte dans le DAG du Metathesaurus.

**Cohérence sémantique : aspects quantitatifs**

**Types sémantiques**. Parmi les 26 584 concepts présentant des différences dans le nombre de descendants calculés par les deux méthodes, 14 787 (56%) présentent également des différences dans les types sémantiques représentés dans les ensembles de descendants. En d'autres termes, dans 44% des cas, les descendants supplémentaires obtenus avec l'approche naïve ont les mêmes types sémantiques que les descendants calculés avec l'approche formelle. Le nombre supplémentaire de types sémantiques varie entre 1 et 68 (médiane = 2). En moyenne, l'approche naïve identifie 49% de types sémantiques en plus parmi les descendants par rapport à l'approche formelle.

**Groupes sémantiques**. Seulement 8 256 (31%) des 26 584 concepts présentent des différences dans les groupes sémantiques représentant les ensembles de descendants. Le nombre de groupes sémantiques supplémentaires varie entre 1 et 11 (médiane = 1). En moyenne, l'approche naïve identifie 127% de groupes sémantiques en plus dans les descendants par rapport à l'approche formelle.

**Cohérence sémantique : aspects qualitatifs**

**Types sémantiques**. Pour les 14 787 concepts présentant des types sémantiques supplémentaires représentés dans les ensembles de descendants calculés par l'approche naïve, les types sémantiques additionnels sont compatibles avec celui (ou ceux) du concept source dans seulement 11% des cas.

**Groupes sémantiques**. Pour les 14 787 concepts présentant des types sémantiques supplémentaires représentés dans les ensembles de descendants calculés par l'approche naïve, seuls 27% de ces concepts appartiennent au même groupe sémantique que le concept source.

## 5.5 Exemple

Nous utilisons le concept *Generally contracted pelvis in pregnancy, labour, and delivery* (C0156969) pour illustrer les différences observées dans les ensembles de descendants obtenus par les deux approches (Figure 7). Les types sémantiques de ce concept sont *Acquired Abnormality* et *Disease or Syndrome*.

Avec l'approche formelle, C0156969 a deux descendants : *Generally contracted pelvis, delivered* (C0156971), catégorisé comme *Disease or Syndrome* et *Generally contracted pelvis, antepartum* (C0156972), catégorisé comme *Acquired Abnormality* et *Disease or Syndrome*. L'approche naïve identifie quatre descendants supplémentaires pour C0156969 : *Generally contracted pelvis, unspecified as to episode of care in pregnancy* (C0156970) catégorisé comme *Acquired Abnormality* et *Disease or Syndrome* et ses trois enfants : *Small pelvic bone* (C0426852), catégorisé comme *Finding, Midpelvic contraction* (C0405009) et *Pelvic disproportion* (C0558374), tous les deux catégorisés comme *Anatomical Abnormality*. La relation existant entre C0156969 et C0156970 a été éliminée par l'approche formelle à cause de la présence du mot "unspecified" dans l'un des termes (du vocabulaire ICD9CM[7]) constituant le concept [4].

Les descendants directs de C0156969 sont cohérents et compatibles avec le concept source car les deux types sémantiques représentés dans ce groupe sont les mêmes que ceux du concept source. Par contre, les types sémantiques additionnels catégorisant les descendants de niveau 2 incluent *Anatomical Abnormality* et *Finding*, qui ne sont pas des descendants des types sémantiques du concept source. La cohérence sémantique de ces descendants est donc faible car elle existe uniquement au niveau des groupes sémantiques (en effet, les six descendants de C0156969 appartiennent au groupe sémantique ***Disorders***).

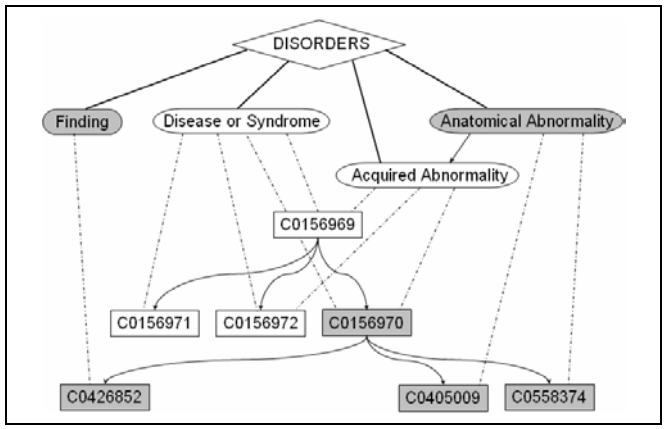


*Figure 7 – Les descendants, types sémantiques, et groupe sémantique de C0156969 (les éléments grisés sont spécifiques à l'approche naïve)*

# 6 Discussion et conclusion

## 6.1 Avantages et inconvénients de chaque approche

Cette étude vérifie en partie nos hypothèses. L'approche formelle réduit le nombre de descendants par rapport à l'approche naïve. Nous avons vu que l'approche naïve obtient tous les descendants de l'approche formelle, plus certains descendants qui lui sont propres. L'approche formelle améliore aussi, mais assez modestement, la cohérence sémantique des ensembles de descendants par rapport à l'approche naïve. Ces résultats sont contrastés par le fait qu'ils ne s'appliquent qu'à 12% des concepts avec descendants, c'est-à-dire 2,6% de l'ensemble des concepts de l'UMLS.

Nous avons également démontré que l'approche formelle sélectionne de manière systématique et cohérente les relations qui doivent être supprimées. Au contraire, c'est l'ordre dans lequel le graphe est parcouru qui détermine quels liens sont ignorés avec l'approche naïve. Il faut également noter que seule l'approche formelle est reproductible. Finalement, nous avons mis en évidence qu'en pratique, elle nécessite moins de ressources pour cons-

[7]http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm

truire les ensembles de descendants et, plus généralement, pour parcourir le graphe du Metathesaurus de l'UMLS. Avec l'approche naïve, des profondeurs de plus de 50 niveaux sont au contraire assez communes à cause des relations hiérarchiques non filtrées dans le Metathesaurus, produisant des graphes souvent plus larges, complexes et ainsi plus difficiles à exploiter.

## 6.2 Minimalité vs. Sémantique

La révision de connaissance est une notion introduite lors de la création incrémentale ou bien la mise à jour d'une base de connaissance [6]. Un certain nombre de caractéristiques sont attendues pour effectuer une révision efficace dans une base de connaissance, notamment la préservation du maximum de connaissance aussi définie comme « minimalité ». Notre travail peut s'apparenter à une révision focalisée sur l'élimination de cycles dans la base de connaissance UMLS. Mais ici, le critère de minimalité n'est pas toujours considéré. En effet, quand une relation est impliquée dans un cycle direct mais aussi dans un cycle indirect, l'approche formelle supprime la relation causant le cycle direct indépendamment de son implication dans le cycle indirect, dérogeant parfois au principe de minimalité. Par exemple (Figure 8), si un cycle indirect existe entre trois concepts C1, C2 et C3 ainsi qu'un cycle direct entre C1 et C3, la prise en compte du critère de minimalité conduirait à éliminer la relation *C3 est-père-de C1*, supprimant ainsi les deux cycles. Au contraire, notre approche qui considère d'abord le cycle direct, est susceptible de supprimer la relation *C1 est-père-de C3* et ainsi de préserver le cycle indirect. Il faut cependant préciser que l'approche formelle a été implémentée ainsi pour privilégier l'aspect sémantique à l'aspect structurel. En effet, on préfère considérer le fait qu'une relation donnée a plus de signification (ou est plus exacte) que sa relation réflexive pour déterminer quelle relation supprimer dans un cycle. Il faut ajouter que c'est aussi pour des raisons pratiques que les cycles directs sont éliminés avant d'identifier les cycles indirects. En effet, l'élimination (automatisée) des cycles directs réduit très largement la complexité des graphes d'ascendants en éliminant un grand nombre de cycles indirects auxquels les relations incorrectes impliquées dans des cycles directs auraient pu participer si elles n'avaient pas été préalablement supprimées.

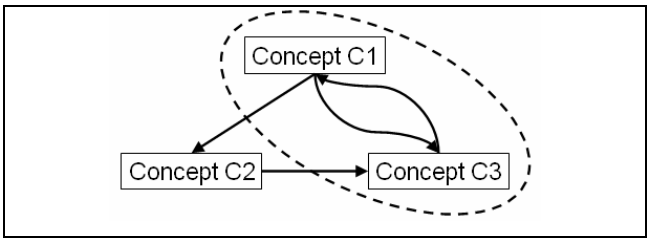


*Figure 8 – Cycles direct et indirect entre les concepts C1, C2 et C3, le cycle direct (entouré en pointillé) étant traité prioritairement par l'approche formelle*

Par contre, la phase effectuée manuellement par l'expert suit le critère de minimalité. En effet, si un cycle est identifié entre les concepts C1, C2 et C3 et un autre entre les concepts C2, C3 et C4 (Figure 9), l'expert choisit d'éliminer la relation *C2 est-père-de C3*, supprimant ainsi les deux cycles.

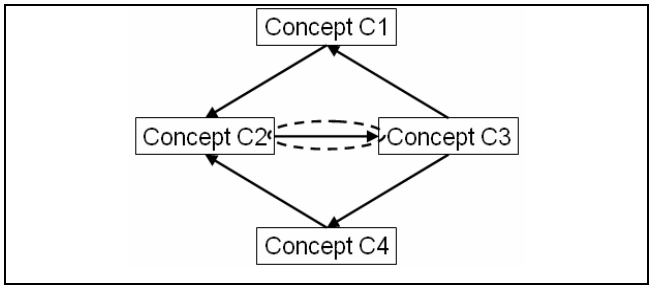


*Figure 9 – Suppression de la relation entre C2 et C3 pour éliminer les deux cycles présents dans ce graphe*

Le critère de minimalité n'est donc pas appliqué systématiquement dans le cadre de l'approche formelle parce que cette dernière privilégie l'aspect sémantique inhérent des relations impliquées dans les cycles. L'objectif est de favoriser plutôt la suppression de relations sémantiquement erronées à l'origine de problèmes structurels.

## 6.3 Applications

La définition d'approches permettant de supprimer les cycles dans des systèmes terminologiques est indispensable pour de nombreuses applications. C'est le cas de celles qui exploitent la hiérarchie des terminologies et requièrent des graphes acycliques.

**L'expansion de requêtes** en recherche d'information en est un exemple concret. Elle permet d'élargir la recherche classique d'un terme donné à des termes proches de celui-ci [7], notamment hiérarchiquement. En effet, un utilisateur qui voudrait retrouver des documents indexés par le terme *T* pourrait également être intéressé par les documents indexés par les enfants de *T*, et plus généralement ses descendants. Dans ce cas, le calcul de l'ensemble de descendants de *T* correspond à la fermeture transitive de la relation hiérarchique, et nécessite de disposer d'un graphe acyclique, comme nous l'avons souligné auparavant.

L'utilisation de l'approche naïve mènerait à des résultats aléatoires, alors que l'on veut pouvoir garantir la reproductibilité des résultats dans le cadre d'applications nécessitant le calcul de fermetures transitives. En effet, reprenons l'exemple de la partie 4.1 et supposons que l'on souhaite faire une expansion de requêtes à partir du concept UMLS *Desire for food*. L'approche naïve propose dans ce cas, un certain nombre de concepts descendants incluant *Appetite Regulation* et son fils *Food Intake Regulation* (Figure 5). Mais si on réalise une recherche à partir du concept *Food Intake Regulation,* les concepts descendants proposés ne devraient pas comprendre ses deux ascendants calculés précédemment. Pourtant, en considérant ce concept comme concept d'origine,

l'approche naïve obtient *Desire for food* et son fils *Appetite Regulation* parmi son ensemble de descendants. Au contraire, l'approche formelle supprime systématiquement la relation *Food Intake Regulation est-père-de Desire for food* garantissant qu'on obtient des ensembles de descendants cohérents dans les deux cas. Cette approche permet donc de calculer des ensembles de descendants stables et assure ainsi la reproductibilité des résultats.

Le projet *Indexing Initiative* [2], développé à la National Library of Medicine aux Etats-Unis, vise à automatiser au maximum la phase d'indexation des articles contenus dans la base de données MEDLINE. L'un des algorithmes développés pour ce projet (*restrict To MeSH* [3]) recherche les termes du vocabulaire MeSH associés à chaque concept de l'UMLS. Une des étapes de cet algorithme exploite la hiérarchie de l'UMLS pour trouver des termes MeSH parmi les ascendants du concept UMLS d'origine. Là aussi, on voit clairement la nécessité de disposer d'une version acyclique du graphe du Metathesaurus de l'UMLS.

## 6.4 Limitations

**Approche formelle ou heuristique ?**

Les cycles causés par des relations circulaires directes et réflexives sont supprimés de manière automatique par les règles et les heuristiques de l'approche formelle. Il reste malgré tout certains cas complexes nécessitant une intervention manuelle de la part d'un expert. À ce titre, on pourrait plutôt qualifier cette approche de heuristique. Il est cependant important de préciser que les cas qui sont traités manuellement suivent également des règles qui pourraient être appliquées automatiquement. Par exemple, comme nous l'avons décrit au paragraphe 6.2, une relation impliquée dans au moins deux cycles indirects sera éliminée en priorité du graphe par l'expert. De même, les relations de type *est-fils-de/est-plus-spécifique-que* reliant les concepts UMLS contenant le qualificatif « NEC » (Not Elsewhere Classified) avec des concepts plus spécifiques sont inappropriées et donc systématiquement supprimées à la main. La raison essentielle pour laquelle ces cas sont traités manuellement est qu'ils sont relativement peu nombreux (quelques dizaines après suppression des cycles directs) par rapport au travail assez lourd d'implémentation des règles nécessaires pour les traiter automatiquement. Au total, on gardera donc le qualificatif de « formel » pour cette approche qui, même si elle nécessite parfois une intervention manuelle, reste basée sur des principes appliqués de manière systématique. Ici, toutefois, « formelle » prend un sens différent de celui donné dans d'autres approches pour l'étude des cycles terminologiques (par exemple [13]). Ces dernières reposent sur une représentation formelle des terminologies dans des langages proches des logiques de description et ne sont pas applicables aux terminologies des l'UMLS pour lesquelles le niveau de formalisation des relations est très largement inférieur.

**Intégration terminologique**

Parmi les avantages qu'elle a sur l'approche naïve, nous avons démontré que l'approche formelle filtre de nombreux descendants illégitimes. Toutefois, la compatibilité sémantique des descendants qui restent est loin d'être complète. En effet, 59% des descendants obtenus par l'approche formelle pour les 26 584 concepts considérés en détail dans cette étude sont incompatibles au niveau des types sémantiques avec leur concept source respectif. Par exemple, les descendants de *Accidents* (C0000924), catégorisé comme *Phenomenon or Process* (groupe sémantique ***Phenomena***), incluent le concept *Accident prevention* (C0000918), catégorisé comme *Therapeutic or Preventive Procedure* (groupe sémantique ***Procedures***) et qui est donc sémantiquement incompatible avec le type sémantique du concept source. Cette relation non hiérarchique dans son vocabulaire d'origine n'est filtrée par aucune des deux approches dans la mesure où elle n'introduit pas d'incohérence structurelle dans le graphe. En d'autres termes, notre méthode est efficace pour s'assurer que les relations hiérarchiques impliquées dans des cycles sont supprimées du Metathesaurus de manière consistante. Cependant, pour que le Metathesaurus soit cohérent non seulement structurellement (i.e. défini sous la forme d'un DAG), mais aussi sémantiquement, une analyse sémantique de toutes les relations hiérarchiques serait nécessaire [11], ce qui dépasse largement le cadre de cette étude.
La présence de telles relations reflète la cohérence sémantique limitée du Metathesaurus en général. Il faut malgré tout noter que la relation *parent/enfant* existant entre *Accidents* et *Accident prevention* dans le système MeSH, quoiqu'ontologiquement incorrecte, joue un rôle bénéfique dans un contexte de recherche d'information. En effet, les utilisateurs intéressés par des documents sur les accidents risquent d'être intéressés par des documents relatifs à leur prévention. Cet exemple illustre la différence entre les systèmes terminologiques et les systèmes de représentation des connaissances. Plus généralement, il est incorrect de considérer le Metathesaurus comme une ontologie du domaine biomédical, bien que cette idée soit fréquemment avancée, puisque c'est en fait le produit d'une intégration terminologique, dans lequel les relations ne sont, à dessein, filtrées ni structurellement, ni sémantiquement.

En conclusion, l'approche naïve est simple à mettre en œuvre mais ne garantit pas la pertinence sémantique des relations supprimées. L'approche formelle améliore la cohérence sémantique des ensembles de descendants, mais elle est complexe à mettre en œuvre et requiert un expert du domaine. Enfin, il faut noter que la cohérence sémantique des ensembles de descendants reste imparfaite, même avec l'approche formelle, ce qui traduit une limitation de l'intégration terminologique dans l'UMLS.

## Remerciements

Ce travail a été financé en partie par la Région Bretagne (PRIR).

## Bibliographie


[1] Aho A, Ullman J. Concepts fondamentaux de l'Informatique. DUNOD, Paris, 1993:493-500

[2] Aronson AR, Bodenreider O, Chang HF, Humphrey SM, Mork JG, Nelson SJ, Rindflesch TC, Wilbur WJ. The NLM Indexing Initiative. Proc AMIA Symp 2000:17-21

[3] Bodenreider O, Nelson SJ, Hole WT, Chang HF. Beyond synonymy: exploiting the UMLS semantics in mapping vocabularies. Proc AMIA Symp 1998:815-819

[4] Bodenreider O. Circular hierarchical relationships in the UMLS: etiology, diagnosis, treatment, complications and prevention. Proc AMIA Symp 2001:57-61

[5] Cimino JJ. Auditing the Unified Medical Language System with semantic methods. J Am Med Inform Assoc 1998;5(1):41-51

[6] Crampé I, Euzenat J. Révision interactive dans une base de connaissance à objets, Actes RFIA, Rennes, 1996:615-623

[7] Efthimiadis EN. Query expansion. In Williams ME, Annual Review of Information Systems and Technology (ARIST), 1996:31:121-187

[8] Hahn U, Schulz S. Boosting the Medical Knowledge Infrastructure — A Feasibility Study on Very Large Terminological Knowledge Bases. Proc Symp on Engineering of Intelligent Systems 2004

[9] Lindberg DA, Humphreys BL, McCray AT. The Unified Medical Language System. Methods Inf Med; 1993;32(4):281-291

[10] McCray AT, Burgun A, Bodenreider O. Aggregating UMLS semantic types for reducing conceptual complexity. Medinfo 10(Pt 1):216-220

[11] McCray AT, Bodenreider O. A conceptual framework for the biomedical domain. In: Green R, Bean CA, Myaeng SH, editors. The semantics of relationships: an interdisciplinary perspective. Boston: Kluwer Academic Publishers; 2002. p. 181-198

[12] Mendonca EA, Cimino JJ, Campbell KE, Spackman KA. Reproducibility of interpreting "and" and "or" in terminology systems. Proc AMIA Symp. 1998;:790-4.

[13] Nebel B. Terminological cycles: semantics and computational properties. In J. F. Sowa, editor, Principles of Semantic Networks. Morgan Kaufmann, Los Altos, 1991:331--361

[14] Névéol A, Soualmia LF, Douyère M, Rogozan A, Thirion B, Darmoni SJ. Using CISMeF MeSH "Encapsulated" Terminology and a Categorization Algorithm for Health Resources. International Journal of Medical Informatics 2004:73(1);57-64

[15] Pisanelli DM, Gangemi A, Steve G. An ontological analysis of the UMLS Methathesaurus. Proc AMIA Symp 1998:810-814

[16] Zweigenbaum P. Encoder l'information médicale : des terminologies aux systèmes de représentation des connaissances. Innovation Stratégique en Information de Santé, (2-3):27-47, 1999

[17] Zweigenbaum P. L'UMLS entre langue et ontologie : une approche pragmatique dans le domaine médical. Revue d'Intelligence Artificielle, 2004:18:111-137