

# Utilisation de connaissances supplémentaires pour la découverte de mappings dans le système *TaxoMap*

C. Reynaud, B. Safar

*Université Paris-Sud, CNRS (L.R.I.) & INRIA (Futurs), Orsay, France*  
{chantal.reynaud, brigitte.safar}@lri.fr

## 1 Introduction

La recherche de documents pertinents sur le Web est une tâche encore souvent laborieuse. Le Web sémantique devrait faciliter ce travail en réalisant un appariement sémantique entre les requêtes des utilisateurs et les documents auxquels sont associées des méta-données issues d'ontologies décrivant leur contenu. Notre travail contribue à faciliter cet appariement en proposant des techniques de mise en correspondance entre ontologies, mises en œuvre dans le système *TaxoMap* (Kéfi (2006)).

Les schémas en entrée du processus de mise en correspondance sont des taxonomies. Une taxonomie  $(C, H_C)$  comprend un ensemble de concepts  $C$  et une hiérarchie de subsomption entre concepts  $H_C$ . Un concept est ainsi défini uniquement par son label et les relations de sous-classes qui le relie à d'autres concepts.

L'approche que nous proposons est particulièrement bien adaptée aux taxonomies ayant les caractéristiques suivantes :

- des taxonomies dissymétriques : une taxonomie étant très peu structurée, voire pas du tout (simple ensemble de termes), alors que la seconde l'est davantage.
- des taxonomies comportant exclusivement des relations de subsomption entre concepts.
- des taxonomies très spécifiques comportant des descriptions très fines de domaines d'application, ce qui se traduit au niveau des concepts par : (1) des concepts appartenant à un même domaine circonscrit, (2) des labels correspondant à des expressions composées de plusieurs mots, (3) des labels des concepts généraux inclus dans les labels des concepts plus spécifiques.

Le processus d'alignement a pour objectif de mettre en correspondance des concepts de la taxonomie la moins structurée, la taxonomie source ( $T_{Source}$ ), avec les concepts de la taxonomie la plus structurée, la taxonomie cible ( $T_{Cible}$ ). Ce processus génère des mappings 1-1 qui sont des relations d'équivalence ou des relations de sous-classes. Une relation d'équivalence  $isEq$  est un lien entre un concept dans  $T_{Source}$  et un concept dans  $T_{Cible}$  dont les noms sont considérés comme étant similaires. Les relations de spécialisation  $isA$  sont des liens usuels de sous-classe/super-classe.

Cette approche a des applications dans le contexte du Web. A titre d'illustration, nous présenterons lors de l'exposé deux cas d'utilisation :

1. L'interrogation élargie de portails Web .

## Utilisation de connaissances de support

Le système *TaxoMap* peut être utilisé pour aligner la taxonomie associée à des documents externes ( $T_{Source}$ ) avec celle d'un portail Web ( $T_{Cible}$ ) de façon à augmenter le nombre de documents accessibles à partir de ce portail. La recherche de documents s'appuie en général sur des ontologies très simples, des taxonomies. Les taxonomies de concepts sur lesquelles se basent les portails Web sont en général bien structurées, celles des autres documents auxquels on souhaiterait avoir accès via ce portail ne le sont pas toujours.

### 2. L'annotation sémantique de documents Web

L'approche peut être utile pour lier des termes isolés, extraits de documents Web, à ceux d'une ontologie, et générer des annotations sémantiques basées sur cette ontologie.

Après avoir décrit l'approche générale mise en œuvre dans *TaxoMap*, nous présenterons tout d'abord les aspects favorisant son application à grande échelle. Deux caractéristiques seront plus particulièrement mises en avant.

- l'application séquentielle de différentes techniques (terminologiques et structurelles) complémentaires, exécutables indépendamment les unes des autres. L'ordre d'exécution a été déterminé de façon à ce que le processus d'alignement soit le plus efficace possible, étant donnée la nature des taxonomies rapprochées (Kefi et al. 2006). L'indépendance des techniques autorise cependant la sélection de celles qui sont les plus appropriées dans un contexte d'utilisation donné, ou en fonction des caractéristiques des taxonomies à rapprocher (domaine couvert, longueur des labels, ...).

- la proposition de techniques différentes par rapport à leur degré d'automatisation. Les techniques que nous proposons se décomposent en deux catégories par rapport aux critères de précision et de rappel. Les techniques de la première catégorie permettent la découverte de mappings probables, i.e les éléments dont les mappings ont de fortes chances d'être pertinents. Leur précision est très grande au détriment du rappel (qui reste malgré tout significatif) (Reynaud et Safar (2006b)). Les techniques de la seconde catégorie génère des mappings dont le rappel est meilleur mais la précision inférieure. Cette distinction permet de considérer que les mappings proposés par la 1<sup>ère</sup> catégorie sont acceptables tels quels sans être validés par un expert. Le processus d'exécution peut-être complètement automatisé. En revanche, les mappings générés par la seconde catégorie de techniques correspondent plus à des suggestions de mappings potentiels faites à l'utilisateur. Le processus de découverte de cette catégorie de mappings doit être complété par une phase de validation à la charge de l'expert. Le recours à ces techniques dépend de la possibilité ou non d'effectuer cette phase de validation.

Nous nous focaliserons ensuite sur une des techniques de *TaxoMap*, spécifique au rapprochement de modèles ontologiques pauvres sémantiquement. Plus précisément, nous traiterons de l'utilisation de connaissances supplémentaires pour pallier l'insuffisance de connaissances contenues dans les modèles alignés. Nous présenterons la technique STR<sub>w</sub> de *TaxoMap* basée sur l'exploitation de la structure de WordNet. Nous effectuerons une analyse comparative de cette technique par rapport à l'état de l'art, en particulier par rapport aux travaux récemment réalisés par Sabou et al. (2006) et Aleksovski et al. (2006a, 2006b).

## 2 Utilisation de connaissances de support

Pour identifier des mappings entre les concepts de deux ontologies,  $O_{Src}$  et  $O_{Tar}$ , de nombreux travaux portent actuellement sur l'utilisation de connaissances supplémentaires dites de "background" ou de support, représentées le plus souvent sous la forme d'une 3<sup>ème</sup> ontologie,  $O_{BK}$  (voir Aleksovski et al., 2006a, 2006b, Sabou et al., 2006, Reynaud et Safar, 2006a, 2006b). L'objectif de ces travaux est de compléter les techniques classiques d'appariement qui exploitent la structure ou la richesse du langage de représentation des ontologies, et qui ne s'appliquent plus quand les ontologies à apparier sont faiblement structurées ou se limitent à des simples hiérarchies de classification.

Nous présentons tout d'abord, de façon globale, l'approche générale commune à ces différents travaux et nous reviendrons ensuite sur les aspects qui les différencient.

### 2.1 Approche générale

Pour identifier l'existence d'un mapping de la forme ( $X_{Src}$  relation  $Y_{Tar}$ ) où  $X_{Src} \in O_{Src}$ ,  $Y_{Tar} \in O_{Tar}$ , et  $relation \in T$ , l'ensemble des relations exprimables entre deux concepts dans les ontologies considérées, l'approche générale suivie par ces différents travaux se décompose en 2 phases : l'ancrage et la dérivation.

L'**ancrage** consiste tout d'abord à apparier chacun des 2 concepts  $X_{Src}$  et  $Y_{Tar}$ , pris indépendamment l'un de l'autre, avec un ou des concepts de la 3<sup>ème</sup> ontologie ( $O_{BK}$ ), c'est-à-dire, à identifier des mappings de la forme ( $X_{Src}$  relation  $X_{BK}$ ) et ( $Y_{Tar}$  relation  $Y_{BK}$ ) où  $X_{BK}$  et  $Y_{BK} \in O_{BK}$  et sont appelés des *ancres* ou *points d'ancrage*.

La **dérivation** consiste ensuite à s'appuyer sur la structuration de  $O_{BK}$  pour rechercher s'il existe des relations entre les différents points d'ancrage identifiés, puis essayer d'en dériver des relations entre les éléments des ontologies à apparier.

L'ensemble  $T$  des relations utilisées dans ces différents travaux est l'ensemble  $\{\leq, \geq, =\}$  où  $X \leq Y$  peut se lire suivant les cas comme « $X$  isA  $Y$ », « $X$  part-of  $Y$ » ou plus généralement « $X$  narrower-than  $Y$ » et les mappings ( $X_{Src}$  relation  $Y_{Tar}$ ) cherchés sont dérivés en utilisant des règles de la forme :

Si ( $X_{Src} \leq X_{BK}$ ) et ( $X_{BK} \leq Y_{BK}$ ) et ( $Y_{BK} \leq Y_{Tar}$ ) alors dériver ( $X_{Src} \leq Y_{Tar}$ )

Si ( $X_{Src} \geq X_{BK}$ ) et ( $X_{BK} \geq Y_{BK}$ ) et ( $Y_{BK} \geq Y_{Tar}$ ) alors dériver ( $X_{Src} \geq Y_{Tar}$ ).

Ces règles utilisent aussi la relation d'équivalence,  $=$ , en considérant que l'existence d'une relation de type  $A = B$  permet de rajouter les deux relations  $A \leq B$  et  $A \geq B$  et qu'inversement, le fait d'avoir pu dériver les deux relations  $X_{Src} \leq Y_{Tar}$  et  $X_{Src} \geq Y_{Tar}$  permet de dériver la relation  $X_{Src} = Y_{Tar}$ .

Si l'on considère que l'appariement d'ontologies est une fonction sur 2 ontologies qui retourne un ensemble de relations entre leurs concepts,  $f : (O_1, O_2) \rightarrow \{(X \text{ relation } Y) \mid X \in O_1, relation \in T, Y \in O_2\}$ , l'approche générale suivie par ces différents travaux revient donc à faire globalement trois appariements d'ontologie. En effet, la phase d'ancrage comporte

Utilisation de connaissances de support

deux appariements d'ontologies  $f(O_{Src}, O_{BK})$  et  $f(O_{Tar}, O_{BK})$  et la phase de dérivation, un appariement d'une ontologie sur elle-même  $f(O_{BK}, O_{BK})$ .

Pour effectuer la phase d'ancrage vers les éléments de  $O_{BK}$ , les auteurs s'appuient sur des heuristiques terminologiques simples qui travaillent sur les labels et les synonymes des termes désignant des concepts. Par exemple, en utilisant une mesure de type *edit-distance* et en considérant que si les labels de deux concepts ne se différencient pas par plus de deux caractères, les concepts considérés peuvent être reliés par une relation d'équivalence ou en utilisant une heuristique d'inclusion de labels qui consiste à dire que si tous les mots du label ou du synonyme d'un concept A se trouvent dans le label ou le synonyme d'un concept B, alors B sera considéré comme plus spécialisé que A ( $B \leq A$ ).

A partir de ce schéma général, les travaux se différencient sur les caractéristiques des ontologies employées comme support et sur les stratégies de mise en œuvre des deux phases.

## 2.2 Les travaux d'Aleksovski et al.

L'idée de base qui sous-tend ces travaux est que l'ontologie de support  $O_{BK}$  est plus complète et plus détaillée que les deux ontologies à rapprocher, et qu'elle contient une description en compréhension du domaine des 2 autres.

Les deux phases d'ancrage et de dérivation sont réalisées globalement : l'ancrage consiste tout d'abord à essayer d'apparier chacun des concepts des 2 ontologies initiales ( $O_{Src}$  et  $O_{Tar}$ ) avec les concepts de la 3<sup>ème</sup> ( $O_{BK}$ ). La dérivation consiste ensuite à rechercher au sein de  $O_{BK}$  les relations qui existent entre les différents points d'ancrage identifiés, puis d'en dériver des relations entre les éléments des ontologies à apparier. De multiples points d'ancrage pouvant être trouvés pour un concept (certains trivialement faux d'ailleurs, quand l'heuristique d'inclusion de labels est employée brutalement sans garde-fou), si les points d'ancrage de 2 concepts X et Y sont reliés de façons similaires, cela renforce le lien entre les deux concepts. S'ils sont reliés de façon incompatible (ex  $X^1_{BK} \leq Y^1_{BK}$  et  $X^2_{BK} \geq Y^2_{BK}$ ) aucune relation de subsomption ne peut être inférée entre X et Y, mais cela révèle quand même que les 2 concepts ont un lien avec une certaine proximité sémantique.

Les auteurs soulignent le fait que des concepts issus de 2 ontologies différentes sont rarement équivalents mais partagent en fait un certain recouvrement sémantique, et qu'identifier de tels recouvrements est utile dans la tâche d'intégration.

Dans Aleksovski et al. (2006a), les concepts à rapprocher sont des éléments issus de 2 listes de vocabulaires plats, non structurés. L'ontologie  $O_{BK}$  utilisée pour rechercher les dérivations est une ontologie représentant des points de vue multiples (ou aspects), ce qui permet d'identifier plusieurs dérivations entre 2 points d'ancrage, suivant les différents aspects. Un ensemble de mappings (Gold Standard) a été élaboré avec le concours manuel d'un expert. Puis, les auteurs ont réalisé 2 expérimentations : l'une, appelée lexical matching (LM), en recherchant directement des appariements entre les termes de  $O_{Src}$  et  $O_{Tar}$ , l'autre (Semantic Matching, SM) en recherchant d'abord les ancrages dans  $O_{BK}$ , puis les dérivations entre les paires d'ancres trouvées. Rien n'est dit de très précis dans ce papier sur le type de

relation définie entre les termes d'un appariement, ni sur la façon dont on les obtient ou les combine.

Bien qu'ils aient remplacé un simple problème de recherche d'appariement (de  $O_{Src}$  vers  $O_{Tar}$ ) par un double problème (de  $O_{Src}$  vers  $O_{BK}$ , et de  $O_{Tar}$  vers  $O_{BK}$ ), les auteurs observent une amélioration de la précision des mappings obtenus. Ceci peut s'expliquer par l'existence de multiples dérivations obtenues dans  $O_{BK}$ , qui permet d'identifier des proximités sémantiques non identifiables par de simples rapprochements terminologiques.

Remarquons aussi que l'étape d'ancrage de  $O_{Tar}$  vers  $O_{BK}$  n'a pas été effectuée automatiquement mais manuellement, ce qui introduit probablement un biais, car l'étape d'ancrage des termes de  $O_{Src}$  vers  $O_{BK}$  n'apparaît pas si simple. Avec les mêmes techniques terminologiques, les auteurs ont ancré moins de termes de  $O_{Src}$  dans  $O_{BK}$  qu'ils n'avaient trouvé d'appariements directs de  $O_{Src}$  vers  $O_{Tar}$  ! (moins mais mieux car la précision est légèrement meilleure).

Remarquons aussi que les 2 expérimentations ont été effectuées de façon exclusive, sans essayer de combiner les 2 approches.

Dans Aleksovski et al. (2006b), les concepts à rapprocher appartiennent à 2 ontologies vraiment structurées par des relations du type « *X narrower-than Y* » et « *X Broader-than Y* » ( $\{\leq, \geq\}$ ) et l' $O_{BK}$  contient des relations de type *is-a* et *part-of*.

Ces 2 relations permettront d'inférer des relations de type *narrower-than*, dans la recherche de dérivation entre 2 ancres en utilisant les règles suivantes :

Si ( $X_{BK} \text{ isA } Y_{BK}$ ) alors dériver ( $X_{BK} \leq Y_{BK}$ ) et si ( $X_{BK} \text{ part-of } Y_{BK}$ ) alors dériver ( $X_{BK} \leq Y_{BK}$ ).

Les auteurs utilisent la fermeture transitive de relations :

Si ( $X^1_{BK} \text{ isA } X^2_{BK}$ ) et ( $X^2_{BK} \text{ isA } X^3_{BK}$ ) et .. et ( $X^{n-1}_{BK} \text{ isA } X^n_{BK}$ ) alors dériver ( $X^1_{BK} \leq X^n_{BK}$ ).

Cette fermeture s'applique aussi aux relations *part-of* et peut mêler les relations *isA* et *part-of* ou au contraire imposer de n'utiliser les relations *isA* qu'après avoir utilisé tous les *part-of*.

De nouvelles expérimentations sont effectuées dans ce contexte, la 1<sup>ère</sup> en recherchant directement des appariements entre les termes de  $O_{Src}$  et  $O_{Tar}$ , et les suivantes par dérivation, en utilisant ou pas la fermeture transitive de relation, et sans imposer ou en imposant des contraintes sur l'ordre d'utilisation des relations *isA* et *part-of* lors de la fermeture. Pour pallier l'absence de mappings de référence, les évaluations de ces expérimentations ont été faites en choisissant au hasard 30 concepts de  $O_{Src}$  et en évaluant manuellement la correction des relations trouvées. La dernière technique de dérivation est celle qui donne les meilleurs résultats, toutes les relations identifiées ayant été jugées correctes.

## 2.3 Les travaux de Sabou et al.

A l'opposé des travaux précédents, les auteurs considèrent qu'il n'existe pas a priori une ontologie qui soit plus complète et plus détaillée que les deux ontologies à rapprocher, et qui puisse seule servir de support. Ils proposent donc d'utiliser l'ensemble des ontologies accessibles sur le Web par l'intermédiaire du moteur de recherche sémantique Swoogle. Pour identifier l'existence d'un mapping de la forme ( $X_{Src}, \text{relation}, Y_{Tar}$ ), les auteurs proposent de

## Utilisation de connaissances de support

rechercher à la volée les ontologies qui permettent l'ancrage simultané des deux concepts à apparier, puis de chercher s'il existe une dérivation entre les deux ancres dans les ontologies considérées.

L'approche peut paraître beaucoup plus coûteuse que la précédente puisqu'elle travaille séquentiellement sur toutes les paires de concepts possibles et impose a priori de rechercher plusieurs fois l'ancrage d'un même concept dans une même ontologie, théoriquement autant de fois que de concepts avec lequel on essaye de le mettre en relation. Mais elle permet d'identifier à la volée, sans choix manuel préalable, les ontologies susceptibles de servir de background même à un seul mapping et elle est parallélisable. De plus, l'approche est présentée comme complémentaire à d'autres techniques d'appariement et n'est donc utilisée que pour les concepts qui n'ont pas pu être apparés par ces autres techniques, donc sur un nombre limité de concepts.

Si aucune ontologie ne permet l'ancrage simultané des deux concepts à apparier, l'approche précédente peut être étendue récursivement en travaillant sur plusieurs ontologies à la fois. Les auteurs proposent ainsi d'ancrer  $X_{Src}$  dans une première ontologie, puis de rechercher, pour tous les concepts  $Y_{BK}$  en relation avec l'ancre dans cette ontologie s'ils sont en relation avec  $Y_{Tar}$  dans d'autres ontologies. Même si elle peut être parallélisée, cette dernière stratégie est bien sur encore plus coûteuse que la précédente.

## 2.4 Utilisation de connaissances de support dans *TaxoMap*

Comme dans Aleksovski et al. (2006a, 2006b), nous n'utilisons à ce jour dans notre approche qu'une seule ressource support  $O_{BK}$  identifiée manuellement au préalable, en l'occurrence WordNet. Bien évidemment, les ontologies à apparier étant très spécifiques et comportant des descriptions très fines du domaine d'application, avec des concepts très spécialisés, WordNet ne peut pas être considérée, et de loin, comme plus complète et plus détaillée que ces deux ontologies. Mais notre technique d'utilisation d'une  $O_{BK}$  n'étant comme dans Sabou et al. (2006) qu'une technique complémentaire à d'autres techniques d'appariement, la plupart des appariements portant sur les concepts très spécialisés auront été identifiés par les autres techniques et cette dernière ne s'appliquera que sur les concepts de  $X_{Src}$  non encore apparés.

Par rapport aux autres travaux, du fait de notre contexte d'application, nous limitons le nombre de relations à identifier, en ne cherchant à apparier les concepts de  $O_{Src}$  qu'avec des concepts de  $O_{Tar}$  considérés comme plus généraux, i.e. nous recherchons des mappings orientés de la forme  $X_{Src} \leq Y_{Tar}$  et pas ceux de la forme  $X_{Src} \geq Y_{Tar}$ .

Notre approche est donc la suivante : nous commençons par identifier manuellement avec un expert, le concept de WordNet noté  $root_A$ , qui sera le concept le plus spécialisé de WordNet qui généralise a priori tous les concepts du domaine des ontologies à apparier (*food* dans notre exemple). Puis nous réalisons l'ancrage dans WordNet de tous les concepts de  $O_{Tar}$  et de l'ensemble des concepts de  $O_{Src}$  non encore apparés.

Notre technique se différencie sur la recherche des dérivations. Au lieu de rechercher les dérivations entre les ancres des deux ontologies, nous recherchons d'abord les dérivations

qui mènent à la racine  $root_A$  précédemment identifiée. Ces dérivations sont construites en recherchant dans WordNet les hypernymes de chacune des ancres, jusqu'à atteindre  $root_A$  ou l'une des racines de la hiérarchie WordNet. Par exemple, le résultat de la recherche sur le concept cantaloupe donne les deux ensembles de généralisants suivants qui forment deux dérivations correspondant à deux sens différents du terme :

- Sens 1: cantaloupe → sweet melon → melon → gourd → plant → organism → Living  
 Sens 2 : cantaloupe → sweet melon → melon → edible fruit → green goods → food

Seules les dérivations contenant  $root_A$  sont retenues car elles correspondent au seul sens pertinent pour l'application. Un sous-graphe,  $T_{WN}$ , composé de l'union des concepts et des relations des dérivations sélectionnées (cf. FIG. 1) est alors obtenu. Il se compose du concept racine le plus général de l'application,  $root_A$ , des feuilles correspondant aux ancres des concepts issus des deux ontologies initiales (cercles sur FIG.1) et des généralisants intermédiaires extraits de WordNet qui peuvent, ou non, appartenir à l'une des deux ontologies.

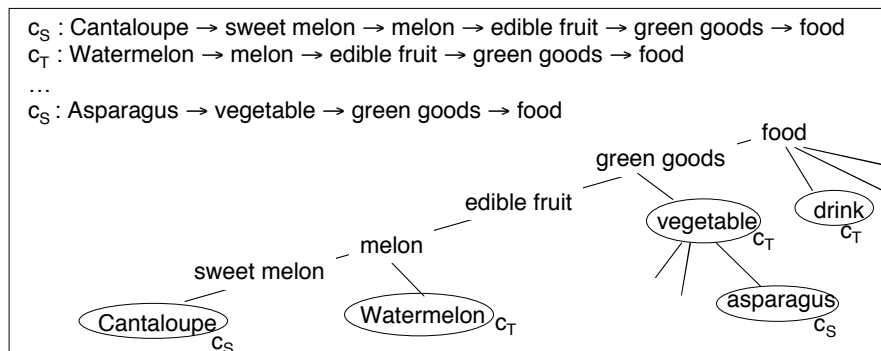


Fig 1. Un exemple de sous graphe  $T_{WN}$

L'identification des mappings pertinents est faite en recherchant pour chaque concept  $c_S$  de  $O_{Src}$  le plus proche concept  $c_T$  de  $O_{Tar}$  qui apparaisse sur une dérivation menant à la racine  $root_A$ . Ainsi à partir du sous graphe de la figure ci-dessus, on peut dériver le mapping asparagus  $isA$  vegetable. Notre technique est comparable dans ces résultats à celle mise en œuvre dans Aleksovski et al. (2006b), si ce n'est que nous ne travaillons que sur les relations de type  $isA$  de WordNet (et pas sur les liens *part-of*) et que nous ne conservons que les mappings orientés de la forme  $c_S isA c_T$ .

Remarquons qu'aucune des techniques que nous venons de présenter ne permet de dériver de mapping sémantique pour le concept cantaloupe puisque aucun de ses ancêtres n'est une ancre d'un concept de  $O_{Tar}$  (tous sont des termes intermédiaires issus de WordNet). En revanche, on aimerait bien être capable de le « rapprocher » du concept Watermelon puisque ces deux concepts sont deux sortes de melon et donc sémantiquement très proches.

Ce rapprochement peut être effectué en utilisant sur le sous graphe  $T_{WN}$ , une mesure de similarité entre nœuds d'un même graphe. Nous utilisons la mesure proposée par Wu et Palmer (1994) selon laquelle la similarité entre deux nœuds  $c_1$  et  $c_2$  est fonction de leur profondeur,  $depth(c_i)$ ,  $i \in [1,2]$ , i.e. leur distance à la racine en nombre d'arcs, et de celle de leur plus petit ancêtre commun ( $LCA$ ).

Utilisation de connaissances de support

$$Sim_{W\&P}(c_1, c_2) = \frac{2 * depth(LCA(c_1, c_2))}{depth(c_1) + depth(c_2)}$$

L'intérêt de cette mesure est double. D'une part, elle est plus précise qu'une mesure basée sur une simple distance des nœuds, car plus la profondeur du *LCA* de deux concepts est importante, plus les deux concepts partagent de caractéristiques communes et plus ils sont proches. D'autre part, une étude de ses propriétés, cf. Reynaud et Safar (2006b), nous a permis d'identifier et de mettre en œuvre une stratégie de recherche qui permet de retrouver très efficacement dans  $T_{WN}$  le concept  $c_T$  de  $O_{Tar}$  qui sera évalué comme le concept le plus similaire d'un concept donné  $c_S$  de  $O_{Src}$ .

Il est clair que les rapprochements effectués à partir d'une mesure de similarité de ce type ne permettent pas d'établir de mappings « sémantiques », c'est-à-dire reliant explicitement deux concepts par un lien de type *isA* ou *Eq*, et qui puissent être justifiés et prouvés par des mécanismes d'inférences cf. Sabou et al. (2006). Il est tout aussi clair qu'il serait dommage de ne pas exploiter l'information identifiée ! Nous proposons donc de retenir les rapprochements de ce type comme des mappings potentiels devant être validés par un expert et de les étiqueter par une nouvelle relation notée '*isClose*'.

Le choix fait dans *TaxoMap* consiste donc à rechercher pour chaque concept  $c_S$  de  $O_{Src}$  qui restait à appairer, le concept  $c_T$  de  $O_{Tar}$  qui lui est le plus similaire suivant la mesure de Wu et Palmer, qui sera noté  $c_{Sim}$ , et de construire le mapping potentiel associé  $c_S$  *isClose*  $c_{Sim}$ . Puis nous extrayons, comme nous l'avons décrit plus haut, l'ensemble des mappings sémantiques lisibles sur les branches du sous graphe  $T_{WN}$ . Si un concept  $c_{Sim}$  apparaît relié à un même concept  $c_S$  à la fois dans un mapping sémantique et dans un mapping potentiel, nous ne conservons que le mapping sémantique.

Par exemple, le concept *vegetable* de  $O_{Tar}$  étant le concept le plus similaire du concept *asparagus* de  $O_{Src}$ , nous construisons le mapping potentiel *asparagus isClose vegetable*. Mais comme nous avons aussi pu construire le mapping sémantique *asparagus isA vegetable* seul ce dernier est conservé. Comme aucun mapping sémantique ne peut être construit pour le concept *cantaloupe*, nous conserverons le mapping potentiel *cantaloupe isClose Watermelon*.

Nous avons réalisé deux expérimentations de cette technique dans le domaine de la microbiologie. Dans la première, la technique a été utilisée directement sur tous les concepts de  $O_{Sr}$  en utilisant une technique d'ancrage du type inclusion de labels et dans la deuxième, la technique a été utilisée en complément d'autres techniques (donc sur les seuls concepts non encore appariés).

La non pertinence des résultats obtenus dans la première expérimentation est largement due à la longueur des labels des concepts de notre domaine (ex : *home-style salad (reduced calorie mayonnaise with chicken)*) qui ne sont bien sûr pas reconnus directement par WordNet et qui sont incorrectement ancrés par l'heuristique d'inclusion de labels (les 3 ancres identifiés pour l'exemple précédent sont : *salad*, *mayonnaise*, *chicken*). En revanche, dans la deuxième expérimentation, comme les autres techniques de *TaxoMap* tirent justement parti de la longueur des labels pour exploiter leur similarité, la technique n'a du être appliquée que sur les seuls concepts non encore appariés, avec plus souvent des labels courts, et les résultats sont plus pertinents.

D'autres expérimentations ont été réalisées sur des taxonomies servant de test dans la communauté appariement. Elles ont montré que cette technique n'est pas adaptée pour

aligner des taxonomies dont le domaine d'application est trop large. En effet, la technique construit un sous-arbre à partir de tous les nœuds hypernymes de WordNet jusqu'à atteindre le nœud le plus général de l'application. Dans le cas d'un domaine très grand, le concept le plus général est un nœud placé très haut dans la hiérarchie WordNet, si ce n'est le nœud racine.  $T_{WN}$  est donc très gros. Il mêle des sens de termes différents et conduit à générer des mappings qui ne sont absolument pas pertinents. Des améliorations seraient possibles si plusieurs sous-arbres étaient construits, un par sous-domaine en supposant que les différents sous-domaines puissent être identifiés.

### 3 Conclusion et perspectives

Les différents travaux présentés montrent bien l'intérêt d'utiliser des connaissances supplémentaires pour la découverte automatique de mappings. La comparaison effectuée nous a permis d'identifier les similitudes et les complémentarités des différentes approches, ainsi que de possibles directions de recherche. Par exemple, nous pourrions essayer de valider les mappings potentiels identifiés dans notre approche en utilisant la technique proposée par Sabou et al. Inversement, ces mappings potentiels pourraient être utilisés dans les travaux de Sabou, dans la phase de recherche récursive pour diriger les recherches et ordonner les tests effectués.

Ces comparaisons nous ont permis de mettre en lumière quelques problèmes posés par le passage à l'échelle. Le recours à des connaissances externes est très intéressant lorsqu'on connaît précisément le contexte au sein duquel les éléments manipulés doivent être interprétés. Il est plus délicat lorsque le contexte est plus large ou inconnu au départ. Ainsi des techniques génériques, applicables quel que soit le domaine d'application et ayant recours de façon dynamique à de telles connaissances externes, sont face à un réel problème d'identification du contexte d'étude. C'est le cas lorsque la connaissance de support est obtenue via Swoogle ou lorsqu'on utilise WordNet.

Le deuxième problème identifié au travers des travaux présentés concerne le traitement des relations présentes dans l'ontologie supplémentaire utilisée. Comment tirer parti de toute la richesse des relations représentées dans une ontologie ? Doit-on s'autoriser à combiner des relations sémantiques différentes, si oui comment ? Comment interpréter les résultats obtenus suite à ces combinaisons ?

### Références

- Aleksovski, Z., Klein, M., Ten Kate, W., Van Harmelen, F. (2006a). Matching Unstructured Vocabularies using a Background Ontology, Proceedings of the 15<sup>th</sup> International Conference on Knowledge Engineering and Knowledge Management (EKAW'06), Springer-Verlag.

## Utilisation de connaissances de support

- Aleksovski, Z., Klein, M., Ten Kate, W., Van Harmelen, F. (2006b). Exploiting the Structure of Background Knowledge used in Ontology Matching. ISWC'06 Workshop on Ontology Matching (OM-2006), Athens, Georgia, USA.
- Kéfi, H. (2006). Ontologies et aide à l'utilisateur pour l'interrogation de sources multiples et hétérogènes. Thèse de doctorat de l'Université Paris-Sud.
- Kéfi, H., Safar, B., Reynaud, C. (2006). Alignement de taxonomies pour l'interrogation de sources d'information hétérogènes. RFIA. Tours.
- Lin, D. (1998). An Information-Theoretic Definition of Similarity. ICML, Madison, pp. 296-304.
- Reynaud, C., Safar, B. (2006a) When usual structural alignment techniques don't apply. ISWC '06 Workshop on Ontology Matching (OM-2006), Poster, Athens, Georgia, USA.
- Reynaud, C., Safar, B. (2006b). Structural Techniques for Alignment of Taxonomies: experiments and evaluation, In TR 1453, LRI, Université Paris-Sud, Juin 2006.
- Sabou, M., D'Aquin, M., Motta, E. (2006). Using the Semantic Web as Background Knowledge for Ontology Mapping, ISWC'06 Workshop on Ontology Matching (OM-2006), Athens, Georgia, USA.
- Wu, Z., Palmer, M. (1994). *Verb semantics and lexical selection*. Computational Linguistics. Las cruces, pp. 133-138.

## Summary

This paper deals with the alignment techniques between ontologies performed by the *TaxoMap* system. We focus on the applicability on a large scale and we compare one of our techniques with close works using background knowledge.