Techniques d'alignement d'ontologies basées sur la structure d'une ressource complémentaire

Brigitte Safar, Chantal Reynaud, François Calvier

Université Paris-Sud, CNRS (LRI, UMR 8623) & INRIA (Futurs),
91405 Orsay, France

{brigitte.safar, chantal.reynaud, francois.calvier}@lri.fr

RÉSUMÉ. *Pour identifier des mappings entre les concepts de deux ontologies, de nombreux travaux récents portent sur l'utilisation de connaissances complémentaires dites de "background" ou de support, représentées le plus souvent sous la forme d'une 3ème ontologie. Leur objectif commun 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 de simples hiérarchies de classification. Cet article présente une analyse comparative des travaux utilisant des connaissances de support, en commençant par leur schéma général commun, suivi par une analyse des travaux en fonction du type de connaissance de support utilisée. Nous étudions ensuite les problèmes rencontrés lorsque la connaissance de support est WordNet puis nous montrons comment notre système TaxoMap résout ces difficultés.*

ABSTRACT. *A lot of alignment systems providing mappings between the concepts of two ontologies rely on additional source, called background knowledge, represented most of the time by a third ontology. The common objective is to complement current matching techniques which exploit structure or features represented in ontology representation languages and which fail when ontologies are only hierarchies or weakly structured models. In this paper, we present a comparative analysis of works using background knowledge. A common general scheme is first introduced followed by an analysis of two kinds of work that differ by the kind of background knowledge they use. Then we present the difficulties encountered when using WordNet as background knowledge. Finally, we show how the* TaxoMap *system we implemented can avoid those difficulties.*

MOTS-CLÉS : *Alignement d'ontologies, ressource complémentaire.*

KEYWORDS: *Ontology Mapping, Background Knowledge*

1. Introduction

L'explosion du nombre de sources d'informations accessibles via le Web multiplie les besoins de techniques permettant l'intégration de ces sources. En définissant les concepts associés à des domaines particuliers, les ontologies sont un élément essentiel des systèmes d'intégration, car elles permettent à la fois de décrire le contenu des sources à intégrer et d'expliciter le vocabulaire utilisable dans les requêtes des utilisateurs. La tâche d'alignement d'ontologies (recherche de mappings, appariements ou mises en correspondance) est particulièrement importante dans les systèmes d'intégration puisqu'elle autorise la prise en compte conjointe de ressources décrites par des ontologies différentes. Ce thème de recherche a donné lieu à de très nombreux travaux (Shvaiko & Euzenat, 2005).

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 complémentaires, dites de "background" ou de support, représentées le plus souvent sous la forme d'une 3ème ontologie, O_{BK} (voir Aleksovski *et al.,* 2006a, 2006b, Sabou *et al.,* 2006, Kalfoglou & Hu 2005, Reynaud & Safar, 2006, 2007). 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 de simples hiérarchies de classification.

Cet article présente une analyse comparative de ces différents travaux. La section 2 introduit tout d'abord le schéma général commun, en deux étapes, l'ancrage et la dérivation, puis la section 3 analyse plus précisément deux travaux particuliers qui diffèrent sur le type de connaissances complémentaires utilisées et la stratégie de recherche de mappings qui en découle. Les difficultés rencontrées lors de l'utilisation de WordNet comme ressource complémentaire sont présentées en section 4, en particulier le problème de contresens dû aux différents sens d'un terme. La section 5 détaille la façon dont notre système d'alignement de taxonomies *TaxoMap* (Reynaud & Safar, 2007) dépasse cette difficulté, en limitant le sens des termes pris en compte lors des appariements. Les résultats expérimentaux présentés montrent le gain de précision des appariements obtenus par cette limitation du sens des termes. La section 6 conclut le papier.

2. Schéma général commun

Le processus d'alignement entre ontologies a pour objectif de mettre en correspondance les concepts d'une des ontologies, dite ontologie source (O_{Src}), avec les concepts d'une autre ontologie, dite ontologie cible (O_{Tar}). Pour simplifier la présentation générale des différents travaux référencés, nous considèrerons que chaque ontologie O ne comprend qu'un ensemble de concepts C et un ensemble de relations R entre ces concepts.

Pour identifier l'existence d'un mapping de la forme (X_{Src} *relation* Y_{Tar}) où $X_{Src} \in C_{Src}$, $Y_{Tar} \in C_{Tar}$, et *relation* $\in R$, l'ensemble des relations exprimables entre deux concepts appartenant respectivement à l'une et à l'autre des ontologies considérées, l'approche générale suivie se décompose en 2 phases : l'ancrage et la dérivation (cf. Fig.1).

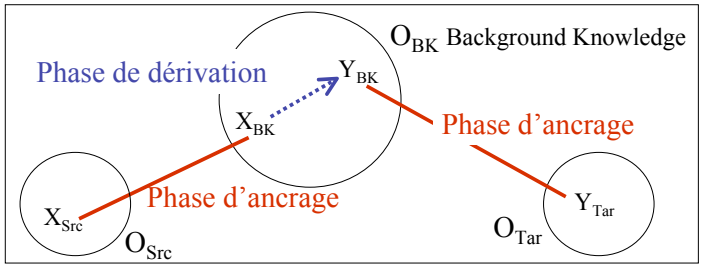

Figure 1. *Schéma général de dérivation d'un mapping (X_{Src} relation Y_{Tar})*

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 3ème 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 $Y_{BK} \in C_{BK}$ et sont appelés des *ancres* ou *points d'ancrage*.

La **dérivation** consiste ensuite à s'appuyer sur la structuration de O_{BK} pour :

- soit rechercher s'il existe des relations entre les différents points d'ancrage X_{BK}, Y_{BK} identifiés, afin d'essayer d'en dériver des relations (des mappings « sémantiques ») entre les éléments des ontologies à apparier,
- soit utiliser une mesure de similarité entre nœuds d'un même graphe, pour identifier pour chaque ancre X_{BK} d'un concept de l'ontologie source, l'ancre Y_{BK} du concept de l'ontologie cible qui lui est le plus similaire. Remarquons que la relation identifiée par cette mesure de similarité est une relation de proximité qui ne contient pas d'information sur la sémantique du lien unissant les deux concepts.

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 : $(O_1,O_2) \rightarrow \{(X \textit{ relation } Y) \mid X \in C_1, \textit{relation} \in R, Y \in C_2\}$, l'approche générale suivie par ces différents travaux revient donc à faire globalement trois appariements d'ontologies. En effet, la phase d'ancrage comporte 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 portent sur les labels et les synonymes des termes désignant les concepts. Une première heuristique utilise une mesure de type *edit-distance* et considère 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. Une deuxième heuristique s'appuie sur l'inclusion de labels et 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 stratégies de mise en œuvre des deux phases et sur les caractéristiques des ontologies employées comme support. Nous étudierons tout d'abord les travaux qui, dans la phase de dérivation, recherchent dans O_{BK} des relations entre les points d'ancrage, puis ceux qui s'appuient sur une mesure de similarité et emploient comme support la ressource WordNet.

3. Recherche de relations entre les ancres dans O_{BK}

Dans la recherche de mappings de la forme (X_{Src} *relation* Y_{Tar}), l'ensemble R des relations utilisées est l'ensemble $\{\leq, \geq, \equiv\}$ où $X \leq Y$ peut se lire, suivant les cas, *«X isA Y»*, *«X part-of Y»* ou plus généralement *«X narrower-than Y»*. Les mappings cherchés sont dérivés en exploitant des règles de la forme :

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

Ces règles utilisent aussi la relation d'équivalence, $\equiv$, en considérant que l'existence d'une relation de type $A \equiv B$ 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 $X_{Src} \equiv Y_{Tar.}$

Ces différentes règles permettent ainsi de dériver des mappings « sémantiques », i.e. des mappings reliant deux concepts par un lien de type *isA* ou *isEq* dont la sémantique est bien définie et qui peuvent être justifiés et prouvés par des mécanismes d'inférences (Sabou *et al.,* 2006).

Les travaux basés sur des règles de dérivation se différencient sur le type d'ontologie de support employée. Aleksovski dans (Aleksovski *et al.*, 2006a et 2006b), suppose que la recherche de dérivation peut s'effectuer sur une ontologie de support unique, préalablement identifiée et qui couvre a priori tous les concepts des ontologies à apparier. A l'inverse, les travaux décrits dans (Sabou *et al.,* 2006) font l'hypothèse opposée : la recherche de dérivation ne peut s'effectuer qu'au sein de multiples ontologies de support, sélectionnées dynamiquement. Les deux sections suivantes décrivent plus précisément l'approche suivie par ces travaux.

3.1. *Recherche de dérivation dans une ontologie de support unique et complète.*

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 3ème (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 de O_{Src} et O_{Tar}.

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 en recherchant directement des appariements entre les termes de O_{Src} et O_{Tar}, l'autre en recherchant d'abord les ancrages dans O_{BK}, puis les dérivations entre les paires d'ancres trouvées. Les auteurs observent une amélioration de la précision des mappings obtenus, dans la 2ème expérimentation. 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.

Dans (Aleksovski *et al.*, 2006b), les concepts à rapprocher appartiennent à 2 ontologies structurées par des relations du type « *X narrower-than Y* » et « *X Broader-than Y* » ($\{\leq, \geq\}$) et O_{BK} contient des relations de type *is-a* et *part-of*. Ces 2 relations permettent d'inférer des relations de type *narrower-than*, dans la recherche de dérivation entre 2 ancres en s'appuyant sur les règles suivantes :

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

Les auteurs utilisent la fermeture transitive de relations : Si (X^1_{BK} *isA* X^2_{BK}) et (X^2_{BK} *isA* X^3_{BK}) et .. et (X^{n-1}_{BK} *isA* X^n_{BK}) alors dériver ($X^1_{BK} \leq X^n_{BK}$), qui s'applique aussi aux relations *part-of* et peut mêler les relations *isA* et *part-of* ou au contraire imposer de n'employer les relations *isA* qu'après avoir exploité tous les *part-of*.

De nouvelles expérimentations sont effectuées dans ce contexte, la 1ere 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.

3.2. *Recherche de dérivation dans un ensemble d'ontologies de support*

A l'opposé des travaux précédents, les auteurs de (Sabou *et al.,* 2006) considèrent qu'il n'existe pas a priori, dans tous les domaines, 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'exploiter 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} *relation* Y_{Tar}), les auteurs proposent de rechercher à la volée une ontologie qui permette l'ancrage simultané des deux concepts à apparier, puis de chercher s'il existe une dérivation entre les deux ancres dans l'ontologie considérée. Si une telle dérivation n'existe pas, le processus repart dans la recherche automatique d'une nouvelle ontologie permettant l'ancrage des deux concepts.

L'approche peut paraître beaucoup plus coûteuse que la précédente puisqu'elle travaille séquentiellement sur toutes les paires de concepts possibles et qu'elle conduit a priori à répéter *n* fois la phase d'ancrage d'un même concept dans la même ontologie si on essaye de l'apparier à *n* concepts différents. Cependant, 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 appariés 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 sûr encore plus coûteuse que la précédente.

Pour conclure l'analyse de ces travaux, on remarquera qu'utiliser une unique ontologie de support permet de supposer que toutes les ancres des concepts à mettre en relation peuvent être identifiées en une passe avant d'effectuer la recherche de dérivation proprement dite. La complexité de la phase de dérivation diminue alors puisque cette phase consiste, dans ce cas, à rechercher les liens entre chaque point d'ancrage des éléments de O_{Src} et l'ensemble des points d'ancrage des éléments de O_{Tar}.

4. Utilisation de la ressource WordNet

WordNet est une ressource lexicale de langue anglaise, disponible sur internet, qui regroupe des termes (noms, verbes, adjectifs et adverbes) en ensembles de synonymes appelés *synsets*. Un synset regroupe tous les termes dénotant un concept donné. Le terme associé à un concept est représenté sous une forme lexicalisée, sans marque de féminin ni de pluriel. Les synsets sont reliés entre eux par des relations sémantiques : relation de généralisation/spécialisation (…is a kind of…), relation composant/composé (this is a part of…). Une interface d'interrogation permet à un

utilisateur de rechercher un terme *t* dans la base de WordNet et renvoie une définition en langue naturelle, ainsi que ses généralisants, ses spécialisations et les termes auxquels il est lié par une relation de composition, pour les différents sens de ce terme (les différents synsets auxquels il appartient).

WordNet peut être utilisé de différentes façons pour la recherche de mappings. Une première technique consiste comme dans (Bach *et al.,* 2004), à étendre systématiquement le label d'un concept avec les synonymes appartenant au synset de chaque terme du label dans WordNet, ce qui permet par exemple, de rapprocher « person » de « human ».

La technique la plus couramment utilisée consiste à s'appuyer sur les relations de généralisation/spécialisation pour considérer WordNet comme une hiérarchie de concepts au sein de laquelle il est possible d'utiliser une mesure pour calculer la similarité entre deux de ses nœuds. Un outil de ce type, *WordNet::Similarity* (Pedersen *et al.* 2004), est ainsi disponible sur le Web. Il permet de calculer de façon interactive une similarité numérique entre deux concepts quelconques, en choisissant la mesure de similarité employée parmi plusieurs.

Cette technique est utilisée par (Kalfoglou *et al.,* 2005) dans un de ses modules d'appariement, *WNNameMatcher*, pour identifier pour chaque ancre d'un concept de O_{Src}, l'ancre du concept de O_{Tar} qui lui est le plus similaire. La mesure de similarité employée est celle 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*).

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

Nous avons utilisé une technique basée sur la même mesure et nous avons remarqué qu'elle donnait des résultats pertinents quand les domaines d'application des ontologies à comparer étaient proches et très focalisés. En revanche, l'expérience a montré que si les domaines d'application étaient plus larges et ne se recoupaient pas, les résultats étaient beaucoup moins satisfaisants. Le problème est dû aux différents synsets auxquels un même terme peut appartenir et donc aux contresens et aux rapprochements erronés qui peuvent en découler.

Pour illustrer le problème rencontré, nous présentons Fig.2 les rapprochements identifiés lors d'expérimentations menées sur une paire de taxonomies[1] mise à la disposition de la communauté sur le site internet Ontology Matching[2] : les taxonomies Russia-A (O_{Tar}) et Russia-B (O_{Src}), qui décrivent la Russie à des fins touristiques, sa géographie, ses monuments et de plus dans Russia-B, ses moyens de transport.

---

[1] http://www.atl.external.lmco.com/projects/ontology/ontologies/russia/

[2] http://oaei.ontologymatching.org/2007/

Fig.2 représente le sous-graphe de WordNet mobilisé dans la recherche des ancres des termes de O_{Tar} (entourés dans la figure par un ovale gris) les plus proches des termes de O_{Src} (entourés d'un ovale blanc) correspondant à des véhicules. Aucun des termes de O_{Tar}, n'évoquant les moyens de transport, tous les termes de O_{Src} correspondant à des véhicules vont être rapprochés de la ville de Berlin par la technique considérée, puisque le terme « Berlin » appartient dans WordNet à 3 synsets distincts : l'un correspondant à la capitale de l'Allemagne, le second à un musicien et le troisième à la berline, une sorte de voiture.

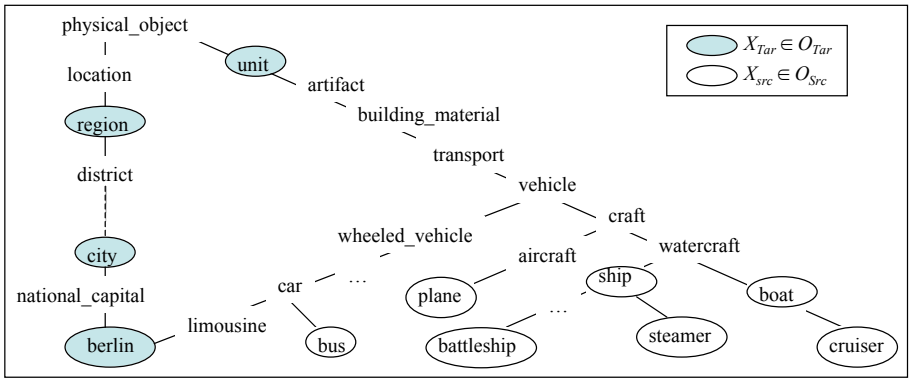

Figure 2. *Sous graphe de WordNet mobilisé dans la recherche des véhicules*

Nous montrerons dans le paragraphe suivant comment, dans le système *TaxoMap*, nous évitons ce problème en limitant les sens des termes de WordNet qui seront pris en compte dans la recherche.

5. Utilisation de WordNet comme connaissance de support dans *TaxoMap*

Comme dans les travaux d'Aleksovski, Taxomap n'utilise à ce jour qu'une seule ressource support O_{BK}, WordNet. Dans les premières expérimentations réalisées dans le domaine du risque alimentaire[3], les ontologies à apparier étant très spécifiques et comportant des descriptions très fines du domaine d'application, les concepts sont très spécialisés et la plupart ne sont pas reconnus par WordNet qui ne peut donc pas être considéré comme une ressource plus complète et plus détaillée que les deux ontologies de départ. Notre technique d'utilisation d'une ressource O_{BK} ne peut donc pas être utilisée seule comme méthode d'alignement et n'est qu'une technique complémentaire à d'autres techniques d'appariement qui seront appliquées au préalable, comme dans (Sabou *et al.,* 2006). Au moment de l'utilisation de WordNet, la plupart des appariements portant sur les concepts très

[3] Le projet e.dot (Entrepôt de Données Ouvert sur la Toile) est un projet de recherche RNTL (2003-2005).

spécialisés auront déjà été effectués par les autres techniques et le recours à O_{BK} ne s'effectuera que sur les concepts de O_{Src} non encore appariés.

Pour éviter les contresens et les rapprochements erronés dus aux multiples sens possibles d'un même terme, notre technique se décompose en deux phases. Nous commençons par extraire de WordNet, un sous arbre composé des seuls synsets correspondant aux sens supposés pertinents pour le domaine des ontologies considérées. Nous identifions ensuite, au sein de ce sous arbre, les mappings sémantiques et les rapprochements basés sur la mesure de similarité.

5.1. *Extraction du sous-arbre de WordNet pertinent pour le domaine*

Notre approche est donc la suivante. Si les domaines d'application des ontologies à apparier sont proches et ciblés, 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). Si les domaines d'application sont plus larges et distincts, nous identifions avec l'expert, plusieurs racines qui couvriront les thèmes des concepts de l'ontologie cible. Puis nous réalisons l'ancrage dans WordNet de tous les concepts de O_{Tar} et de l'ensemble des concepts de O_{Src} non appariés au préalable par les techniques d'appariement précédentes de notre système.

Notre technique se différencie aussi 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. 3) 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.3) et des généralisants intermédiaires extraits de WordNet qui peuvent, ou non, appartenir à l'une des deux ontologies.

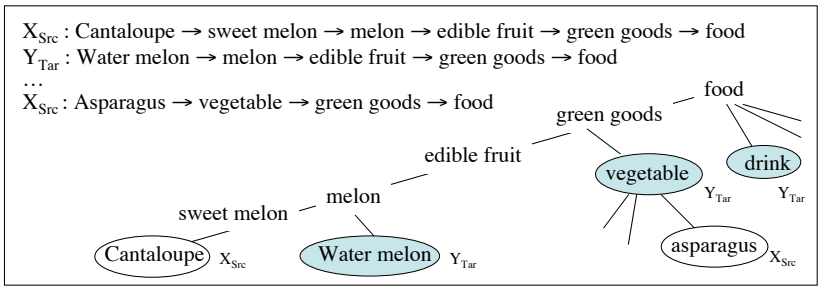

Figure 3. *Un exemple de sous graphe T_{WN} de racine food*

Dans l'expérimentation menée sur les ontologies Russia évoquée précédemment, les racines choisies pour couvrir les thématiques des termes de la cible, (*Location, Living Thing, Structure* et *Body of Water*), ne sont des généralisants pour aucun des termes de O_{Src} évoquant des véhicules. Aucune des dérivations issues de ces termes n'est donc retenue et aucun d'eux n'appartient finalement au sous graphe T_{WN}. Aucun appariement ne peut donc être proposé pour eux. Nous préférons éviter de reconnaître des termes plutôt que de mal les reconnaître. Le rappel des appariements finalement identifiés sera ainsi plus faible mais la précision bien meilleure.

5.2. *Identification des mappings*

Par rapport aux autres travaux, du fait de notre contexte d'application, nous limitons le nombre de relations à identifier. En effet, le processus d'alignement d'ontologies s'effectue dans le cadre de la recherche de documents à partir d'un portail Web. L'objectif est d'augmenter le nombre de documents accessibles par un portail sans en modifier l'interface d'interrogation. L'interface du portail s'appuyant sur une taxonomie cible (O_{Tar}), aligner cette taxonomie avec celles de documents issus de sources externes au portail permet d'augmenter le nombre de documents accessibles. Dans ce contexte, les seuls mappings pertinents sont ceux qui apparient les concepts des taxonomies des sources externes (O_{Src}) avec des concepts de la taxonomie du portail (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 ($X_{Src} \geq Y_{Tar}$).

L'identification des mappings sémantiques pertinents est faite en recherchant, pour chaque ancre d'un concept de O_{Src}, la plus proche ancre d'un concept de O_{Tar} qui apparaisse sur une dérivation menant à la racine $root_A$. Ainsi à partir du sous graphe Fig. 3., on peut dériver le mapping (asparagus *isA* vegetable). Notre technique est comparable dans ses résultats à celle mise en œuvre par Aleksovski 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 (X_{Src} *isA* Y_{Tar}).

Remarquons que cette étude des dérivations ne permet pas d'identifier 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, comme Kalfoglou, une mesure de similarité entre nœuds d'un même graphe, mais seulement sur le sous graphe T_{WN}. Nous avons aussi fait le choix d'utiliser dans *TaxoMap* la mesure proposée par Wu et Palmer, car une étude des propriétés de cette mesure, (Reynaud et Safar, 2007), nous a permis d'identifier et de mettre en œuvre une stratégie de recherche permettant de retrouver très efficacement dans T_{WN}, le concept de O_{Tar} qui sera évalué comme le concept le plus similaire d'un concept donné 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 dont la sémantique est bien définie, de type *isA* ou *Eq*. Il est tout aussi clair qu'il serait dommage de ne pas exploiter l'information trouvée ! 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 X_{Src} de O_{Src} qui reste à apparier, le concept Y_{Tar} de O_{Tar} qui lui est le plus similaire suivant la mesure de Wu et Palmer, qui sera noté Y_{Sim}, et de construire le mapping potentiel associé (X_{Src} *isClose* Y_{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 Y_{Sim} apparaît relié à un même concept X_{Src} 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 mapping est conservé. Comme aucun mapping sémantique ne peut être construit pour le concept cantaloupe, nous conservons, en revanche, le mapping potentiel (cantaloupe *isClose* Watermelon).

5.3. *Expérimentations*

Nous avons tout d'abord réalisé trois expérimentations de cette technique dans le domaine du risque alimentaire. Le domaine des deux ontologies à comparer étant identique et très ciblé, l'identification d'une unique racine (*food*) couvrant tous les concepts des deux ontologies était immédiat. Dans la première expérimentation, la technique a été utilisée directement et seule, sur tous les concepts de O_{Src} en

s'appuyant sur une méthode d'ancrage du type inclusion de labels. Dans la deuxième et la troisième expérimentation, la technique a été utilisée en complément d'autres techniques (donc sur les seuls concepts non encore appariés). Dans ces deux dernières expérimentations, la technique a été appliquée, dans un cas, avec une méthode d'ancrage du type inclusion de labels et dans l'autre cas, sans méthode d'ancrage particulière, en ne s'appuyant que sur la capacité d'ancrage de WordNet, celui-ci permettant par exemple d'ancrer directement le terme poultries à sa forme lexicalisée poultry.

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 identifiées 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 dû être appliquée que sur les seuls concepts non encore appariés, avec le plus souvent des labels courts, et les résultats sont plus pertinents. Sans méthode d'ancrage particulière, sur 29 concepts testés non encore appariés, *TaxoMap* identifie correctement 6 mappings de spécialisation (lamb *isA* meat, frankfurter *isA* sausage, broccoli *isA* vegetable,…), et 3 mappings potentiels (cantaloupe *isClose* Watermelon, broccoli *isClose* cauliflower,..). Avec l'ancrage par inclusion de labels, 9 mappings supplémentaires sont identifiés (25% cider *isA* drink, almound paste *isA* ingredient, … pumkin pie *isA* pastry) mais 2 sont incorrects (apple cider *isClose* vegetable, pumkin pie *isClose* meat pie).

Une série d'autres expérimentations a été réalisée sur des taxonomies servant de test dans la communauté appariement. Toutes ces expérimentations ont montré que si le domaine d'application des ontologies était très large, l'utilisation d'une unique racine n'était pas adaptée. En effet, si le domaine est très large, le concept de WordNet généralisant les concepts à apparier est très général, (*entity)* et le sous graphe construit, T_{WN} est très gros. T_{WN} est ainsi composé de tous les nœuds de la hiérarchie de WordNet sans aucune restriction. Il mêle des sens de termes différents et conduit la technique à générer des mappings qui ne sont absolument pas pertinents.

Nous présentons dans Tab.1 le nombre de mappings obtenus sur les taxonomies Russia. Avec une racine unique *entity*, (cf. la première colonne de Tab.1), la technique utilisée sans méthode d'ancrage particulière permet d'identifier 61 mappings de type *isA* et 15 de type *isClose* parmi les 162 termes de Russia-B non appariés par les heuristiques préalables de *TaxoMap* (sur 370 termes au départ). En l'absence d'une liste complète des mappings de référence, nous avons évalué les résultats manuellement. Seuls 29 des 61 mappings *isA* et 8 des 15 mappings *isClose* nous ont paru corrects. En particulier, tous les mappings relatifs aux véhicules de Russia-B sont faux, comme nous l'avions vu Fig.2.

Une amélioration sensible de la précision des mappings retournés a été obtenue en identifiant plusieurs racines pour couvrir les sous domaines traités par l'ontologie cible et en construisant en même temps plusieurs sous arbres distincts, un par sous domaine. Sur Russia, avec les racines *Location, Living Thing, Structure* et *Body of Water*, la technique employée sans méthode d'ancrage, (cf. la 2ème colonne de Tab1), permet d'identifier 35 mappings de type *isA* et 11 de type *isClose.* 29 des 35 mappings *isA*, en particulier tous les mappings géographiques concernant des noms de villes, de pays, de régions et de fleuves et 9 des11 mappings *isClose* nous ont paru corrects.

	Avec une racine unique (Entity)	Avec plusieurs racines et sans méthode d'ancrage	Avec plusieurs racines et une phase d'ancrage
# *isA* mappings trouvés (corrects)	61 (29)	35 (29)	42 (32)
# *isClose* mappings trouvés (corrects)	15 (8)	11 (9)	12 (10)
Total des mappings (corrects)	76 (37)	46 (38)	54 (42)
Recall (Précision)	0,23 (0,49)	0,23 (0,83)	0,26 (0,78)

Table 1. *Nombre de mappings trouvés parmi les 162 termes de Russia-B*

Bien que le même nombre (29) de mappings *isA* corrects apparaisse dans ces deux premières expérimentations, les mappings corrects sont différents. Par exemple, dans la deuxième expérimentation, le mapping (alcohol *isA* drink) n'est pas identifié puisque le concept drink de O_{Tar} n'est pas couvert par les racines choisies. En revanche, le mapping (pine *isA* plant) est correctement identifié alors que, dans la première expérimentation, sans limitation du sens, un mapping incorrect (pine *isA* material) est trouvé.

Une troisième expérimentation menée avec les mêmes racines mais en utilisant une méthode d'ancrage basée sur l'inclusion de labels montre une dégradation de la précision des mappings identifiés (cf. 3ème colonne de Tab.1., 7 mappings *isA* supplémentaires sont identifiés mais 5 de ces 7 nouveaux mappings sont erronés). Nous en concluons que l'inclusion de labels n'est pas une méthode d'ancrage adaptée à WordNet, en particulier quand les labels des concepts sont des expressions composées de plusieurs mots.

Un choix plus fin des racines permettrait très certainement d'améliorer le rappel. Dans notre contexte applicatif, la phase d'identification de ces racines dans WordNet peut être faite uniquement en référence aux concepts apparaissant dans O_{Tar}. Cette tâche ne doit donc être effectuée qu'une seule fois et les racines identifiées pourront être réutilisées quelles que soient les taxonomies sources devant

être alignées avec O_{Tar}. De ce fait, l'identification des racines mériterait d'être faite avec soin pour identifier précisément tous les sous domaines couverts par O_{Tar}. Les premiers résultats présentés dans cet article nous paraissent déjà très encourageants même s'ils ont été obtenus sans que toutes les racines pertinentes n'aient été identifiées.

6. Conclusion

Dans cet article, nous avons discuté de l'intérêt d'utiliser des connaissances complémentaires pour la découverte automatique de mappings. Nous tirons de cette étude les conclusions suivantes.

Le recours à des connaissances externes est délicat lorsque le domaine d'application est très large ou inconnu au départ. Ainsi des techniques génériques, applicables quel que soit ce domaine et ayant recours de façon dynamique à de telles connaissances externes, sont face à un réel problème d'identification du contexte d'étude. En revanche, l'usage de connaissances de support est très intéressant lorsqu'on connaît précisément le contexte au sein duquel les éléments manipulés doivent être interprétés et qu'il est possible de le spécifier. Il permet de découvrir des mappings sémantiques et de dépasser les limites des approches syntaxiques.

WordNet est souvent utilisé comme ressource complémentaire car il est une bonne source d'information sur les synonymies et fournit une hiérarchie de concepts basée sur les relations de généralisation/spécialisation. En revanche les expériences ont montré qu'il est difficile d'en extraire des relations pertinentes si le sens précis des termes cherchés n'est pas préalablement défini. Le résultat de nos expérimentations avec WordNet montre que notre approche basée sur l'identification de racines multiples est une solution prometteuse quand le domaine de l'application est très large. Quel que soit ce domaine, il est possible avec notre technique, d'extraire de WordNet le sous arbre regroupant les seuls concepts pertinents pour l'application et d'éviter les problèmes de contresens. De plus, la phase de recherche de dérivation peut être effectuée efficacement dans ce sous arbre, puisque toutes les ancres sont connues. Nous avons montré comment identifier au sein du sous arbre les mappings sémantiques quand ils existent et comment identifier sinon des mappings potentiels traduisant des relations de proximité.

Nous avons aussi montré dans nos expérimentations que la technique d'utilisation de connaissances de support ne peut pas être employée comme seule méthode d'alignement. Dans *TaxoMap,* différentes techniques sont appliquées en séquence, et celle qui s'appuie sur WordNet intervient après les techniques terminologiques et structurelles. Elle permet d'identifier, avec une bonne précision, des mappings supplémentaires qui ne peuvent pas être identifiés par les autres techniques, et permet ainsi d'augmenter l'efficacité globale du système. Mais c'est une technique complémentaire aux autres qu'elle ne remplace pas à elle seule.

7. Bibliographie

Aleksovski Z., Klein M., Ten Kate W., Van Harmelen F. «Matching Unstructured Vocabularies using a Background Ontology », *Proceedings of the 15th International Conference on Knowledge Engineering and Knowledge Management (EKAW'06),* October 2006, Springer-Verlag, pp. 182-197.

Aleksovski Z., Klein M., Ten Kate W., Van Harmelen F. «Exploiting the Structure of Background Knowledge used in Ontology Matching ». *ISWC'06 Workshop on Ontology Matching (OM-2006),* November 2006, Athens, Georgia, USA, pp. 13-24.

Bach T.L., Dieng-Kuntz R., Gandon F., « On Ontology Matching Problems - for Building a Corporate Semantic Web in a Multi-Communities Organization ». ICEIS (4) 2004: pp. 236-243.

Kalfoglou Y.,Hu B.,« Crosi Mapping System (CMS) Result of the 2005 Ontology Alignment Contest». *K-Cap'05 Integrating Ontologies workshop*, 2005, Banff, Canada, pp. 77-85.

Lin D. « An Information-Theoretic Definition of Similarity ». ICML, Madison, 1998, pp. 296-304.

Pedersen, T., Patwardhan, S., Michelizzi J. « WordNet::Similarity - Measuring the Relatedness of Concepts », AAAI-04, July 2004, San Jose, CA.

Reynaud C., Safar B. « When usual structural alignment techniques don't apply ». *ISWC '06 Workshop on Ontology Matching (OM-2006),* Poster, Athens, Georgia, USA.

Reynaud C., Safar B. « Techniques structurelles d'alignement pour portails Web », Revue RNTI, à paraître, 2007.

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.

Shvaiko P., Euzenat J. « A Survey of Schema-based Matching Approaches », *Journal on Data Semantics,* 2005, pp. 146-171.

Wu Z., Palmer M. *Verb semantics and lexical selection.* In 32nd Annual Meeting of The Association for Computational Linguistics, 1994, Las Cruces, pp. 133-138.