
Alignement d'ontologies basé sur des ressources complémentaires : illustration sur le système *TaxoMap*

Brigitte Safar^{*,} – Chantal Reynaud^{*,**}**

^{*} LRI ; Univ. Paris-Sud 11, CNRS UMR 8623
Bât. 490, 91405 Orsay, France

^{**} INRIA Saclay - Île-de-France - Projet Gemo
Parc Orsay Université
4 rue Jacques Monod; Bât. G
91893 Orsay, France

{safar, chantal.reynaud}@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 sous la forme d'une 3^{ème} 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 taxonomies. Cet article comporte deux parties. La première présente une étude de différents 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. Une seconde partie est consacrée au système d'alignement *TaxoMap*. Nous présentons le système et son contexte d'utilisation. Nous décrivons ensuite l'utilisation de *WordNet* comme connaissance de support ainsi que les résultats d'expérimentation obtenus.

ABSTRACT. A lot of alignment systems providing mappings between the concepts of two ontologies rely on the use of 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. This paper has two parts. First, we present a state-of-the-art of research work using background knowledge. A common general scheme is first introduced followed by an analysis of works that differ by the kind of background knowledge they use. The second part is dedicated to *TaxoMap*. We present the use context and the general architecture of the system. Then, we describe the way *WordNet* is exploited in *TaxoMap* as support knowledge together with experimental results.

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

KEYWORDS: Ontology Matching, 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'explicitier 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).

Nous nous intéressons, dans cet article, aux travaux ayant pour objectif d'aligner deux ontologies et utilisant des connaissances complémentaires, dites de "background" ou de support, représentées le plus souvent sous la forme d'une troisième ontologie. Les techniques basées sur l'exploitation de connaissances complémentaires complètent les techniques plus 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. Ces techniques fournissent des résultats intéressants à condition, bien sûr, que de telles connaissances complémentaires soient disponibles.

L'utilisation de connaissances complémentaires pour aligner deux ontologies passe par la mise en œuvre d'un processus d'alignement dont le schéma général est aujourd'hui bien identifié. Par contre, la façon dont le processus est mis en œuvre et les problèmes posés dépendent en grande partie de la nature des connaissances complémentaires utilisées, en particulier s'il s'agit de connaissances générales ou de connaissances spécifiques à des domaines d'application particuliers. Si le recours à des connaissances généralistes est possible quel que soit le domaine des ontologies à apparier, comment être sûr que les contextes d'interprétation des concepts manipulés sont identiques ? Si le recours à des connaissances plus ciblées évite de mêler plusieurs contextes d'interprétation, comment obtenir de telles connaissances ? Si plusieurs ontologies sont utilisées afin de couvrir l'ensemble du domaine des ontologies à aligner, comment gérer cette multiplicité ? Toutes ces questions sont importantes et seront abordées dans cet article. Nous illustrerons le recours à des connaissances généralistes par l'exploitation de WordNet. Nous proposerons une solution au problème de contextes d'interprétation multiples dans WordNet qui a été mise en œuvre dans le système *TaxoMap* et qui a donc pu faire l'objet d'expérimentations.

Dans cet article, nous adoptons le plan suivant. En section 2, nous présentons les travaux en alignement utilisant des connaissances complémentaires. Nous décrivons tout d'abord le schéma général adopté dans ces travaux pour

l'alignement. Nous présentons ensuite les travaux basés sur l'utilisation d'une ressource généraliste, WordNet, puis les travaux s'appuyant sur des ontologies spécifiques à des domaines d'application plus ciblés. En section 3, nous décrivons *TaxoMap* en précisant son contexte d'utilisation et l'architecture du système. L'utilisation de WordNet comme connaissance de support dans *TaxoMap* fait l'objet de la section 4, qui détaille la solution proposée pour éviter les rapprochements erronés dus aux multiples sens d'un même terme. La section 5 est consacrée à l'expérimentation et à l'analyse des résultats montrant le gain de précision des appariements obtenus par l'application de la solution proposée. La section 6 conclut le papier.

2. Travaux sur l'alignement utilisant des ressources complémentaires

2.1. Description générale du processus d'alignement

Le processus d'alignement d'ontologies a pour objectif de mettre en correspondance deux ontologies, une ontologie dite source (O_{Src}) et une ontologie dite cible (O_{Tar}), portant a priori sur le même domaine applicatif. Pour simplifier cette présentation, nous considérerons que chaque ontologie O ne comprend qu'un ensemble de concepts C et un ensemble de liens de subsomption reliant ces concepts. Une mise en correspondance, encore appelée mapping, consiste à mettre en relation un concept de l'ontologie source, $X_{Src} \in C_{Src}$, avec un concept de l'ontologie cible, $Y_{Tar} \in C_{Tar}$, pour obtenir une relation de la forme (X_{Src} relation Y_{Tar}) où *relation* appartient le plus souvent à l'ensemble $\{\leq, \equiv\}$ et où $X \leq Y$ peut se lire, suivant les cas, « X isA Y », « X part-of Y » ou « X narrower-than Y » généralisant les relations *isA* et *part-of*.

Baser ce processus sur l'exploitation de ressources complémentaires consiste à exploiter des mises en correspondance déjà présentes ou identifiables dans des ressources différentes de celles qu'il s'agit d'aligner. La description du processus présentée ci-dessous et le schéma qui l'illustre (Fig.1) s'appuient largement sur les travaux de (Sabou *et al.*, 2006) et (Aleksovski *et al.*, 2006a).

L'approche générale suivie par la plupart des travaux utilisant des connaissances complémentaires se décompose en 2 phases : l'**ancrage** et la **dérivation**.

L'**ancrage** consiste à appairer chacun des 2 concepts X_{Src} et Y_{Tar} , pris indépendamment l'un de l'autre, avec un ou des concepts de la 3^{me} 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 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 appairer,

- 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.

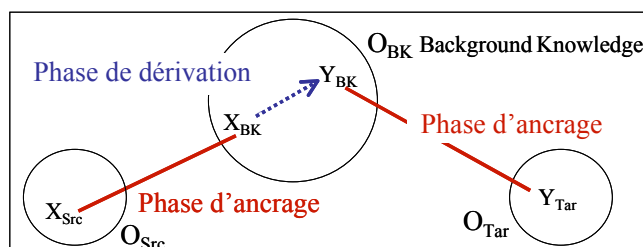


Figure 1. Schéma général de l'identification d'un mapping (X_{Src} relation Y_{Tar})

La phase d'ancrage vers les éléments de O_{BK} s'appuie le plus souvent sur des heuristiques terminologiques simples qui portent sur les labels des 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 d'un label d'un concept A se trouvent dans l'un des labels d'un concept B, alors B sera considéré comme plus spécialisé que A ($B \leq A$).

A partir de ce schéma général, les travaux exploitant des ressources complémentaires se différencient en fonction des caractéristiques des ressources employées comme support, suivant que celles-ci sont des ontologies à la thématique ciblée ou au contraire une ressource généraliste comme WordNet. Nous présenterons tout d'abord les travaux s'appuyant sur la ressource généraliste WordNet, puis ceux ayant recours à des ontologies ciblées.

2.2. Travaux basés sur l'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 ».

Une autre utilisation consiste à n'exploiter que les relations de généralisation/spécialisation entre les différents synsets (assimilables à des concepts) et à considérer WordNet comme une hiérarchie. Dans (Giunchiglia *et al.* 2006), des relations d'équivalence entre deux nœuds sont inférées lorsque leur distance en nombre d'arcs, en suivant les liens de généralisation/spécialisation, est inférieure à un certain seuil.

D'autres travaux s'appuient sur cette hiérarchie pour calculer une mesure de similarité entre deux nœuds. C'est le cas de l'outil *WordNet::Similarity* (Pedersen *et al.* 2004) 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 également utilisée par (Kalfoglou *et al.*, 2005) dans un des modules d'appariement, *WNNNameMatcher*, 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 & 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*).

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

Cette même mesure a été utilisée dans une des techniques d'alignement mises en œuvre dans *TaxoMap* (Reynaud & Safar, 2007). Nous montrons qu'elle donne des résultats pertinents quand les domaines d'application des ontologies à comparer sont proches et très focalisés. En revanche, lorsque les domaines d'application sont larges et ne se recoupent pas, les résultats sont beaucoup moins satisfaisants. Le problème est dû au phénomène de polysémie, i.e. aux différents synsets auxquels un même terme peut appartenir. Cette polysémie entraîne fréquemment des contresens et donc des rapprochements erronés. Pour illustrer ce phénomène, nous nous appuyons sur des expérimentations menées sur une paire de taxonomies¹ 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

¹ <http://www.atl.external.lmco.com/projects/ontology/ontologies/russia/>

² <http://oaci.ontologymatching.org/2007/>

touristiques, sa géographie, ses monuments et en plus dans Russia-B, ses moyens de transport.

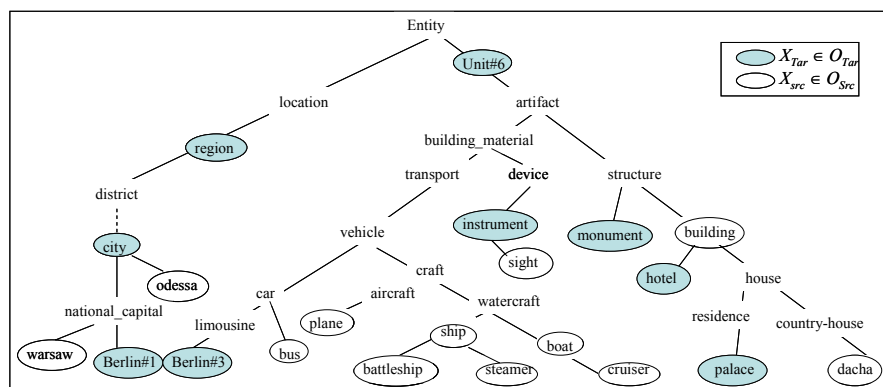


Figure 2. Sous graphe de WordNet mobilisé dans la recherche des véhicules

Dans ces expérimentations, des techniques d'alignement terminologiques ont été appliquées avant la technique basée sur WordNet, ce qui a permis de trouver les alignements entre termes syntaxiquement proches. Fig.2 correspond à l'application de la technique basée sur WordNet. Elle représente une partie de la hiérarchie WordNet après ancrage des concepts des deux ontologies non encore appariés. Les ancres de O_{Tar} sont représentées par des ovales gris et celles de O_{Src} par des ovales blancs. Dans les expérimentations réalisées, la non gestion de la polysémie conduit, par exemple, à rapprocher le terme « bus », correspondant à l'ancre d'un concept de O_{Src} , du terme « Berlin » correspondant à une ancre d'un concept de O_{Tar} qui est le nœud le plus proche de cette hiérarchie qui correspond à un véhicule. Or, cet appariement est erroné. En effet, le terme « Berlin » appartient dans WordNet à 3 synsets : *Berlin#1* qui correspond à la capitale de l'Allemagne, *Berlin#2* qui correspond à un musicien et *Berlin#3* qui correspond à une sorte de voiture (la berline). Un des concepts de O_{Tar} est bien ancré avec « Berlin » mais dans le sens du synset *Berlin#1*, compte tenu du domaine couvert par O_{Tar} , bien que cela ne soit pas explicite.

Une deuxième occurrence de ce même phénomène apparaît Fig.2 pour le terme « unit » qui possède 6 sens différents. Un des concept de O_{Tar} a été ancré avec « unit » dans le sens *unit#1* qui correspond à une unité de mesure. Or « unit » apparaît aussi comme hyperonyme de différentes ancres de O_{Src} , il sera donc retenu dans les alignements proposés pour les concepts associés à ces ancres, alors que le sens de cet hyperonyme est ici *unit#6*, la notion de tout, et non pas l'unité de mesure. On aboutit, là encore, à des alignements erronés du fait de la non gestion de la polysémie.

Le problème de la polysémie est traité par le système *S-Match* (Giunchiglia *et al.* 2004). Ce système calcule des relations logiques (telles que l'équivalence ou la

subsumption) entre les concepts et traite le problème d'alignement entre deux nœuds de deux taxonomies comme un problème de satisfiabilité de formules de la logique propositionnelle. L'approche proposée se base sur deux notions clés, la notion de **concept de label** et la notion de **concept de nœud**. A un concept de label est associé l'ensemble des documents qui portent sur le label du nœud pris de façon isolé alors qu'à un concept de nœud est associé l'ensemble des documents que l'on voudrait classer sous le nœud compte tenu de sa position dans la taxonomie. Le calcul d'un concept de nœud correspond à la conjonction du concept de label de ce nœud et de tous les concepts de label des nœuds le reliant à la racine de la taxonomie.

Le système S-Match utilise WordNet pour définir les concepts de labels. En présence de labels exprimés par des expressions composées de plusieurs mots, telles que « Food and Cheese » par exemple, S-Match extrait de WordNet les sens de chacun des mots du label et traduit les prépositions, signes de ponctuation ou de conjonction en connecteurs logiques. Ainsi, le « and » étant traduit par le connecteur logique « \cup », les sens associés au concept de label $C_{\text{FoodAndCheese}}$ sont représentés par l'union des 3 sens de Food et des 4 sens de Cheese : $\langle \text{Food}, \{\text{senses\#3}\} \rangle \cup \langle \text{Cheese}, \{\text{senses\#4}\} \rangle$.

Le système raisonne ensuite sur les sens calculés pour chaque concept de label pour identifier les relations sémantiques valides entre les concepts de label des deux taxonomies. Les relations sémantiques identifiées, relations d'équivalence ($=$), de généralisation (\supseteq), de spécialisation (\subseteq) et disjonction (\perp), sont générées à partir de règles portant sur les relations lexicales de WordNet (synonyme, hyperonyme/hyponyme, méronyme/holonyme, antonyme). Ainsi une relation de spécialisation peut être identifiée entre les concepts C_{Wine} et $C_{\text{FoodAndCheese}}$ puisqu'un des sens de Wine est un hyponyme d'un des sens de Food. L'ensemble des relations valides calculées entre les différents concepts de label est ensuite traduit sous forme d'expressions logiques, et utilisé par un solveur SAT pour calculer les relations valides entre les concepts de nœuds.

Une phase de désambiguïsation plus simple est effectuée dans (Mougin *et al.*, 2006). Dans ces travaux, WordNet est utilisé comme ressource complémentaire pour aider à valider des mappings entre des ressources Web et l'ontologie médicale UMLS. L'ancrage des différents termes est effectué dans WordNet. Lorsque plusieurs mappings sont trouvés, on privilégie ceux qui s'appuient sur des synsets dont la définition ou les hypernymes contiennent des termes appartenant à une liste prédéfinie de mots-clés du domaine bio-médical.

Nous présentons dans le paragraphe 4, comment notre système *TaxoMap* traite la polysémie en limitant les sens des termes de WordNet pris en compte dans la recherche de mappings.

2.3. Les travaux s'appuyant sur des ontologies ciblées

Le problème de la polysémie évoqué dans le paragraphe précédent est évité quand les ressources qui servent de support sont des ontologies ciblées. Dans un contexte clairement défini, les homonymes sont rares et la phase de dérivation peut être plus simple, sans être perturbée par les ambiguïtés.

Les travaux présentés ici recherchent des relations entre les ancres des concepts des ontologies à aligner en s'appuyant sur des règles de dérivation sémantiques. Dans les relations cherchées, de la forme $(X_{Src} \text{ relation } Y_{Tar})$, l'ensemble R des relations est généralement l'ensemble $\{\leq, \geq, =\}$ où $X \leq Y$ peut se lire, suivant les cas, « X isA Y », « X part-of Y » ou « X narrower-than Y ». Les mappings cherchés sont alors 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 $(X_{Src} \geq Y_{Tar})$.

La relation d'équivalence « \equiv » est aussi prise en compte 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 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.

Parmi les travaux exploitant ces règles de dérivation, certains supposent que la recherche peut s'effectuer sur une ontologie de support unique, préalablement identifiée et qui couvre a priori tous les concepts des ontologies à apparier. Ainsi, dans (Aleksovski *et al.*, 2006a et 2006b) ou dans (Zhang *et al.*, 2006), deux travaux dans le domaine médical, l'ontologie de support O_{BK} est plus complète et plus détaillée que les deux ontologies à rapprocher, et contient une description en compréhension du domaine des 2 autres.

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. Dans (Aleksovski *et al.*, 2006b) ou dans (Zhang *et al.*, 2006), 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 $(X_{BK} \text{ isA } Y_{BK})$ alors $(X_{BK} \leq Y_{BK})$ et Si $(X_{BK} \text{ part-of } Y_{BK})$ alors $(X_{BK} \leq Y_{BK})$. 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 $(X^1_{BK} \leq X^n_{BK})$, 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*.

A l'inverse, d'autres travaux considèrent que la recherche de dérivation ne peut s'effectuer qu'au sein de multiples ontologies de support, sélectionnées dynamiquement. Ainsi, dans (Sabou *et al.*, 2006), les auteurs supposent qu'il n'existe pas a priori, pour un domaine donné, 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. Lorsque 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.

Même si elle peut être parallélisée, cette dernière stratégie est bien sûr encore plus coûteuse que la précédente. De plus, le fait de s'appuyer sur des ontologies identifiées à la volée fait resurgir le problème de la polysémie. Par exemple, le mapping (Game *isA* Sport) peut être dérivé à l'aide d'une ontologie présente sur le Web, où le concept Game est défini comme un spécialisant du concept « Recreation or Exercice » et un généralisant du concept « Sport », alors que dans le contexte des ontologies à apparier, ce mapping est incorrect car Game doit être compris comme une spécialisation du concept « Wild animal ». Pour détecter et rejeter les mappings incorrects dus à la polysémie, les auteurs proposent dans (Gracia *et al.*, 2007) de compléter leur approche en introduisant une phase d'étude des synsets associés dans WordNet aux termes intervenant dans les mappings obtenus. Deux termes X et Y sont considérés comme **sémantiquement similaires** s'il existe un synset S' de X qui soit aussi un synset de Y ou qui soit relié à un synset de Y par une suite de relations de généralisation/spécialisation dans WordNet. Pour un mapping de la forme (X_{Src} relation Y_{Tar}), ce mapping doit être rejeté si S' n'existe pas ou s'il existe mais qu'il n'est pas sémantiquement similaire à l'un des ancêtres de X dans son ontologie d'origine. Dans l'exemple du mapping (Game *isA* Sport), on peut trouver le synset S' « diversion activity », comme généralisant de Game et Sport, mais ce synset n'est pas sémantiquement relié aux ancêtres de Game dans son ontologie d'origine. Le mapping est donc rejeté.

3. TaxoMap : Contexte d'utilisation et architecture générale

Nous décrivons ici le cadre dans lequel le système d'alignement, *TaxoMap*, a été conçu. Nous présentons tout d'abord le contexte d'utilisation et les différentes répercussions de ce contexte sur le mécanisme d'alignement mis en œuvre, puis l'architecture générale de l'outil.

3.1. Contexte d'utilisation

Nous présentons ici l'application initiale, puis l'utilisation prévue des alignements à découvrir et les contraintes posées par les utilisateurs de l'application. Le contexte dans lequel a été conçu le système *TaxoMap* est celui de l'interrogation d'un portail d'information dans le domaine du risque alimentaire³. L'interrogation du portail se fait par l'intermédiaire d'une interface de requête qui s'appuie sur un schéma global du domaine. Ce schéma est une taxonomie $O_{Tar} = (C_{Tar}, H_{Tar})$, constituée d'un ensemble de concepts C_{Tar} reliés par des liens de subsumption au sein d'une hiérarchie H_{Tar} . Etant donnée une requête composée d'une conjonction de concepts du schéma global, l'interface retourne l'ensemble des documents annotés par la conjonction de ces concepts ou de leurs spécialisants, si ceux-ci en possèdent dans le schéma. Sur l'exemple présenté Fig. 3, la requête $Query(X)$ portant sur le concept X de O_{Tar} retournera les documents du portail annotés par les concepts X , X_1 , X_2 et X_3 où X_1 , X_2 et X_3 sont les spécialisants de X dans O_{Tar} .

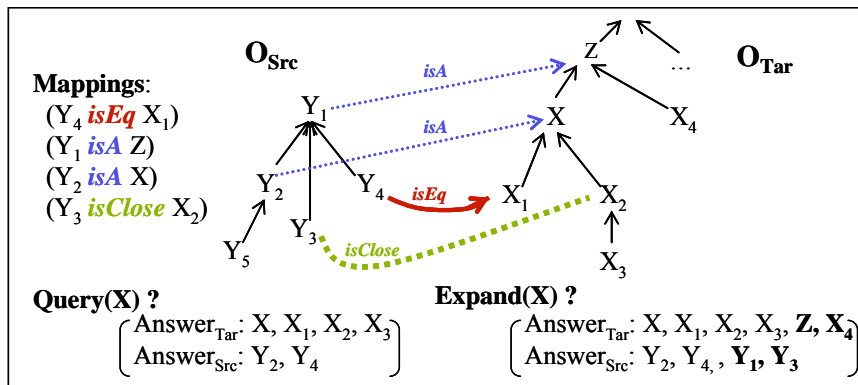


Figure 3. Exemple d'utilisation de mappings dans une requête

Pour augmenter les réponses retournées par le portail, il est intéressant de pouvoir accéder à des documents du domaine n'appartenant pas au portail et qui sont annotés par d'autres ressources sémantiques. Le problème est alors d'être capable d'aligner les concepts utilisés pour l'annotation dans ces ressources

³ Travail effectué dans le cadre du projet ANR RNTL e.dot (Entrepôt de Données Ouvert sur la Toile), 2002-2005.

externes avec ceux du schéma global du portail. *TaxoMap* a été conçu dans ce but et établit des appariements sous la forme de relations d'équivalence (*isEq*), de subsumption (*isA*) ou de proximité (*isClose*). Le processus d'alignement mis en œuvre est un processus orienté pour découvrir des alignements entre les concepts représentés dans la ressource externe (l'ontologie source, O_{Src}), vers l'ontologie du portail (l'ontologie cible, O_{Tar}).

Les utilisateurs du portail ont souhaité que le processus d'interrogation ne soit pas altéré par l'alignement et que le vocabulaire utilisé pour l'interrogation reste celui du schéma initial. De plus, la structuration de l'ontologie source pouvant être très faible ou différente de celle du portail, les liens de subsumption existant dans cette ontologie n'ont pas été pris en compte dans le processus de recherche, à la demande des utilisateurs. Les seules relations utilisées sont donc les relations de subsumption existant dans le schéma du portail, O_{Tar} , et les mappings identifiés. Sur l'exemple Fig. 3, la requête $Query(X)$ retournera les documents de la source annotés par les concepts Y_2 et Y_4 mais pas par Y_5 car ce concept n'est pas relié directement par un mapping à un concept de O_{Tar} .

De plus, les utilisateurs ont souhaité que les appariements construits soient sélectifs puisqu'ils influencent directement la qualité des résultats retournés. Les appariements effectués par une relation d'équivalence devront être sûrs. Si un concept de la source externe est jugé très proche mais non équivalent à un concept de O_{Tar} , les utilisateurs le considèrent au mieux comme un fils ou un frère de ce concept. Par exemple, dans Fig. 3, nous supposons que le concept Y_1 de la source a été jugé proche de X mais pas assez pour lui être équivalent. Il a ainsi été considéré comme son frère, donc rattaché par un lien *isA* à Z , le père de X . Il n'est pas pris en compte par la requête $Query(X)$.

Pour respecter la rigueur de ces contraintes mais permettre néanmoins l'augmentation des réponses retournées par le portail, nous avons proposé aux utilisateurs de prévoir un mécanisme explicite d'élargissement de requête, distinct de l'interrogation habituelle. Comme dans les travaux classiques dans le domaine des *Réponses Coopératives* (Minker, 1998), (Bidault *et al.*, 2000), ce mécanisme d'élargissement s'appuie sur les liens de subsumption de l'ontologie, et élargit par exemple la requête $Query(X)$ en la remplaçant par $Query(Z)$ où Z est l'hyperonyme de X . L'élargissement permet ainsi d'atteindre les documents annotés par Z ou par ses spécialisants, les concepts frères de X , par exemple dans Fig. 3, les concepts X_4 et Y_1 .

Nous avons de plus proposé d'utiliser dans les mappings, en complément des relations *isA* et *isEq*, une nouvelle relation, dénotée *isClose* qui ne sera utilisée que dans les phases d'élargissement de la requête. Cette nouvelle relation nous permet de traduire une proximité entre deux concepts, identifiée par exemple par une mesure de similarité, mais sans qu'on puisse clairement expliciter sa sémantique.

Dans l'exemple présenté Fig. 3, l'élargissement de $Query(X)$ en $Expand(X)$ permet d'atteindre les documents supplémentaires annotés par les concepts

apparaissant en gras dans la figure, i.e. les documents supplémentaires étiquetés par Z , X_4 et Y_1 atteints par $Query(Z)$ et ceux étiquetés par Y_3 , concept associé par la relation de proximité ($Y_3 isClose X_2$) au spécialisant de X, X_2 .

Après avoir présenté le contexte d'utilisation des mappings découverts par notre système d'alignement, nous présentons brièvement son architecture générale.

3. 2. Architecture générale

Comme nous l'avons vu dans la section précédente, *TaxoMap* a été conçu pour découvrir des alignements entre des taxonomies où les concepts sont seulement définis par leur label et les relations de subsomption qu'ils entretiennent avec les autres concepts. Le processus d'alignement est un processus orienté qui cherche à relier chaque concept de la taxonomie source à un unique concept de la taxonomie cible. Nous proposons une approche générique et semi-automatique, un processus totalement automatique n'étant pas concevable du fait de la grande hétérogénéité sémantique entre les sources. Cette approche s'effectue en deux temps. Dans un premier temps, des mappings **probables** sont automatiquement découverts. Dans un deuxième temps, des mappings **potentiels** sont proposés pour aider l'expert du domaine à mettre en correspondance les éléments pour lesquels un mapping probable n'a pu être trouvé automatiquement. La découverte de mappings repose sur des techniques variées : terminologiques et structurelles. Ces différentes techniques sont composées de façon à rendre le processus de génération des mappings le plus efficace possible.

Les techniques terminologiques sont appliquées en priorité. Elles permettent d'exploiter toute la richesse des labels des concepts, en particulier dans les domaines où les homonymes sont rares et où les labels des concepts sont précis et souvent composés de plusieurs mots. Ces techniques permettent d'identifier des mappings reliant les concepts par des relations d'équivalence ou de subsomption. La précision⁴ élevée des mappings obtenus par ces techniques permet de qualifier les mappings de probables.

Les techniques structurelles mises en œuvre dans un deuxième temps, permettent d'identifier des mappings potentiels, i.e. moins sûrs que les précédents et qui devront être validés par un expert, mais indispensables pour compléter les mappings probables. Comme leur nom l'indique, elles s'appuient sur la structure de différentes représentations. Nous proposons trois techniques. La première s'appuie sur la structure de la taxonomie cible qui est supposée être la plus structurée. La deuxième technique s'appuie sur la structure d'une ressource externe, WordNet. La

⁴ Si on dispose d'un alignement de références composé de M mappings, la qualité d'un alignement obtenu composé de N mappings, est évaluée par 2 mesures : la précision et le rappel. Soit C le nombre de mappings corrects parmi ceux obtenus, la précision est le rapport C/N et le rappel est le rapport C/M .

troisième technique exploite la structure de la taxonomie source, qui peut être très pauvre, en la combinant à celle de la taxonomie cible.

C'est la seconde technique structurelle, qui s'appuie sur la structure de la ressource externe WordNet, que nous présentons dans la section suivante.

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

Les techniques terminologiques ne permettent pas de rapprocher deux concepts de labels très différents même s'ils sont sémantiquement très proches comme, par exemple, Cantaloupe et Watermelon. En revanche, l'utilisation d'une ressource linguistique support, comme WordNet, permet de reconnaître ces deux concepts comme deux sortes de melon et met en évidence leur proximité. WordNet est une ressource généraliste, qui ne reconnaît pas les concepts très spécialisés ou avec des labels comprenant plus de deux mots. Quand les ontologies à apparier sont très spécifiques et comportent des descriptions très fines du domaine d'application, les concepts très spécialisés avec des labels très détaillés ne sont pas reconnus. WordNet ne peut donc pas être considéré comme une ressource plus complète et plus détaillée que les deux ontologies de départ et la technique utilisant cette ressource comme connaissance de support ne peut pas être utilisée seule. Dans *TaxoMap* la technique est complémentaire à d'autres techniques appliquées au préalable. Au moment de son utilisation, un certain nombre d'appariements ont déjà été effectués par les autres techniques et le recours à WordNet n'intervient que sur les concepts de O_{src} non encore appariés.

Pour éviter les contresens et les rapprochements erronés dûs aux multiples sens possibles d'un même terme, notre technique se décompose en deux phases. Nous commençons par extraire de WordNet des sous-arbres composés des seuls synsets correspondant aux sens supposés pertinents pour le domaine de l'ontologie cible. Nous identifions ensuite, au sein de ces sous-arbres, les mappings sémantiques et les mappings traduisant une simple proximité entre concepts.

4.1. Extraction des sous-arbres de WordNet pertinents pour le domaine

L'extraction s'effectue en deux étapes : les racines des sous-arbres pertinents, i.e. les synsets de WordNet qui couvrent le domaine, sont tout d'abord identifiées, puis les sous-arbres ayant ces concepts pour racines sont extraits.

4.1.1. Identification des racines

Cette étape d'identification est effectuée manuellement par un expert. Si les domaines d'application des ontologies à apparier sont proches et ciblés, l'expert identifie le concept de WordNet, noté *rootA*, qui est le concept le plus spécialisé de WordNet généralisant a priori tous les concepts du domaine des ontologies à

apparié (food par exemple pour l'application sur le risque alimentaire). Si les domaines d'application sont plus larges et distincts, l'expert identifie plusieurs racines qui couvrent les thèmes des concepts de l'ontologie cible (par exemple pour Russia, Location, Living Thing, Structure et Body of Water). Pour ce faire, l'expert est assisté par une interface (présentée Fig. 4) qui facilite la procédure de sélection.

Cette interface permet à l'expert de saisir, dans la fenêtre supérieure, chacun des termes associés à des concepts-racine. Une fois validés, ces termes apparaîtront dans la fenêtre de gauche. La sélection d'un des termes de cette fenêtre permet de visualiser les définitions associées à ses différents sens (dans Fig.4, la fenêtre en haut à droite affiche quatre sens du terme structure sélectionné dans la fenêtre de gauche). La sélection d'un de ces sens permet de le retenir comme racine. Le sens retenu est alors transféré dans la fenêtre en bas à droite.

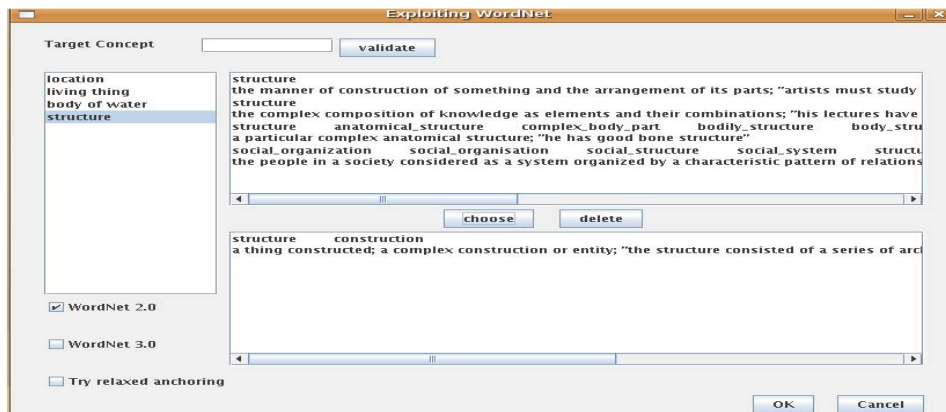


Figure 4 Interface permettant de choisir dans WordNet les synsets correspondant à des concepts-racine

4.1.2. Extraction des sous-arbres

Dans cette seconde étape, nous réalisons tout d'abord 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 précédemment appliquées.

Puis pour chaque ancre, nous recherchons les dérivations qui mènent de celle-ci à la (ou les) racine(s) $rootA$ précédemment identifiée(s). Ces dérivations sont construites en recherchant dans WordNet les hypernymes de chacune des ancres, jusqu'à atteindre la (ou les) racine(s) $rootA$ 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 → ... → Living thing → Entity

Sens 2 : cantaloupe → sweet melon → melon → edible fruit → green goods → food

Seules les dérivations contenant la (ou l'une des) racine(s) $rootA$ sont retenues car elles correspondent au seul sens pertinent pour l'application. Le regroupement des dérivations sélectionnées autour de la (ou des) racine(s) $rootA$ permet de construire un (ou des) sous-arbre(s) T_{WN} (cf. Fig. 5 ou Fig.6). Un sous-arbre T_{WN} se compose d'un concept racine, $rootA$, des feuilles correspondant aux ancres des concepts issus des deux ontologies initiales (cercles blancs pour les ancres des concepts de O_{Src} et cercles grisés pour celles de O_{Tar} dans les Fig.5 et 6) et des généralisants intermédiaires extraits de WordNet qui peuvent, ou non, appartenir à l'une des deux ontologies.

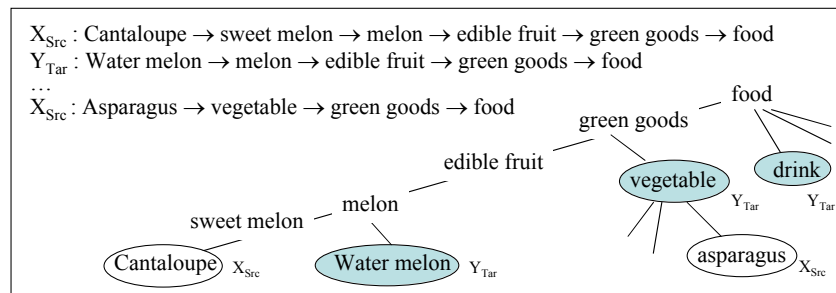


Figure 5. Un exemple de sous graphe T_{WN} de racine *food*

Dans l'expérimentation menée sur les ontologies Russia évoquée précédemment, les racines (Location, Living Thing, Structure et Body of Water) choisies pour couvrir les thématiques de la cible ne sont des généralisants d'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 aux sous-arbres T_{WN} (cf. Fig.6). 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.

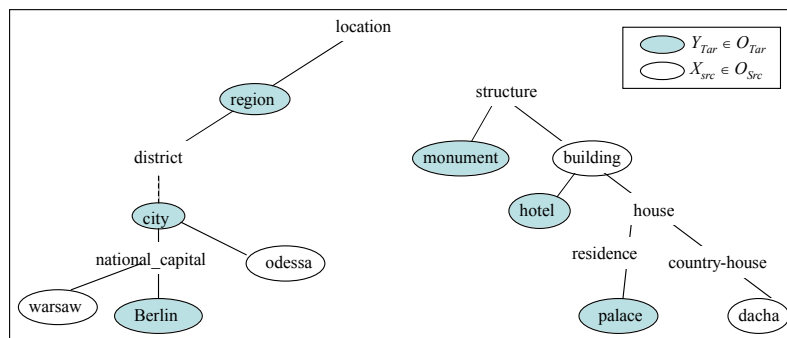


Figure 6. Deux sous-arbres T_{WN} associés aux deux racines *location* et *structure*

4.2. Identification des mappings

Les sous-arbres T_{WN} extraits de WordNet peuvent être exploités de deux façons distinctes :

- en considérant les arcs entre deux nœuds des sous-arbres comme autant de relations d'hyponymie, ce qui permettra d'extraire des mappings « sémantiques », c'est-à-dire reliant explicitement deux concepts par un lien dont la sémantique est bien définie, ici de type '*isA*' ;
- en utilisant une mesure de similarité entre nœuds d'un même graphe qui permettra d'identifier au sein d'un graphe, le nœud le plus proche d'un nœud considéré et d'établir entre les deux nœuds une relation de proximité, dénotée '*isClose*'.

La première méthode consiste ainsi à rechercher dans T_{WN} , pour chaque ancre d'un concept de O_{Src} , son plus proche hyperonyme qui soit une ancre d'un concept de O_{Tar} . Par exemple, le mapping (asparagus *isA* vegetable) peut être dérivé au sein du sous-arbre représenté Fig. 5. Cette méthode est comparable, dans ses résultats, à celle mise en œuvre dans les travaux présentés dans le paragraphe 2.3, 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 reliant un concept de la taxonomie source à un concept de la taxonomie cible considéré comme plus général.

Remarquons que cette première méthode ne permet pas d'identifier de mapping pour le concept cantaloupe car aucun de ses ancêtres dans le sous-arbre 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.

La seconde méthode permet d'effectuer ce rapprochement en utilisant, comme dans (Kalfoglou *et al.*, 2005), une mesure de similarité entre nœuds d'un même graphe. Nous avons aussi choisi dans *TaxoMap* la mesure de Wu et Palmer, car une étude des propriétés de cette mesure 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} (Reynaud et Safar, 2007).

Il est clair que les rapprochements effectués à partir d'une mesure de similarité ne permettent pas d'établir de mappings « sémantiques ». Il est tout aussi clair qu'il serait dommage de ne pas exploiter l'information trouvée ! Les relations de proximité identifiées et dénotées par la relation '*isClose*' ne seront donc retenues que comme des « mappings potentiels » devant être validés par un expert.

Dans *TaxoMap*, nous adoptons une approche combinant l'utilisation de ces deux méthodes. Pour chaque concept X_{Src} de O_{Src} qui reste à apparier, nous recherchons

tout d'abord le concept Y_{Sim} de O_{Tar} qui lui est le plus similaire suivant la mesure de Wu et Palmer, et nous construisons le mapping potentiel associé ($X_{Src} \text{ isClose } Y_{Sim}$). Dans un deuxième temps, nous extrayons, s'il existe, et comme nous l'avons décrit plus haut, le mapping sémantique reliant X_{Src} à son plus proche hyperonyme dans le sous-arbre 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 pouvons aussi 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 conservons, en revanche, le mapping potentiel (*cantaloupe isClose Watermelon*).

5. Expérimentations

Nous avons réalisé plusieurs séries d'expérimentations de la technique présentée. La première série a été effectuée sur une paire de taxonomies portant sur le même domaine, très ciblé, du risque alimentaire, pour laquelle l'identification d'une unique racine (*food*) couvrant tous les concepts des deux taxonomies était immédiat. Plusieurs autres séries d'expérimentations ont été effectuées sur des paires de taxonomies, dont la paire de taxonomies *Russia*, servant de test dans la communauté appariement et de domaine beaucoup plus général, imposant l'utilisation de plusieurs racines.

Dans les différentes séries d'expérimentations, nous avons testé, entre autres, deux façons d'effectuer l'ancrage des concepts dans WordNet : l'une, sans méthode d'ancrage particulière, en ne s'appuyant que sur la capacité d'ancrage de WordNet, qui permet par exemple d'ancrer directement le terme *poulet* à sa forme lexicalisée *poultry*, l'autre, en employant une méthode d'ancrage du type inclusion de labels. Les expérimentations montrent que l'inclusion de labels n'est pas une méthode d'ancrage adaptée à WordNet.

5.1. Appariement de taxonomies relevant d'un domaine ciblé

Le terme *food* semblant un bon généralisant de tous les concepts du domaine, la première expérimentation réalisée dans le domaine du risque alimentaire a consisté à tester si la technique pouvait être employée comme seule technique d'alignement de concepts. Elle a donc été appliquée sur tous les concepts de O_{Src} en s'appuyant sur une méthode d'ancrage du type inclusion de labels. La très faible précision des résultats obtenus est largement due à la longueur des labels des concepts du 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. Pour l'exemple précédent, 3 ancrés (salad, mayonnaise et chicken) sont identifiées qui mènent toutes les trois à des résultats erronés alors que les autres techniques de *TaxoMap* non utilisées dans cette expérimentation donnent des résultats plus pertinents en le rapprochant du concept mixed salad with chicken.

Deux autres expérimentations ont été réalisées en utilisant à chaque fois la technique s'appuyant sur WordNet en complément des autres techniques d'alignement de *TaxoMap* (donc sur les seuls concepts non encore appariés), mais en faisant varier la façon de réaliser les ancrages. Comme les autres techniques d'alignement de *TaxoMap* se basent justement sur la longueur des labels pour déduire leur similarité, la technique a été appliquée uniquement sur les concepts non encore appariés, ayant le plus souvent des labels courts, et les résultats sont plus pertinents. Sans méthode d'ancrage particulière, sur 29 concepts testés non encore appariés, la technique 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 un ancrage par inclusion de labels, 9 mappings supplémentaires sont identifiés (25% cider *isA* drink, almond paste *isA* ingredient, ... pumkin pie *isA* pastry) mais 2 sont incorrects (apple cider *isClose* vegetable, pumkin pie *isClose* meat pie).

5.2. Appariement de taxonomies relevant d'un domaine très général

Les expérimentations précédentes ayant montré que la technique d'utilisation de connaissances de support ne devait pas être employée comme seule méthode d'alignement, les expérimentations suivantes ont été faites en l'intégrant en complément des autres techniques d'alignement de *TaxoMap*, donc en ne l'appliquant que sur les concepts de l'ontologie source non encore appariés par les autres techniques.

Toutes les 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, dans ce cas, le concept de WordNet généralisant les concepts à apparier est l'un des concepts les plus généraux de WordNet, le concept *Entity*, et le sous-arbre construit, T_{WN} , est très gros. T_{WN} est ainsi composé de presque 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 les résultats des trois expérimentations effectuées sur les taxonomies Russia. La première expérimentation est faite en n'utilisant qu'une racine unique *Entity*, les deux suivantes utilisent plusieurs racines et font simplement varier la technique d'ancrage.

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'avons vu Fig.2. Remarquons qu'utiliser une racine unique très générale revient à prendre en compte tous les sens d'un terme et à ne pas faire d'étape de désambiguïsation.

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

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

Une amélioration sensible de la précision⁵ 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 2^{ème} colonne de Tab.1), 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 des 11 mappings *isClose* nous ont paru corrects.

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. 3^{ème} colonne de Tab.1., 7 mappings *isA*

⁵ L'alignement de référence n'étant pas complet, nous ne savons pas combien de mappings auraient dû être identifiés et le rappel est incalculable.

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

Cet article concerne les travaux d'alignement d'ontologies utilisant des connaissances complémentaires, dites de background ou de support, représentées très souvent sous la forme d'une troisième ontologie. L'analyse des travaux présentés montre que l'application de telles techniques est intéressante en complément de techniques terminologiques ou structurelles ou lorsque ces dernières techniques ne s'appliquent pas du fait de la pauvreté des ontologies devant être alignées : ontologies réduites à des taxonomies et/ou peu structurées. Le schéma général du processus d'alignement est aujourd'hui clair pour l'ensemble de ces travaux : une phase d'ancrage pour lier les concepts des ontologies à apparier avec ceux de la 3^{me} ontologie, suivie d'une phase de dérivation qui recherche des relations entre les points d'ancrage pour inférer des relations entre les éléments des ontologies à apparier.

Toutefois, le recours à des connaissances externes est délicat car les résultats de l'alignement obtenu dépendent de la plus ou moins grande proximité sémantique de ces connaissances et de celles des ontologies à aligner. Le recours à des ontologies ciblées facilite cette adéquation lorsque de telles ontologies sont disponibles. L'exploitation de ressources généralistes est plus souple car utilisable quel que soit le domaine d'application des ontologies à aligner mais nécessite d'être attentif au contexte d'interprétation des concepts manipulés pour ne pas mêler des sens différents qui aboutiraient à faire des rapprochements erronés entre concepts. Dans ce papier, nous nous sommes plus particulièrement intéressés à ce problème dans le cadre du système d'alignement *TaxoMap* lorsque la ressource complémentaire utilisée était WordNet.

WordNet est souvent utilisé comme ressource complémentaire car c'est une bonne source d'information sur les synonymies. Il permet, par ailleurs, de disposer

d'une hiérarchie de concepts basée sur les relations de généralisation/spécialisation entre synsets. En revanche, les expériences ont montré qu'il est difficile d'en extraire des relations pertinentes sans prendre en considération le sens précis des termes manipulés. La solution mise en œuvre dans *TaxoMap* pour répondre à ce problème est basée sur l'identification des thèmes traités dans les ontologies à appairer. La phase de dérivation est ensuite restreinte à la seule partie de la hiérarchie de WordNet regroupant l'ensemble des concepts pertinents de l'application. Nous proposons une méthode d'identification de mappings sémantiques au sein de cette partie et, lorsqu'il n'existe pas de mappings sémantiques, nous proposons d'identifier des mappings « potentiels » traduisant des relations de proximité entre concepts.

Les résultats des expérimentations ont montré 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. Elle permet d'identifier, avec une bonne précision, des mappings supplémentaires qui ne peuvent pas être identifiés par les autres techniques, et augmente ainsi l'efficacité globale du système. Par ailleurs, les expérimentations montrent que l'approche est prometteuse même quand le domaine de l'application est très large. En particulier, la solution consistant à identifier les différents thèmes sur lesquels portent les ontologies pour se restreindre aux parties de la hiérarchie de WordNet concernant ces thèmes, améliore de façon sensible les résultats. Ainsi, nous envisageons de continuer à étudier cette voie dont les résultats sont encourageants. Notre objectif est maintenant d'explorer les solutions qui permettraient d'identifier automatiquement les concepts représentant les thèmes traités dans l'ontologie.

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, p. 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, p. 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*: p. 236-243.
- Bidault A., Froidevaux Ch., Safar B., « Repairing Queries in a Mediator approach », In *Proceedings of ECAI'00*, p. 406-410, Berlin, August 2000.
- Giunchiglia F., Shvaiko P., Yatskevich M., « Discovering Missing Background Knowledge in Ontology Matching », In *Proceedings of ECAI 06*, p. 382-386, Trento, Italy.

- Giunchiglia F., Shvaiko P., Yatskevich M., « S-Match: an algorithm and an implementation of Semantic Matching », In *Proceedings of ESWC'04*, p. 61-75
- Gracia J., Lopez V., D'Aquin M., Sabou M., Motta E., Mena E., « Solving Semantic Ambiguity to Improve Semantic Web based Ontology Matching », In *Proceedings of ISWC'07 Workshop on Ontology Matching (OM-07)*, Pusan, Korea, November 2007.
- Kalfoglou Y., Hu B., « Crosi Mapping System (CMS) Result of the 2005 Ontology Alignment Contest ». *K-Cap'05 Integrating Ontologies Workshop*, 2005, Banff, Canada, p. 77-85.
- Minker J., « An overview of Cooperative Answering in Databases », In *Proceedings of FQAS'98*, p. 282-285, 1998.
- Mougin F., Burgun A., Bodenreider O., « Using WordNet to improve the Mapping of Data Elements to ULMS for Data Sources Integration », In *Proc. of AMIA 2006*, p. 574-578.
- 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 W1, *Fouille du Web*, Ed. Cépadués, 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, p. 146-171.
- Wu Z., Palmer M. « *Verb semantics and lexical selection* », In 32nd Annual Meeting of The Association for Computational Linguistics, 1994, Las Cruces, p. 133-138.
- Zhang S., Bodenreider O., « NLM Anatomical Ontology Alignment System. Results of the 2006 Ontology Alignment Contest », In *Proc. of ISWC'06 Ontology Matching 2006*.

Article reçu le xx xx xxxx
Article accepté le xx xx xxxx

Chantal Reynaud est professeur en informatique à l'IUT d'Orsay – Université de Paris-Sud et membre du LRI et du projet GEMO de l'INRIA Saclay îLe-de-France. Ses activités de recherche se situent dans le domaine du Web sémantique et s'articulent autour de deux thèmes principaux : la médiation pour le Web sémantique, la recherche d'information sur le Web à base d'ontologies.

Brigitte Safar est maître de conférences en informatique à l'Université Paris-Sud et membre de l'équipe-projet IASI-GEMO commune au LRI et à l'INRIA Saclay Île-de-France. Ses activités de recherche portent sur l'intégration d'informations et le Web sémantique.