Alignement de taxonomies pour l'interrogation de sources d'information hétérogènes

Taxonomy Alignment for querying heterogeneous information sources

H. Kefi B. Safar C. Reynaud

[1]Université Paris-Sud, CNRS (L.R.I.) & INRIA (Futurs)

Bâtiment 490, 91405 Orsay cedex
cr@lri.fr

Résumé

Intégrer des sources d'information hétérogènes permet un accès unifié sans modification du contenu. Les schémas des sources sont mis en correspondance de façon à ce qu'il soit possible d'accéder à tout un ensemble de documents provenant de sources multiples, à partir d'un système d'interrogation unique. La spécification de ces mises en correspondance, ou mappings, est ainsi au cœur de l'intégration. Nous proposons d'utiliser différentes techniques d'alignement de taxonomies pour automatiser leur génération. Ces techniques ont été implémentées dans un outil logiciel TaxoMap qui recherche les mappings et, en cas d'échec, donne des indications pour aider les utilisateurs à les spécifier eux-mêmes. Nous présentons et discutons des résultats issus d'expériences réalisées dans le domaine de la microbiologie. Nous testons TaxoMap sur différentes taxonomies issues de domaines variés.

Mots Clefs

Ingénierie des connaissances, Ontologies, Mise en correspondances de taxonomies, Intégration de sources d'information.

Abstract

Integration is a way of accessing to multiple heterogeneous information sources without changing their content. Typically schemas of the sources are mapped each other, so that all the sources can be accessed from a unique querying system. One of the challenges is to define the mappings in order to integrate the different sources. We explore some alignment techniques to generate semantic mappings automatically. A tool, TaxoMap, finds mappings or suggests indicators in order to help users find mappings. Experiment results in various domains are discussed.

Keywords

Knowledge Engineering, Ontologies, Mappings of taxonomies, Integration of information sources.

1 Introduction

L'intégration d'information est une problématique importante du fait du nombre croissant de sources d'information disponibles via le Web. Un système d'intégration d'information fournit une interface uniforme pour interroger des ensembles de sources d'information pré-existantes ayant été créées de façon indépendante les unes des autres. Il est basé sur un schéma médiateur unique, l'ontologie, qui représente le vocabulaire du domaine, qui permet de décrire le contenu des sources à intégrer et d'exprimer les requêtes des utilisateurs.

Notre travail se situe dans ce cadre et porte sur l'intégration d'une nouvelle source d'information à un médiateur existant. Le médiateur et la source considérée couvrent un même domaine d'application, mais tous deux ont été développés indépendamment et possèdent leur propre ontologie. La spécificité de notre travail réside dans les faits suivants. D'une part, les utilisateurs sont habitués à formuler leurs requêtes uniquement dans les termes de l'ontologie du médiateur et une exigence forte consiste à préserver ce mode d'interrogation. D'autre part, les fournisseurs d'accès à la nouvelle source ne souhaitent pas modifier leur ontologie. Le problème est donc de lier l'ontologie de la nouvelle source à celle du médiateur afin de pouvoir accéder à son contenu. De tels liens sont possibles grâce à des mises en correspondance, ou mappings, établis entre les éléments des deux ontologies. Nous proposons de calculer ces mappings et de les stocker pour que le moteur d'interrogation puisse ensuite les exploiter.

Dans nos travaux, les entrées du processus de découverte des mappings sont des taxonomies de concepts d'un domaine de spécialité dont les noms peuvent être des expressions composées de plusieurs mots. Les données ne font pas partie des entrées.

Les sorties correspondent aux résultats d'un processus d'alignement de taxonomies. Il ne s'agit pas de fusion de taxonomies. Les taxonomies en entrée ne subissent aucune modification. L'intégration est réalisée à l'aide de

mappings sémantiques qui sont des mises en correspondance de type 1:1 entre les éléments des taxonomies définies en entrée du processus.

Nous proposons une approche générique pour découvrir de façon automatique des mappings entre des éléments complexes de taxonomies. Nous procédons en deux temps, un processus totalement automatique n'étant pas concevable du fait de la très grande hétérogénéité sémantique entre les sources. Dans un premier temps, des mappings probables sont automatiquement découverts. Dans un deuxième temps, des indicateurs sont proposés ainsi que des mappings potentiels 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, structurelles et sémantiques. 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 noms des concepts, en particulier dans les domaines où les homonymes sont rares. Les techniques structurelles permettent de trouver des mises en correspondance potentielles du fait de la position dans la taxonomie des concepts candidats au mapping lorsque l'exploitation de leur syntaxe ne suffit pas. Enfin, lorsque terminologiquement, ou structurellement, il n'est pas possible de trouver de mappings, les techniques sémantiques montrent tout leur intérêt.

Le papier est organisé de la façon suivante. La section 2 décrit les travaux déjà réalisés dans le domaine. L'approche que nous proposons pour aligner des taxonomies de classes est présentée en section 3. Les techniques terminologiques conduisant à la découverte de mappings probables sont présentées en section 4. La section 5 décrit deux autres techniques sémantiques, menant à des indicateurs pour aider l'utilisateur à trouver des mappings complémentaires. La section 6 présente des résultats expérimentaux. Enfin en section 7, nous concluons et présentons quelques perspectives.

2 Travaux proches

Les mappings sont souvent générés manuellement. Ce processus est extrêmement fastidieux même s'il est facilité par des outils d'édition sophistiqués. Plusieurs synthèses de travaux portant sur le matching automatique de schémas [13] ou d'ontologies [6] ont été récemment publiées. Les techniques décrites exploitent différents types d'information, les noms des éléments, les types des données, la structure de la représentation des éléments des schémas, les caractéristiques des données, etc. [8], [18], [2], [15]. Les prototypes réalisés sont presque tous basés sur plusieurs critères de mise en correspondance, par exemple les noms des éléments couplés avec la structure de la représentation.

Ainsi, Noy et Musen dans Anchor-PROMPT rapprochent des ontologies vues comme des graphes au sein desquels les nœuds sont des classes et les liens sont des propriétés. Le système prend en entrée un ensemble d'ancres qui sont des couples d'éléments liés. Un algorithme analyse les chemins dans le sous-graphe délimité par les ancres et détermine quelles classes apparaissent fréquemment dans des positions similaires sur des chemins similaires. Ces classes correspondent ainsi vraisemblablement à des concepts similaires [12]. Maedche et Staab, dans [9], proposent d'exploiter les relations entre les concepts. Des travaux de Ehrig et Staab sont focalisés sur l'efficacité des algorithmes de génération de mappings [3]. Des heuristiques pour restreindre le nombre de mappings candidats sont utilisées et une ontologie permet de classer les mappings candidats selon qu'ils sont plus ou moins prometteurs. Dans un travail de Maedche et Staab, une mesure globale de similarité entre deux hiérarchies est calculée, consistant à comparer les éléments parents et fils de tous les éléments communs [9]. Cette méthode n'est adaptée que si les hiérarchies sont bien structurées et contiennent de nombreux éléments en commun. Enfin Euzenat et Valtchev dans [4] proposent une mesure de similarité dédiée aux ontologies décrites en OWL-Lite, permettant d'agréger différentes techniques de comparaison exploitant les constructeurs de OWL-Lite dans une mesure commune.

D'autres travaux exploitent les données associées aux ontologies et appliquent des techniques d'apprentissage automatique sur ces données [2].

Enfin, des travaux réalisés à l'université de Trento en Italie mettent l'accent sur les aspects sémantiques et proposent d'utiliser WordNet pour aider à mettre en correspondance sémantiquement des éléments d'une hiérarchie de classification [5].

Notre travail de recherche se différencie des travaux cités précédemment par le fait que les schémas en entrée du processus de mise en correspondance sont des taxonomies. Les taxonomies sont des ontologies très sommaires avec des définitions de concepts très pauvres. Les concepts sont principalement définis par référence à leur terminologie. Ils n'ont pas d'attributs. Leur nom est complexe. Les seuls liens représentés sont des liens taxinomiques. Les taxonomies que nous rapprochons ne sont par ailleurs pas identiques du point de vue de leur structure. Notre problème consiste à rapprocher une taxonomie très peu structurée d'une autre qui l'est davantage. Ainsi, la structure des représentations n'est pas le critère premier sur lequel le processus de mise en correspondance peut être basé alors que les techniques terminologiques appliquées sur les expressions des noms

des concepts sont intéressantes. Par ailleurs, les données ne font pas partie des entrées de notre système.

3 Notre approche

L'intégration prend en entrée deux taxonomies. Une taxonomie est un ensemble de concepts reliés par des relations *is-a*. Les taxonomies peuvent être représentées par des graphes orientés acycliques. Les concepts sont représentés par des nœuds du graphe connectés par les liens orientés de type *is-a*.
Le but est de découvrir l'élément dans la taxonomie cible qui peut être mis en correspondance avec chacun des éléments de la taxonomie source. Il s'agit d'un processus orienté. Le processus d'alignement détermine des mappings qui sont des relations de type 1:1 de deux catégories : des mappings d'équivalence et des mappings de spécialisation.

3.1 Deux catégories de mappings

Les mappings d'équivalence sont des relations d'équivalence. Une relation d'équivalence est un lien entre un élément de la taxonomie source et un élément de la taxonomie cible dont les noms sont identiques ou similaires d'un point de vue terminologique. En effet, les taxonomies couvrent les mêmes domaines d'applications et il n'y a pas d'homonymes ou très rarement.

Exemple : Pork sausage (liver) is-equivalent-to Pork liver sausage. Une requête qui contient Pork liver sausage retournera aussi des documents annotés par Pork sausage (liver) car Pork sausage (liver) est lié par une relation d'équivalence à Pork liver sausage.

Les mappings de spécialisation sont des relations *is-a*, les liens usuels sous-classe/super-classe. Quand ils relient un élément de la taxonomie source à un super-élément de la taxonomie cible, le degré de généralité du lien est supposé être le même que dans le lien *is-a* reliant ce super-élément à d'autres sous-éléments dans la taxonomie cible. Une requête contenant un élément de la taxonomie cible, e_T, retournera des documents annotés par cet élément et également par tous les éléments liés à e_T par une ou un enchaînement de relations *is-a*, ces éléments appartenant soit à la taxonomie source, soit à la taxonomie cible.

Exemple : Asparagus *is-a* Fresh fruit and vegetables. L'élément Asparagus de la taxonomie source sera relié à Fresh fruit and vegetables de la taxonomie cible tout comme Carrots, un autre élément de celle-ci. Une requête qui contient Fresh fruit and vegetables retournera les documents annotés par Fresh fruit and vegetables, Carrots et Asparagus.

3.2 La découverte des mappings

Les mappings sont découverts à partir des résultats d'un processus de mise en correspondance qui spécifie les éléments des taxonomies qui peuvent être liés en les accompagnant d'une valeur de similarité. Cette valeur, entre 0 (forte dissimilarité) et 1 (forte similarité) indique la plausibilité de la mise en correspondance entre les deux éléments considérés. Le scénario idéal correspond à l'obtention d'un mapping probable entre deux éléments, c'est-à-dire ayant des chances importantes d'être pertinent. Cette notion de « mapping probable» correspond à trois cas.
Etant donné un élément de la taxonomie source e_S, *m* sera un « mapping probable » si (1) sa valeur de similarité est très forte (égale à 1 ou supérieure à un certain seuil) ou (2) si sa valeur de similarité est la plus élevée et que le nom de l'élément cible de ce mapping *m* est inclus dans le nom de e_S ou (3) si sa valeur de similarité est significativement plus élevée que celle de tous les autres mappings possibles. Ces trois heuristiques ont été implémentées à l'aide de trois techniques terminologiques, toutes basées sur une même valeur de similarité, $Sim_{LIN\text{-}like}$, présentée en section 4.

Comme les mappings de tous les éléments de la taxonomie source ne peuvent pas toujours être découverts automatiquement, nous proposons un processus de découverte des mappings en 3 étapes. Dans un premier temps, on cherche les mappings probables. D'autres mappings, dits potentiels car n'ayant pas une forte plausibilité, et des indicateurs de mappings potentiels sont ensuite calculés pour tous les autres éléments de la taxonomie source non appariés. Ces informations sont délivrées à l'expert du domaine. La 3ème étape est à la charge de cet expert qui, sur la base des informations qui lui auront été communiquées, doit compléter les mappings trouvés automatiquement par le système.
La recherche des mappings probables est basée sur des techniques purement terminologiques exploitant la richesse des labels des concepts. Les techniques structurelles et sémantiques sont utilisées pour suggérer des mappings potentiels à l'expert lorsque l'application des techniques terminologiques a échoué.

La génération de mappings probables est décrite en section 4. La suggestion d'indicateurs à l'expert est détaillée en section 5.

4 La génération de mappings probables

Etant données deux taxonomies, nous essayons d'apparier chaque élément de la taxonomie source à un élément de la taxonomie cible. L'approche proposée est essentiellement terminologique. Trois techniques sont successivement appliquées. Toutes sont basées sur la mesure de similarité $Sim_{LIN\text{-}like}$. Nous présentons cette mesure de similarité en section 4.1. Nous décrivons ensuite chacune des techniques terminologiques mises en œuvre.

4.1 La mesure de similarité $Sim_{LIN\text{-}like}$

La mesure de similarité $Sim_{LIN\text{-}like}$ est basée sur la technique des n-grammes et est définie à partir de la mesure développée par Lin [7].

$$Sim_{Lin}(x,y) = 2 * \frac{\sum_{t \in tri(x) \cap tri(y)} \log P(t)}{\sum_{t \in tri(y)} \log P(t) + \sum_{t \in tri(y)} \log P(t)}$$

Lin mesure la similarité entre deux d'éléments x et y, Sim_{LIN},(x, y), en se basant sur le nombre de tri-grammes partagés par les noms de ces deux éléments. Les tri-grammes d'une chaîne de caractères sont les séquences de 3 caractères obtenues en déplaçant une fenêtre de 3 cases sur la chaîne considérée.
Exemple : tri-gramme(dried milk) = {dri, rie, ied, ed_, d_m, _mi, mil, ilk} où '_ ' représente l'espace.

Dans la formule, tri(x) représente tri-gramme(x), i.e. l'ensemble des tri-grammes de la chaîne de caractères x et P(t) représente la probabilité d'apparition du tri-gramme t dans les termes du corpus. Cette probabilité est supposée indépendante des autres tri-grammes de la chaîne x et est estimée par la fréquence du tri-gramme dans le corpus. Sim_{LIN} est une adaptation aux tri-grammes de la mesure de similarité appelée coefficient de Jaccard [16] : $Sim_{Jaccard}(A,B) = P(A \cap B)/P(A \cup B)$.

La mesure que nous proposons, $Sim_{LIN\text{-}like,}$ est proche de Sim_{LIN} mais prend en compte des aspects linguistiques classiques comme le fait que certains des mots apparaissant dans un nom de concept peuvent être moins importants que d'autres. Par exemple, dans une taxonomie représentant les aliments, les mots fonctionnels 'with, without, and, no, etc' ont moins de sens que les mots propres du domaine, c'est-à-dire les noms d'aliments. De même, dans un concept comme ground beef salad with mayonnaise, le groupe nominal ground beef salad est plus important pour la classification que la composante complément with mayonnaise.

$$Sim_{Lin\_Like}(x,y) = \frac{2 * \sum_{t \in Inter} \log P(t) + 0,5 * \sum_{t \in I'} \log P(t)}{\sum_{t \in tri(y)} \log P(t) + \sum_{t \in tri(y)} \log P(t)}$$

Dans $Sim_{LIN\text{-}like}$ l'ensemble des tri-grammes communs de la formule de Lin est partitionné en deux sous ensembles. I' contient les trigrams communs extraits à partir de mots considérés comme moins importants alors que INTER comprend les autres. Un coefficient, dont la valeur doit être ajustée à partir d'expérimentations, diminue le poids de I' dans la formule. La valeur retenue dans nos propres expérimentations (0,5) est relativement basse par rapport au coefficient (2) dans Sim_{LIN}. En effet, les adjectifs sont nombreux. Plusieurs peuvent être contenus dans le même nom de concept. Il s'agit souvent de chaînes de caractères plus longues que celles des noms. Par ailleurs le suffixe ed à la fin des adjectifs conduit systématiquement à considérer que ed_ est un tri-gramme commun alors qu'il n'est pas pertinent.

Une liste des adjectifs du domaine étudié et une liste des mots fonctionnels ont été établies. Le traitement différencié des mots contenus dans les noms des concepts exploite ces listes et est réalisé par application d'heuristiques générales pouvant s'appliquer à un domaine d'application quelconque. Elles définissent les conditions selon lesquelles le poids des tri-grammes des mots dits « moins importants » doit être diminué. Nous donnons deux de ces heuristiques :

H1 : le poids d'un adjectif commun à deux noms de concepts doit être réduit sauf si un autre mot non fonctionnel est également commun.

Exemple : le poids de l'adjectif cooked dans cooked pork doit être diminué lorsqu'il est comparé à cooked vegetable mais pas lorsqu'on le compare à cooked pork meat.

H2 : le poids d'un mot apparaissant derrière un mot fonctionnel doit être diminué sauf s'il apparaît aussi derrière un mot fonctionnel de même sens dans l'élément comparé et si un autre mot non fonctionnel est commun aux deux éléments.

Exemple : le poids de mayonnaise dans ground beef with mayonnaise doit être diminué lorsqu'il est comparé à t_1 : mayonnaise ou à t_2 : salad with mayonnaise mais pas lorsqu'on le compare à t_3 : mixed beef and mayonnaise. En effet, dans t_1, l'occurrence de mayonnaise n'est pas derrière un mot fonctionnel, et dans t_2, il n'y a pas d'autre mot commun. En revanche dans t_3, l'occurrence de mayonnaise apparaît derrière le mot fonctionnel and, de même sens que with et le mot beef est commun aux deux éléments.

4.2 Les techniques terminologiques

Les techniques terminologiques fournissent l'ensemble des mappings probables. L'algorithme général de recherche de ces mappings est présenté dans Fig. 1.

Dans la première étape, nous calculons la similarité $Sim_{LIN\text{-}Like}$ entre chaque élément de S_S et chaque élément de S_T. Dans la seconde étape, trois techniques sont mises en œuvre

1. Calculer $Sim_{LINLike}$ entre chaque élément de S_s et chaque élément de S_T
2. Pour chaque $e_s \in S_s$
   Si $\exists\ e_T \in S_T$ / TestEquivalent(e_s,e_T) alors stop, retourner : e_s *is-equivalent* e_T
3. Sinon calculer *MC* = MappingCandidates(e_s)
   Si $\exists\ e_T \in$ MC / TestReliable(e_s, e_T) alors stop, retourner : e_s is-a e_T
   Sinon calculer SearchIndicator(e_s)

Fig.1. Algorithme de découverte des mappings d'un élément de S_s

La recherche des mappings d'équivalence

Cette technique recherche les mappings d'équivalence (étape 2 de l'algorithme). S'il existe un élément e_T de S_T tel que sa similarité $Sim_{LIN\text{-}Like}$ avec l'élément concerné e_S de S_S est supérieure à un seuil, alors l'algorithme s'arrête pour cet élément et retourne le mapping d'équivalence, e_S *equivalent-to* e_T.
Exemple: $Sim_{LIN\text{-}Like}(e_S,e_T) = 1.0$ pour e_S = pork liver sausage et e_T = pork sausage liver donc pork liver sausage *is-equivalent* pork sausage liver.

Si pour un élément e_S de S_S, il n'existe pas de mapping d'équivalence avec un élément de la taxonomie cible, les appariements de spécialisation sont recherchés (étape 3 de l'algorithme). Ainsi les étapes suivantes ne sont effectuées que s'il n'y a pas d'élément de S_T qui est équivalent à l'élément e_S considéré.

Les appariements de spécialisation probables correspondent à des relations *is-a*. Ils sont identifiés au sein d'un ensemble de candidats au mapping (*MC*) suivant deux heuristiques : (1) l'une basée sur l'inclusion entre les noms des éléments, (2) l'autre basée sur la similarité relative entre les différents candidats. La première heuristique est appliquée avant la seconde car nous jugeons les appariements obtenus par cette technique plus fiables que ceux issus de la seconde.

La recherche d'inclusion de labels

L'ensemble des candidats au mapping (*MC*) est constitué par l'union de deux sous-ensembles. Le premier est la liste *INC* (pour liste des inclusions) composée des éléments de S_T dont le nom est inclus dans le nom de l'élément e_S considéré, ordonné par valeur de $Sim_{LIN\text{-}Like}$ décroissante. Le second sous-ensemble est composé des trois éléments de S_T qui n'appartiennent pas à *INC* mais qui ont les valeurs de similarité avec e_S les plus élevées (b_1, b_2 et b_3). Nous nommons C_{max} l'élément de *MC* dont la valeur de similarité avec e_S est la plus forte. L'algorithme permettant d'enchaîner ces deux heuristiques est présenté en Fig. 2.

TestReliable(e_S,e_T)
1. Calculer $C_{max} \in MC / \forall C_i \in MC$, $Sim_{LINLike}(e_S,C_{max}) \geq Sim_{LINLike}(e_S,C_i)$
2. Si C_{max} est l'élément avec la plus forte valeur de similarité de *INC*
   alors stop, retourner e_S *is-a* C_{max}
   sinon si $C_{max} = b_1$ et $Sim_{LINLike}(e_S,b_2) / Sim_{LINLike}(e_S,b_1) \leq 0.6$
   alors stop, retourner e_S lié par une relation *is-a* au père de b_1.

Fig.2. Algorithme de découverte des mappings probables d'un élément e_S

Si le label de C_{max} est inclus dans le label de e_S alors e_S est considéré comme une spécialisation de C_{max}. La relation e_S *is-a* C_{max} est dérivée.

Exemple: Soit e_S = beef lean. Les candidats au mapping sont les suivants : *INC* = {beef ($Sim_{LIN\text{-}Like}$: 0.586),…}, b_1 = beef fat ($Sim_{LIN\text{-}Like}$: 0.492), b_2 = roast beef ($Sim_{LIN\text{-}Like}$: 0.373), b_3 = ground beef ($Sim_{LIN\text{-}Like}$: 0.332). C_{max} est le premier élément de *INC*, i.e. beef. Beef est inclus dans le label de e_S qui est beef lean donc nous construisons un appariement de spécialisation entre beef lean et beef.

Notons qu'une heuristique permet d'éviter de considérer un élément inclus comme un père s'il apparaît après un mot fonctionnel ou à l'intérieur d'une expression entre parenthèses. Ainsi, home-style salad with chicken (reduced calorie mayonnaise), par exemple, ne sera pas reconnu comme une spécialisation de mayonnaise ou de chicken mais pourrait être reconnu comme une sorte de salade.

La recherche des meilleurs mappings relativement aux autres possibles

Si le label de C_{max} n'est pas inclus dans celui de e_S ($C_{max} = b_1$) et si la similarité de Cmax est significativement plus élevée que celle de b_2 ($Sim_{LIN\text{-}Like}(e_S, b_2)$/ $Sim_{LIN\text{-}Like}(e_S, b_1) \leq 0.6$), alors une relation *is-a* est construite entre e_S et le généralisant de l'élément b_1 de S_T suivant en cela l'heuristique qui dit que deux éléments similaires mais non équivalents ont un généralisant commun.

Exemple : Soit e_S = cooked cured pork. Les candidats au mapping sont : *INC* = {pork ($Sim_{LIN\text{-}Like}$: 0.354),…}, b_1 = cooked pork meat ($Sim_{LIN\text{-}Like}$: 0.694), b_2 = pork meat ($Sim_{LIN\text{-}Like}$: 0.308), b_3 = pork rind ($Sim_{LIN\text{-}Like}$: 0.292). C_{max} est b_1, i.e. cooked pork meat et $Sim(e_s, b_2)$/ $Sim(e_s, b_1)$ = 0.444 ≤ 0.6. Comme meat based product est le père de cooked pork meat, nous construisons le mapping de specialisation : cooked cured pork *is-a* meat based product.

Si la similarité de C_{max} n'est pas significativement plus élevée que celle de b_2, nous ne pouvons pas identifier de mappings probables. Le choix de l'appariement sera effectué par l'expert du domaine en fonction des indicateurs qui lui seront proposés.

Exemple : Soit e_s = beef adipose tissue. Les candidats au mapping sont : *INC* = {beef (sim : 0.222), …}, b_1 = pork meat tissue ($Sim_{LIN\text{-}Like}$: 0.444), b_2 = beef connective tissue ($Sim_{LIN\text{-}Like}$: 0.434), b_3 = beef fat ($Sim_{LIN\text{-}Like}$: 0.207). Puisque $Sim_{LIN\text{-}Like}(e_s, b_2)/Sim_{LIN\text{-}Like}(e_s, b_1) = 0.974 \geq 0.6$ aucun mapping probable n'est identifié. D'autres techniques doivent être utilisées pour aider l'expert du domaine à choisir le mapping pertinent.

5 Suggestion d'indicateurs

Les deux techniques présentées ci-dessous sont utilisées quand des appariements probables n'ont pu être identifiés. Etant donné un élément e_S de S_S, ces techniques recherchent l'élément e_T de S_T le plus proche, i.e. qui pourrait être un frère ou un père de e_S dans la taxonomie (cf. section 5.1.) ou sans pouvoir préciser le type de relation exacte reliant les deux termes considérés (cf. section 5.2.). A ce jour, les résultats obtenus avec ces techniques sont considérés comme des indicateurs d'appariement possible. Ils sont proposés à l'expert avec l'ensemble des candidats au mapping. L'expert doit décider quel appariement est le plus approprié ou si aucun ne l'est, sur cette base.

5.1 Recherche d'ancêtres communs

Dans cette première technique, nous travaillons sur les candidats au mapping identifiés précédemment. L'idée consiste à utiliser leur position dans la taxonomie cible. Le sous-graphe représentant les éléments de *MC* au sein de S_T est analysé. Dans le meilleur des cas, si tous les éléments ont le même père dans S_T, l'élément e_S considéré a aussi probablement le même père. Dans le cas

contraire, tous les éléments de *MC* n'ayant pas le même père, nous cherchons leur plus petit ancêtre commun. Si cet ancêtre commun est un nœud très haut placé dans la taxonomie, il ne sera pas très significatif car trop général. Pour obtenir des suggestions d'appariement plus pertinentes, nous recherchons alors des ancêtres partiels, i.e. des nœuds qui ne sont l'ancêtre que d'un sous-ensemble d'éléments de *MC*. Ainsi pour chaque sous-ensemble d'éléments de *MC* partageant un ancêtre commun partiel, nous construisons le sous-graphe correspondant qui a pour racine l'ancêtre partiel considéré.

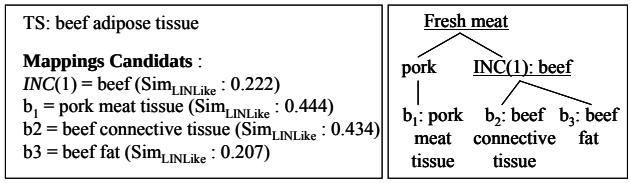

Fig. 3. Recherche d'ancêtres communs partiels

Les sous-graphes sont construits comme suit. Nous identifions le plus petit ancêtre commun (Lowest Common Ancestors, LCA) de chaque paire d'éléments de *MC* et nous étendons l'ensemble considéré élément par élément. Pour un sous-graphe dont la racine est l'ancêtre partiel *Anc*, nous calculons la proximité *P(Anc)* entre les éléments de *MC* correspondant aux nœuds de ce sous-graphe (MC_{Anc}).

$$P(Anc) = \frac{|MC| * \sum_{e_t \in MC_{Anc}} dist(e_t, Anc)}{|MC_{Anc}| * \sum_{e_t \in MC_{Anc}} sim_{LIN\_Like}(e_s, e_t)}$$

La proximité est calculée en fonction du nombre total d'éléments de *MC* sur le nombre d'éléments MC_{Anc} couverts par le sous-graphe. Nous calculons aussi la somme des mesures de similarité des éléments de MC_{Anc} et la distance en nombre d'arcs de chaque élément de MC_{Anc} à l'ancêtre commun partiel. Le sous-graphe qui a la plus faible valeur de proximité est considéré comme le plus pertinent. Dans ce sous-graphe le plus pertinent, le nœud qui a la valeur de similarité la plus élevée est noté C_{MaxAnc} et est considéré comme le candidat au mapping le plus proche de e_S. Si C_{MaxAnc} appartient à l'ensemble des termes inclus, *Inc*, il est considéré comme un père possible de e_S dans S_S. Dans le cas contraire, C_{MaxAnc} est considéré comme un frère possible, et son père (qui n'est pas forcément le nœud racine du sous-graphe) est proposé comme un père possible de e_S.

Par exemple, dans Fig. 3, Fresh meat est le plus petit ancêtre commun des quatre candidats au mapping, avec une distance de 7 (2+2+2+1). Mais un autre sous-graphe de racine beef et regroupant trois des candidats au mapping {beef, beef connective tissue, beef fat} peut être construit avec une distance égale à 2. Le groupement représenté par ce dernier sous-graphe est donc jugé plus pertinent et le nœud de ce sous-graphe qui a la valeur de similarité la plus élevée, beef connective tissue, est proposé comme un frère de beef adipose tissue.

5.2 Utilisation d'une source linguistique

Les techniques décrites précédemment ne sont pas suffisantes quand les concepts sont sémantiquement proches mais que leur nom est différent. Ainsi, aucune de ces techniques ne permet de rapprocher cantaloupe et watermelon alors que l'interrogation d'une source linguistique, telle que WordNet [11], peut indiquer que ces concepts correspondent tous deux à des sortes de melons et sont sémantiquement proches. Il est donc possible, en utilisant WordNet, de trouver pour chaque élément de la taxonomie source de quels éléments de la taxonomie cible il peut être sémantiquement rapproché (ceux avec lesquels ils partagent des généralisants) puis d'identifier l'élément dont il est le plus proche. Toutefois cette méthode ne peut en général pas être utilisée seule pour apparier les termes de deux taxonomies, car WordNet ne contient que des termes du vocabulaire courant alors que les hiérarchies étudiées contiennent de très nombreux termes spécialisés. Ainsi, cette technique n'est appropriée que pour apparier des termes simples dont le label contient peu de mots.

WordNet est une base de données 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).

Un certain nombre d'autres travaux utilisent aussi WorNet pour apparier des concepts. Ainsi dans [15], les auteurs étendent 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 ». Dans notre contexte, l'utilisation des synonymes n'est pas suffisant. Pour reprendre notre exemple initial, les concepts cantaloupe et watermelon ne sont pas des synonymes dans WordNet (aucun n'appartient au synset de l'autre) mais ils sont tous deux des spécialisations d'un même concept, melon.

L'utilisation de WordNet ne peut donc pas être immédiate et s'effectue en deux étapes. La première étape consiste à construire un sous-arbre (appelé S_{WN}) à partir de

l'ensemble des généralisants dans WordNet de chacun des éléments des deux taxonomies S_T et S_S que l'on cherche à rapprocher. La seconde étape consiste à utiliser une mesure de similarité ([14], [17]) pour identifier l'élément de la taxonomie cible S_T le plus proche de l'élément de S_S considéré, au sein de la taxonomie commune S_{WN}.

Pour construire S_{WN}, nous interrogeons WordNet pour chaque élément de chacune des deux taxonomies. Pour chaque élément, et pour chacun de ses sens (chaque synset auquel il appartient), nous extrayons l'ensemble de ses généralisants jusqu'à atteindre le concept le plus général pour l'application considérée. Par exemple, le résultat de la recherche sur le terme cantaloupe donne les deux ensembles de généralisants suivants qui correspondent à deux sens différents du terme.

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

Seul est conservé le sens qui contient le concept racine de l'application étudiée (sens 2 dans l'exemple car il contient food). Les généralisants de l'élément sont alors intégrés pour construire le sous-arbre de WordNet pertinent pour l'application, i.e., dont la racine est le terme le plus général de l'application. Les feuilles sont les termes issus des deux taxonomies initiales (dans des cercles dans la Fig. 4). Les nœuds intermédiaires peuvent être indifféremment des termes de WordNet ou de l'une des deux taxonomies.

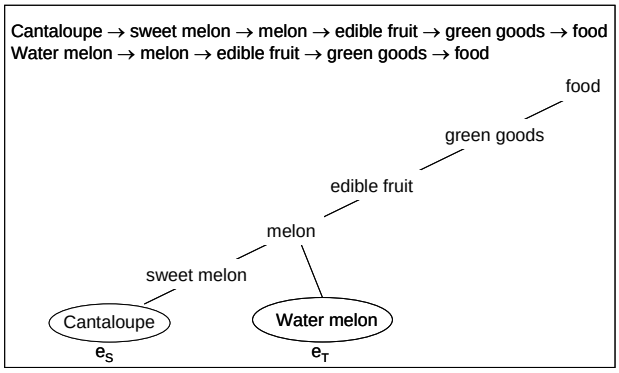

Fig. 4. Un sous-graphe de S_{WN} reliant cantaloupe et water melon dont la racine est food

Une fois S_{WN} construit, nous évaluons la similarité sémantique entre deux nœuds n_1 et n_2 de l'arbre en utilisant la mesure de similarité proposée par Wu & Palmer [17] : $Sim_{WP}(n_1,n_2) = 2 * depth(LCA(n_1,n_2)) /(depth(n_1) + depth(n_2))$ où $depth(n_i)$ correspond au nombre d'arcs apparaissant sur le chemin menant du nœud n_i à la racine de l'arbre et où $LCA(n_1, n_2)$ est le plus petit ancêtre commun des deux nœuds n_1 et n_2 comme nous l'avons défini plus haut.

Ainsi pour chaque élément e_S de la taxonomie cible reconnu par WordNet, la méthode renvoie l'élément de la taxonomie source e_T pour lequel la mesure de similarité de Wu & Palmer est la plus élevée. Si l'élément e_S n'est pas reconnu par WordNet, ce qui est le cas de la quasi totalité des éléments dont le label contient plus de trois mots, des heuristiques permettant de découper le label et de re-soumettre automatiquement à WordNet les différentes parties ont été mises en œuvre, mais les résultats ne sont pas concluants pour l'instant, et ne sont pas pris en compte par le système.

6 Expériences

Notre travail de recherche a été guidé par son application dans le cadre du projet e.dot. Nous présentons le contexte du projet dans une première sous-section. Nous présentons et discutons ensuite les résultats obtenus dans les expérimentations réalisées.

6.1 Contexte d'expérimentation

Le but du projet e.dot[1] est de proposer des solutions génériques pour enrichir un entrepôt de données avec les données du Web. Le domaine d'application cible est le risque alimentaire. Dans le cadre de ce projet, notre objectif a été de permettre l'interrogation unifiée des sources d'information Sym'Previus et Com'Base, contenant toutes deux des documents utiles aux experts en microbiologie. L'interrogation doit se faire à partir du moteur d'interrogation MIEL [1] basé sur l'ontologie Sym'Previus. Une requête formulée uniquement à l'aide du vocabulaire de Sym'Previus doit retourner des documents de Sym'Previus annotés avec les termes de la requête ou des termes plus spécifiques et également des documents de Com'Base annotés avec des termes liés par des mappings à ceux de la requête ou aux termes plus spécifiques de Sym'Previus.

L'ontologie de Sym'Previus est une taxonomie composée de 460 termes reliés par des liens de subsomption. Les termes sont organisés selon 7 niveaux. Chaque terme de Sym'Previus dispose d'une liste de synonymes et d'une traduction anglaise.

Com'Base a été construite par une équipe anglaise. Le schéma associé comprend 172 termes liés par des liens de subsomption. Les termes sont organisés selon une taxonomie qui n'est pas très structurée. Seuls deux niveaux existent, le premier niveau comprenant uniquement 12 termes.

Les deux taxonomies ont des parties qui se recouvrent mais ne sont pas structurées de la même façon. D'une part, elles représentent des points de vue propres. D'autre part, le niveau de granularité choisi n'est pas le même car le nombre des termes représentés est différent.

[1] projet RNTL (2003-2005).

6.2 Résultats d'expérimentations

Un prototype, TaxoMap, a été implémenté en Java. Il a permis de réaliser des expérimentations sur les taxonomies Sym'Previus et Com'Base dans le cadre de e.dot ainsi que sur d'autres taxonomies publiées sur le Web [19] pour lesquelles certains mappings sont fournis.

Afin d'évaluer les résultats obtenus en sortie de TaxoMap, un expert en microbiologie a défini manuellement les mappings entre Com'Base (S_{So}) et Sym'Previus (S_{Ci}), soit pour 172 éléments (nombre d'éléments de Com'Base). Le tableau ci-dessous donne le nombre d'éléments de Com'Base pour lesquels un mapping a été trouvé, manuellement et automatiquement en précisant la catégorie du mapping trouvé.

Catégorie des mappings	Nombre de mappings trouvés manuellement	Nombre de mappings trouvés automatiquement
Equivalence	44	28
Classification	121	115
Indéterminé	-	20
Inexistant	7	9

Tab. 1. Type et nombre de mappings donnés par l'expert et le système

Plusieurs expérimentations ont été réalisées avec TaxoMap. Les résultats qui figurent dans le tableau Tab. 1. ont été obtenus avec un seuil de 1.0 ce qui signifie que les mappings d'équivalence ont une valeur de similarité $Sim_{LIN\text{-}Like}$ de 1. La figure 5.a présente quelques mappings d'équivalence indiqués par l'expert mais non trouvés avec un seuil de 1.0 et la figure 5.b présente des mappings d'équivalence faux qui seraient trouvés avec un seuil de 0.8.

.

$Sim_{LINLike}$(watermelon, water melon) = 0.848
$Sim_{LINLike}$(broccoli, brocoli) = 0.709
$Sim_{LINLike}$(tofu, soya tofu) = 0.679
$Sim_{LINLike}$(beefburger, hamburger) = 0.543
$Sim_{LINLike}$(yogurt, yoghurt) = 0.417

Fig. 5a. Faux négatifs avec un seuil de 1.0

$Sim_{LINLike}$(milk chocolate, chocolate) = 0.901
$Sim_{LINLike}$(25% apple juice, apple juice) = 0.899
$Sim_{LINLike}$(broth (culture medium), culture medium) = 0.852
$Sim_{LINLike}$(cooked pork, cooked pork meat) = 0.874
$Sim_{LINLike}$(water, water of cow) = 0.811

Fig. 5b. Faux positifs avec un seuil de 0.8

Le seuil de 1.0 est restrictif, les chaînes de caractères comparées devant être rigoureusement identiques. En contrepartie, ces mappings sont sûrs (absence de faux positifs c'est-à-dire de mappings faux découverts automatiquement) dans le contexte dans lequel on travaille (absence d'homonymes). Les 28 mappings d'équivalence automatiquement découverts sont tous pertinents comme le montre Tab. 2.

Les mappings non découverts par cette technique du fait du seuil élevé (16 mappings. cf. Tab. 2.), mais proches terminologiquement (cf. Fig. 5a), le seront par les deux autres techniques terminologiques. Il ne s'agira alors pas de mappings d'équivalence mais de mappings de classification. Par exemple, yogurt et yoghurt ne seront pas considérés comme étant équivalents mais comme deux fils d'un même père, soit deux spécialisations d'un même concept. Le nombre plus élevé de mappings d'équivalence trouvés par l'expert provient aussi du fait qu'il considère comme équivalents des concepts sémantiquement identiques tout en étant terminologiquement différents. TaxoMap identifient certains d'entre eux via WordNet (cf. 5.2) lorsque les labels comportent au maximum 3 mots.

	Mappings d'équivalence	Mappings de classification
Nombre de mappings découverts manuellement	44	121
Nombre de mappings découverts par le système	28	115
Nombre de mappings pertinents découverts par le système	28	96
Précision	1.00	0.83
Rappel	0.64	0.79

Tab. 2. Précision et rappel par catégorie de mappings

L'avantage de TaxoMap est de fournir une catégorisation des mappings selon la façon dont ils ont été obtenus. Ceci est important aux yeux de l'expert puisque chaque ensemble de mappings n'a pas la même vraisemblance. Une telle catégorisation peut accélérer le processus de validation.

Le tableau Tab. 3. permet d'analyser les mappings de classification découverts automatiquement. On constate que les deux techniques terminologiques (inclusion de labels et similarité relative) ne fournissent que très peu de mappings non pertinents.

Techniques	Nombre de mappings découverts	Nombre de mappings pertinents découverts	Précision
Inclusion de labels	62	58	0.93
Similarité relative	11	10	0.91
Ancêtres communs	42	28	0.67

Tab. 3. Analyse des mappings de classification par technique

La technique des ancêtres communs est moins sûre (précision de 0.67), c'est pourquoi les résultats obtenus ne font pas partie des mappings dits « probables ». Ce sont

des indicateurs délivrés à l'expert pour qu'il juge de leur pertinence. Les mappings non pertinents trouvés par la technique des ancêtres communs s'expliquent en partie par la difficulté à traiter les labels comportant beaucoup de mots et surtout à distinguer, parmi ces mots, ceux qui font référence au concept sous-jacent et ceux qui ne font que le caractériser. Par exemple, dans rice with chicken protein, le mot pertinent est rice (4 caractères), les autres mots sont secondaires tout en étant des termes de la taxonomie Sym'Previus. Des erreurs peuvent aussi venir du fait que peu de mots de l'élément de S_S sont des mots d'éléments de S_T. C'est le cas pour home-style macaroni salad car seul salad est un mot de Sym'Previus. Enfin, de mauvais rapprochements sont effectués lorsque la seule chaîne commune est courte et est incluse dans d'autres mots plus longs. Par exemple, past est commun à fig past et almost paste qui sont tous deux liés à pasteurized.

La technique des ancêtres communs délivre deux types d'information : (1) un mapping potentiel (sélectionné étant donnés les critères sur lesquels la technique est basée) et (2) un ensemble de mappings candidats. Dans les expérimentations réalisées, un seuil de 0.15 a été retenu pour sélectionner les mappings candidats à communiquer. Ce seuil semble bien adapté. Il évite à la fois d'avoir des ensembles de mappings vides et des ensembles trop importants composés de concepts très éloignés sémantiquement.

La technique sémantique basée sur l'utilisation de WordNet s'est révélée être tout à fait complémentaire des techniques précédemment utilisées. 20 mappings ont été trouvés. Cette technique a pris tout son intérêt dans l'expérimentation lorsqu'il s'agissait de labels comportant au plus 3 mots. Ainsi, 50% cider a pu être lié à drinks et frankfurter à sausage.

9 mappings n'ont pas été découverts automatiquement. Les expérimentations ont permis de les expliquer. Il s'agit très souvent de labels contenant des mots rares ou des sigles qui ne sont pas dans Sym'Previus.

Dans un second temps, nous avons testé les techniques terminologiques et structurelles de TaxoMap sur d'autres ontologies publiées sur le Web [19] pour lesquelles des mappings étaient donnés (cf. Tab. 4.).

	Yahoo	Standard	Washington	Cornell
Nombre de termes	114	332	167	167
Nombre de mappings fournis	29		54	

Tab. 4. Description des taxonomies Yahoo-Standard-Washington-Cornell

Techniques	Yahoo-Standard	Washington-Cornell
Equivalence	17	12
Inclusion de labels	1	3
Similarité relative	4	10
Ancêtres communs	7	25

Tab. 5. Nombre de mappings trouvés par technique

Le tableau Tab. 5. indique le nombre de mappings pertinents retrouvés avec chacune des techniques, tant terminologiques que structurelles. On constate que les techniques terminologiques permettent de retrouver l'intégralité des mappings fournis pour Yahoo/Standard et seuls 4 mappings sur 54 ne sont pas retrouvés pour Washington/Cornell. Cela montre la complémentarité de ces techniques, chacune contribuant à la découverte d'une catégorie particulière de mappings. Le processus de génération de mappings ainsi proposé, pour ce qui concerne les taxonomies testées et relativement aux mappings fournis, s'est avéré efficace.

7 Conclusion et perspectives

Notre objectif est de construire un système qui génère automatiquement des mappings entre deux taxonomies. Nous avons implémenté un prototype, TaxoMap, qui utilise une approche exploitant la syntaxe des noms des concepts, la structure des taxonomies et un thesaurus, WordNet, pour établir automatiquement des mappings entre des éléments de deux schémas. Les mappings se distinguent selon leur plausibilité. Seuls les mappings probables sont découverts automatiquement. Lorsque le système ne peut pas identifier de mapping probable, il aide l'utilisateur à le faire en lui indiquant un ensemble de mappings candidats. TaxoMap a été testé dans le cadre du projet e.dot avec deux taxonomies réelles dans le domaine de la microbiologie et également sur des taxonomies servant de tests aux travaux réalisés dans le domaine [2], [5].

Ce travail est original parce qu'il pose le problème de la génération de mappings lorsque l'on dispose de peu de critères. En effet, les schémas sont des taxonomies au sein desquelles les concepts sont définis surtout par rapport à la terminologie du nom qui leur est associé et dont les niveaux de profondeur peuvent ne pas être très nombreux. Les expériences réalisées montrent qu'une approche enchaînant des techniques particulières adaptées au contexte décrit, terminologiques, structurelles et sémantiques, peut donner de bons résultats.

L'approche que nous proposons dans ce papier s'applique à des taxonomies. Nous envisageons de poursuivre ce travail en testant encore l'approche sur d'autres domaines d'application. Dans un second temps, nous pensons adapter la méthode à des taxonomies très structurées afin d'exploiter de façon plus importante les techniques structurelles. Enfin, à plus long terme, l'approche sera élargie pour générer automatiquement des mappings entre ontologies plus riches, non restreintes à des hiérarchies de concepts, et représentées en RDF(S) ou OWL.

Bibliographie

[1] P. Buche, J. Dibie-Barthélemy, O. Haemmerlé, M. Houhou. "Towards Flexible Querying of Imprecise Data in a Data warehouse Opened on the Web". In Proceedings of the 6th International Conference on Flewible Query Answering Systems, FQAS'04, Lyon,

France, Lecture Notes in AI #3055, Springer, pp. 28-40, 2004.

[2] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, A. Halevy, "Learning to match ontologies on the Semantic Web", The VLDB Journal, 12:303-319, 2003.

[3] M. Ehrig, S. Staab, "QOM – Quick Ontology mapping", in proc. of the 3rd International Semantic Web Conference (ISWC2004), November 7 to 11, Hiroshima, Japan, 2004.

[4] J. Euzenat, P. Valtchev, "An integrative proximity measure for ontology alignment", in Proc. ISWC 2003.

[5] F. Giunchiglia, P. Shvaiko, M. Yatskevich, "S-Match: an algorithm and an implementation of semantic matching", In proceedings of the European Semantic Web Symposium, LNCS 3053, pp. 61-75, 2004.

[6] Y. Kalfoglou, M. Schorlemmer, "Ontology mapping: the state of the art", in Knowledge Engineering Review, Vol. 18, pp. 1-31, 2003.

[7] D. Lin, "An Information-Theoretic Definition of Similarity", in proc. of the International Conference on Machine Learning – ICML-98, Madison, July 98, pp. 296-304, 1998.

[8] J. Madhavan, P.A. Bernstein, E. Rahm, "Generic Schema Matching with Cupid", VLDB 2001.

[9] A. Maedche, S. Staab, "Measuring similarity between Ontologies", in proc. of the European Conference on Knowledge Acquisition and management – EKAW-2002, Madrid, Spain, October 1-4, LNCS/LNAI 2473, Springer, 2002, pp. 251-263, 2002.

[10] S. Melnick, H. Molina-Garcia, E. Rahm, "Similarity flooding: a versatile graph matching algorithm", in proc. of the International Conference on Data Engineering – ICDE, San José, pp. 117-128, 2002.

[11] G.A. Miller, "WordNet: A lexical Database for English", Communications of the ACM, Vol. 38, N°11, p. 39-45, November, 1995.

[12] N. Noy, M. Musen, "Anchor-PROMPT: Using Non-Local Context for Semantic Matching", IJCAI 2001.

[13] E. Rahm, P. Bernstein, "A survey of approaches to automatic schema matching", VLDB Journal, 10(4), 334-350, 2001.

[14] P. Resnik, "Using information content to evaluate semantic similarity in a taxonomy", in proc. of the International Joint Conference on Artificial Intelligence – IJCAI-95, Montreal, pp. 448-453.

[15] B. Than Le, R. Dieng-Kuntz, F. Gandon, « On Ontology Matching Problems - for Building a Corporate Semantic Web in a Multi-Communities Organization ». ICEIS (4) 2004: pp. 236-243.

[16] CJ. Van Rijsbergen, "Information Retrieval", 2nd edn. Butherworths, London, 1979.

[17] Z. Wu, M. Palmer, "Verb semantics and lexical selection", in proc. of the 32nd Annual Meeting of Computational Linguistics, Las Cruces, 1994, pp. 133-138, 1994.

[18] L.L. Yan, R.J. Miller, L.M. Haas, R. Fagin, "Data-Driven Understanding and Refinement of Schema Mappings". ACM SIGMOD, 2001.

[19] http://anhai.cs.uiuc.edu/archive, Illinois Semantic Integration Archive.