# Alignement des ontologies : utilisation de WordNet et une nouvelle mesure structurelle.

**Aissa Fellah[1], Mimoun Malki[1], Ahmed ZAHAF[1]**

[1] Université Djillali Liabes de Sidi Bel abbés, département d'informatique, Algérie.
{ammfellah@yahoo.fr}

*RÉSUMÉ. L'interopérabilité sémantique entre sources d'information hétérogènes est une problématique importante du fait du nombre croissant de sources d'information disponibles sur le web. L'utilisation des ontologies est une voie très prometteuse pour permettre l'interopérabilité, seulement les ontologies eux même peuvent être hétérogènes. L'alignement des ontologies est le noyau de cette interopérabilité, cependant la génération automatique des correspondances entre deux ontologies est d'une extrême difficulté qui est dû aux divergences (conceptuelle, habitudes, etc.) entre communautés différentes de développement des ontologies. Ce travail est une proposition d'un algorithme d'alignement de deux ontologies de même domaine en utilisant différentes techniques, en particulier nous utilisons WordNet et nous introduisons une nouvelle mesure de similarité structurelle entre deux entités de deux ontologies déférentes qui est basée sur la position structurelle des entités à comparer au sein de leurs ontologies.*

*MOTS-CLÉS : Ontologie, Hétérogénéité, Interopérabilité sémantique, Alignement, Correspondances, Mesure de similarité structurelle.*

*ABSTRACT. Semantic interoperability between heterogeneous sources of information is significant problems because of the number growing of sources of information available on the Web. The use of ontology is a very promising way to allow interoperability, only ontology them self can be heterogeneous. The alignment of ontology is the core of this interoperability, however the automatic generation of the mappings between two ontology is of an extreme difficulty which is due to the divergences (conceptual, practices, etc.) between communities different of development of ontology. This work is a proposal of an algorithm of alignment of two ontology of the same field by using various techniques, in particular we use WordNet and we introduce a new measure of structural similarity between two entities of two deferent ontology which is based on the structural position of the entities to compare within their ontology.*

*KEYWORDS: Ontology, Heterogeneity, Semantic Interoperability, Alignment, Mappings, Structural Similarity measure.*

## 1. Introduction

L'alignement des ontologies est une tache cruciale dans plusieurs domaines d'application. A titre non exhaustif, on peut citer : le web sémantique, la communication dans les SMA (système muti-agents), data warehouse (Shvaiko et al.,2005), l'intégration des schéma/Ontologies, etc. Une ontologie est définie comme la conceptualisation des objets reconnus comme existant dans un domaine, de leurs propriétés et des relations les reliant (Furst,2004). Le problème actuel est qu'étant donné un même domaine ou des domaines connexes, il est possible que plusieurs ontologies soient disponibles (car développées simultanément par plusieurs communautés différentes). La comparaison de deux ontologies, le passage de l'une à l'autre ou de les intégrer devient donc nécessaire. Plusieurs techniques d'alignement, basées sur des critères différents, sont actuellement proposées dans la littérature, des travaux récents (Shvaiko et al.,2005), (Kalfoglou et al.,2003), (Euzenat et al.,2004) présentent une synthèses des techniques. Le choix d'une technique ou d'une autre ou la composition de plusieurs d'entre elles n'est pas une tâche aisée. Plusieurs travaux complètent leurs résultas d'alignement par l'utilisation de WordNet(Miller et al.,1993) comme ressource externe. Nous combinons l'utilisation de WordNet avec une généralisation d'une mesure de proximité sémantique pour améliorer la pertinence des correspondances résultats du procédé d'alignement. Ce travail est organisé de la façon suivante : la section 2 est un état de l'art qui présente l'alignement et se focalise sur les travaux proches. Dans la section 3, nous décrivons notre approche d'alignement au sein de laquelle s'insère notre nouvelle mesure de similarité structurelle. La section 4 est une étude expérimentale qui discute les résultats et les performances de notre méthode. Enfin nous concluons dans la section 5.

## 2. Etat de l'art

### 2 .1 Alignement

L'objectif du processus d'alignement est de gérer le plus automatiquement possible, des appariements sur des ontologies, consiste à trouver des correspondances entre les connaissances spécifiées dans les deux ontologies, de manière à pouvoir les exploiter conjointement dans le même système (Euzenat et al.,2004). En pratique, il s'agit d'identifier des concepts (ou des relations) de la première ontologie avec des concepts (ou des relations) de la seconde. Dans les deux cas, la connexité des deux domaines de connaissance modélisés par les ontologies est requise, sans quoi aucun lien ne peut être établi entre concepts (Kefi et al.,2006). De plus, les formalismes de représentation d'ontologie utilisés doivent être au moins compatibles, ainsi que les paradigmes conceptuels (Furst,2004). Les méthodes appliquées pour repérer les similarités entre concepts et/ou relations sont (Euzenat et al.,2004) : -les méthodes terminologiques qui comparent les **labels** désignant deux concepts ou deux relations; -les méthodes qui comparent les **propriétés internes** des concepts et relations (attributs des concepts, portée d'une relation, etc..); -les méthodes qui comparent les

**propriétés externes** des concepts et relations (subsomptions, relations entre concepts, etc.); -les méthodes qui comparent les **extensions** des concepts et relations; -les méthodes qui comparent la **sémantique** des concepts et relations. Ces méthodes peuvent bien entendu être combinées entre elles. Elles peuvent parfois recourir à des ressources extérieures aux ontologies à aligner.

### 2.2 Travaux proches

Les mappings sont souvent générés manuellement. Ce processus est extrêmement fastidieux même s'il est facilité par des outils d'édition sophistiqués. Les techniques 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. 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 (Noy et al.,2001). 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 (Maedche et al.,2002). Enfin Euzénat et Valtchev dans (Euzenat et al.,2003) 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.

## 3. Approche

Le processus d'alignement est orienté d'une ontologie source vers une ontologie cible, et il a pour objectif de générer deux types de relations : des relations d'équivalence et des relations de spécialisation. Notre technique commence par un prétraitement sur les termes de chaque ontologie, ce travail préalable est important pour améliorer les résultats d'alignement. La première phase est basée sur les chaînes de caractères en particulier la mesure de Levenstein (Levenshtein,1966) peut déterminer les relations d'équivalence et l'heuristique d'inclusion des labels peut déterminer les relation de spécialisation entre concepts. La deuxième phase se concentre sur l'utilisation d'une ressource externe qui est WordNet, ainsi par l'utilisation de la synonymie on peut déduire des équivalences entre concepts dont les labels ne sont pas syntaxiquement similaires. Finalement on utilise une technique structurelle qui exploite les résultats des phases précédentes, et combine la structure des deux ontologies et WordNet, pour trouver d'autres mappings potentiels. Pour la génération des correspondances lors de la phase structurelle, on propose une nouvelle mesure structurelle entre deux ontologies qui est une généralisation de la mesure de Wu & Palmer (Wu et al.,1994 ).

### 3.1 Technique terminologique

Les méthodes terminologiques comparent des chaînes de caractères, plusieurs idées ont été développées dans la littérature en utilisant les comparaisons linguistiques des termes. Elles peuvent être appliquées au nom, à l'étiquette ou aux commentaires au sujet des entités pour trouver ceux qui sont semblables. Cette technique inspirée de l'hypothèse suivante **: deux termes sont similaire veut dire qu'ils dénotent les concepts similaires** (Euzenat et al.,2004). Il y a plusieurs manières d'évaluer la similarité entre deux entités. La manière la plus commune est de définir une quantité à une mesure de cette similarité. En premier lieu on applique des techniques qui considèrent les labels comme une suite de caractères et permettent de relier les concepts dont les labels sont rigoureusement identiques syntaxiquement, l'application de la mesure de similarité Edit distance (distance de Levenstein), cette mesure est basée sur la même hypothèse : **deux termes sont similaires s'ils partagent assez d'éléments importants.** Une version normalisée de (Maedche et al., 2002) :

$$\mathbf{Sim}_{\mathbf{syn}}\mathbf{(t1, t2) = max(0, } \frac{\mathbf{min(|t1|,|t2|) - ed(t1, t2)}}{\mathbf{min(|t1|,|t2|)}}\mathbf{)} \ ;$$

Si Simsyn(c1,c2)=val > **seuil** alors on peut déduire une correspondance d'équivalence de la forme (c1, c2 , = , val). Après expérimentation un seuil est définit pour accepter les couples de termes syntaxiquement rapprochés.

### 3.2 Utilisation de WordNet

WordNet est une ressource lexicale de langue anglaise, 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. Les synsets sont reliés entre eux par des relations sémantiques: relation de généralisation / spécialisation, relation composant/composé. Les techniques basées sur les chaînes de caractères ne sont pas suffisantes quand les concepts sont sémantiquement proches et quand leurs noms sont différents, l'interrogation d'une ressource linguistique tel que WordNet peut indiquer que les concepts sont similaires. Pour le calcul de la similarité linguistique la fonction Syn(c) calcul l'ensemble des Synsets de WordNet du concept c; soit S=Syn(c1) ∩ Syn(c2) l'ensemble des sens communs entre c1 et c2 à comparer, la cardinalité de S est :

$$\lambda\,(S) = |Syn(c1) \cap Syn(c2)| \ ;$$

Soit min(|Syn(c1)|,| Syn(c2)|) le minimum entre les cardinalité des deux ensembles Syn(c1) et Syn(c2) alors la similarité entre deux concepts c1 et c2 sera définie comme suit :

$$\mathbf{Sim}_{\mathbf{ling}}(c1, c2) = \frac{\lambda(S)}{\min(|Syn(c1)|, |Syn(c2)|)} ;$$

Cette mesure retourne 1.0 si au moins c1 est le seul synonyme de c2 ou c2 est le seul synonyme de c1.

### 3.3 Technique structurelle

Les techniques structurelles consistent à exploiter la structure de l'ontologie à comparer, souvent représentées sous forme de graphes, et la comparaison de similarité entre deux entités de deux ontologies peut être basée sur la position des entités dans leurs hiérarchies. Ces techniques implémentent diverses heuristiques et sont basées sur l'hypothèse suivante : (H)-s**i deux entités de deux ontologies sont semblables, leurs voisins le sont également d'une certaine façon**(Euzenat et al.,2004). Notre mesure structurelle est une généralisation de la mesure de Wu & Palmer sur deux hiérarchies. On propose le calcul de la similarité structurelle entre les entités de deux ontologies, on s'inspirant des travaux de (Abolhassani et al.,2006). Cette mesure est évaluée en fonction des relations déjà établies entre les entités des deux ontologies.

**Base intuitive et théorique**

D'après (H), la similarité entre deux concepts c1∈ O1, et c2∈ O2 dépend de la similarité entre leurs voisinages. Autrement dit, si le voisinage d'un concept c1 noté V(c1) est similaire au voisinage du concept c2 noté V(c2) alors c1 et c2 sont similaires d'une certaine manière. Pour simplifier notre étude on considère que le voisinage directe d'un concept qui est constitué de l'ensemble {Père, Fils}, pour le choix des voisinages on prend que le voisinage qui possède un appariement avec un voisinage de la même position relative dans l'ontologie cible (père - père, fils-fils) ; ce qu'on nous appelons voisinage relatif noté VR(c1,c2). On définit le voisinage d'un concept c d'une ontologie O par :

**V( c ) = { tout concept c'∈ O / c'=père( c ) ou c' ∈ fils( c ) };**

Et l'ensemble du voisinage relatif du couple (c1,c2) (avec c1∈ O1 et c2 ∈ O2) est définit par les couples (rc1,rc2) tel que les conditions suivantes seront vérifiées : 1)-rc1 ∈ O1 et rc2 ∈ O2 ; 2)-rc1 et rc2 sont déjà identifiés comme similaires l'un à l'autre soit par une entrée utilisateur soit par une mesure linguistique dont la similarité est supérieure à un seuil définit ;3)-la position de c1 relative à rc1 est la même que pour c2 par rapport à rc2.

$$\mathbf{VR(c1,c2) = \{(rc1,rc2) \ telque \ rc1 \in V(c1) \ et \ rc2 \in V(c2) \ et \ Sim(rc1,rc2) >= seuil\}}$$

Par la comparaison structurelle on essai de quantifié la similarité structurelle notée $Sim_{str}(c1,c2)$ entre c1∈ O1 et c2∈ O2 qui dépend naturellement de la similarité des couples du voisinage relatif $(cr1_i,cr2_i) \in VR(c1,c2)$ avec i=1..|VR(c1,c2)| selon

l'heuristique structurelle exposée ci-dessus. Notre quantification de cette mesure prend en compte les mesures entre voisinages relatives et la mesure de la proximité sémantique entre concept et son voisinage relatif. Il est difficile de choisir entre les mesures de similarité sémantique, laquelle fournira les résultas les plus pertinents dans notre cas. Nous nous basons ici sur la mesure de Wu & palmer appliquée à WordNet, choisie on raison de sa simplicité d'implémentation.
Mesure de Similarité de Wu & Palmer ($Sim_{wp}$):

$$\mathbf{Sim_{wp}(c1,c2) = \frac{profondeur(LCA(c1,c2))}{profondeur(c1) + profondeur(c2)}\ ;}$$

LCA(c1,c2) est le petit ancêtre commun (Lowest Common Ancestors de c1 et c2 et profondeur(LCA(c1,c2)) est nombre d'arcs séparant LCA(c1,c2) de la racine et profondeur( c1) est la longueur du chemin séparant c1 de la racine en passant par LCA(c1,c2). Notre mesure est ainsi formulée :

$$\mathbf{Sim_{str}(c1,c2) = \frac{\sum_{(rc1,rc2)\in VR(c1,c2)} sim_{ling}(rc1,rc2) * sim_{wp}(c1,c2) * (1 - |d1 - d2|^2)}{|VR(c1,c2)|}}$$

avec : $\mathbf{d1 = sim_{wp}(c1,rc1)}$ et $\mathbf{d2 = sim_{wp}(c2,rc2)}$

$Sim_{ling}(c1,c2)$ est soit la similarité syntaxique (mesure de Levenstein) ou la similarité lexicale calculer en utilisant WordNet ou encore une similarité introduite par l'utilisateur. Si d1=d2 (même granularité) et $sim_{ling}(c1,c2)=1$ alors notre mesure coïncide avec la mesure de Wu & Palmer, $sim_{str}(c1,c2)= sim_{wp}(c1,c2)$
**Exemple :** on tente de rapprocher les deux concepts Catalog et (DataBase ou Book) par notre méthode structurelle :

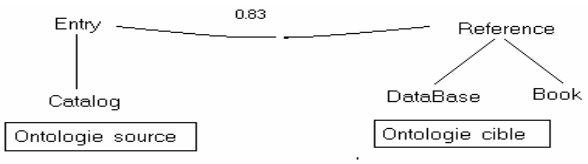


**Figure 1.** *Exemple d'appariement structurel*

Soit c1=Catalog et c2=Book ou DataBase;
supposons que $Sim_{lin}$(Entry,Reference)=0.83 (sont déjà similaires) ;
-**Cas-1 :** VR(c1,c2)=VR(Catalog,Book ) ={(Entry,Reference)}; |VR(c1,c2)|=1 ;
$Sim_{WP}$ (Entry, Catalog)=0.66=d1; $Sim_{WP}$ (Reference, Book)=0.95=d2;
$sim_{WP}$ (Catalog, Book)=0.95;
Et finalement $Sim_{str}$(Catalog,Book ) =0.83*0.95*(1-|0.66-0.95|$^2$ )=0.722
-**Cas-2 :** VR(c1,c2)=VR(Catalog, DataBase ) ={(Entry,Reference)}; |VR(c1,c2)|=1 ;
$Sim_{WP}$ (Entry, Catalog)=0.66=d1; $Sim_{WP}$ (Reference, DataBase)=0.66=d2;
$sim_{WP}$ (Catalog, DataBase)=0.88;

Et finalement $Sim_{str}$ (Catalog, DataBase) =0.83*0.88*(1-|0.66-0.66|$^2$ )=0.730

## 4. Expérimentations

Pour évalué les performances de notre algorithme, un prototype nommé OA, réalisé en JAVA. Notre outil supporte en entée deux ontologies décrites en OWL (étant donné que le langage OWL constitue un standard pour la représentation des ontologies), et produit un fichier XML, contient les mappings résultats. On a procédé à une série de tests on utilisant quelques tests fournies dans la base Benchmark mise à la disposition de la communauté internationale par la compétition EON (Eon, 2007). L'ontologie de base est constituée par un ensemble de références bibliographiques. Le tableau suivant récapitule les résultats des premiers tests obtenus par notre algorithme d'alignement en se basant sur les valeurs des métriques de la qualité d'alignement (mesures de précision, rappel, Fallout et Fmesure) :

| Test | Précision | Rappel | Fallout | Fmesure |
|---|---|---|---|---|
| **101-103-104-203-204-207-221** | 1,00 | 1,00 | 0,00 | 1,00 |
| **201** | 1,00 | 0,97 | 0,00 | 0,98 |
| **202** | 0,25 | 0,03 | 0,75 | 0,05 |
| **205** | 0,94 | 1,00 | 0,06 | 0,97 |
| **222** | 0,85 | 0,97 | 0,15 | 0,91 |
| **224-225-230-232-233-236-237** | 1,00 | 1,00 | 0,00 | 1,00 |
| **239** | 0,88 | 1,00 | 0,12 | 0,94 |
| **247** | 0,92 | 1,00 | 0,08 | 0,96 |
| **249** | 0,33 | 0,03 | 0,67 | 0,06 |
| **258** | 0,67 | 0,15 | 0,33 | 0,25 |
| **259** | 0,63 | 0,19 | 0,38 | 0,29 |
| **301** | 0,95 | 0,82 | 0,05 | 0,88 |
| **304** | 0,97 | 0,91 | 0,03 | 0,94 |

**Tableau 1.** *Résultats de nos tests*

La technique structurelle s'avère complémentaire pour les techniques terminologiques.

## 5. Conclusion

L'alignement des ontologies représente un grand intérêt pour plusieurs domaines d'applications qui manipulent des connaissances hétérogènes. Dans ce travail nous avons proposé un algorithme d'alignement des ontologies, qui combine plusieurs techniques, les techniques terminologiques et l'utilisation de WordNet comme ressource complémentaire, avec l'introduction d'une nouvelle mesure structurelle

entre deux ontologies qui est une généralisation de la mesure de Wu & Palmer. Notre travail est loin d'être terminé, plusieurs améliorations sont possibles pour rendre notre technique plus pertinente.

## Bibliographie

Abolhassani H. , B.B. Hariri, S. H. Haeri, "On Ontology Alignment Experiments", Webology, Volume 3, Number 3, September, 2006

Eon, « EON 2007 : Evaluation of Ontology for the Web », *Proceedings of the 5th International EON Workshop* http://oaei.ontologymatching.org/2007/benchmarks, Ontology Alignment Evaluation Initiative Test library ,2007

Euzenat J., Barrasa J., Bouquet P., Dieng R., Ehrig M., Hauswirth M., Jarrar M., Lara R., Maynard D., Napoli A., Stamou G., Stuckenschmidt H., Shvaiko P., Tessaris S., van Acker S., Zaihrayeu I., Bach T. L., D2.2.3: State of the art on ontology alignment. Technical report,NoE Knowledge Web project delivable, 2004. http://knowledgeweb.semanticweb.org/.

Euzenat J., Valtchev P., Similarity-based ontology alignment in OWL-lite. In *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, pages 333–337, 2004.

Furst F., Contribution à l'ingénierie des ontologies : une méthode et un outil d'opérationnalisation, thèse de doctorat,Université de Nantes ,2004.

Kalfoglou Y., Schorlemmer M., Ontology mapping: the state of the art. *The Knowledge Engineering Review Journal (KER)*, (18(1)):1–31, 2003.

Kefi H. , Safir B., Reynaud C.," Alignement de taxonomies pour l'interrogation de sources d'information hétérogènes" INRIA, 2006.

Levenshtein, I. (1966), 'Binary code capable of correcting deletions, insertions and reversals', Cybernetics and Control Theory 10(8), 707–710.

Maedche A., 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.

Miller G., al., Introduction to WordNet: An on-line lexical database, MIT Press,1993.

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

Shvaiko P., Euzenat J., A survey of schema-based macthing approaches. *Journal on Data Semantics (JoDS)*, IV, 2005.

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