EDOLA : Une nouvelle méthode d'alignement d'ontologies OWL-Lite

Sami Zghal[*,**] — **Karim Kamoun**[**] — **Sadok Ben Yahia**[**]
Engelbert Mephu Nguifo[*] — **Yahya Slimani**[**]

[*] *CRIL CNRS FRE 2499, Université d'Artois, IUT de Lens*
Rue de l'Université - S.P. 16, 62307 Lens Cedex, France
{sami.zghal, mephu}@cril.univ-artois.fr

[**] *Département des Sciences de l'Informatique, Faculté de Sciences de Tunis*
Campus Universitaire, 1060 Tunis, Tunisie
{sadok.benyahia, yahya.slimani}@fst.rnu.tn

RÉSUMÉ. *L'alignement d'ontologies revêt toute son importance dans des applications nécessitant la prise en compte d'une interopérabilité sémantique. Plusieurs approches d'alignement d'ontologies existent dans la littérature. Elles sont basées sur les mesures de similarités. Dans ce papier, une nouvelle méthode d'alignement d'ontologies OWL-Lite est décrite. Le module d'alignement implémente une nouvelle approche d'alignement d'ontologies qui définit un modèle global de calcul de similarité, tout en remédiant au problème de l'intervention de l'utilisateur dans le processus d'alignement. Les tests expérimentaux réalisés sur les ontologies de benchmark montrent une nette amélioration des métriques de rappel et de précision.*

ABSTRACT. *Ontologies have been established for knowledge sharing and are of extensive use as a means for conceptually structuring domains of interest. Thus, in order to guarantee a fluent global communication and knowledge exchange between local knowledge sketched by ontologies, the alignment of ontologies has emerged as a compelling topic to address. In this paper, we introduce a new approach for aligning OWL-Lite ontologies. The main originality of the alignment method stands in the fact that it palliates the main drawbacks appearing in the literature approaches, i.e., problem of user-parameter settings. Carried out experimental results pointed out a sharp improvement in the precision and recall evaluation metrics.*

MOTS-CLÉS : *Ontologie, Alignement d'ontologies, mesures de similarité, OWL-Lite.*

KEYWORDS: *Ontology, Alignment of ontologies, similarity measure, OWL-Lite.*

CORIA 2007, pp. 351–366
Saint-Étienne, 28–30 mars 2007

1. Introduction

Le terme ontologie a été utilisé initialement en philosophie depuis le $19^{ième}$ siècle. Dans ce domaine, il désigne *l'étude de ce qui existe*, i.e., *l'ensemble des connaissances sur le monde* (Welty *et al.,* 2001). Dans le domaine de la représentation des connaissances, les ontologies ne sont considérées que relatives aux différents domaines de connaissances. Elles répondent aux problèmes de représentation et de manipulation des connaissances. L'ontologie est "*est une spécification explicite d'une conceptualisation*" (Gruber, 1993). Les ontologies sont très utilisées dans la représentation des connaissances sur le Web (Charlet *et al.,* 2004). De nos jours, l'ontologie matérialise une connaissance experte d'un domaine. Partant du fait que plusieurs connaissances peuvent prendre des représentations différentes, on trouve de nos jours plusieurs ontologies de domaine pour un même champ d'application. Les techniques d'alignement représentent un cadre général, dans lequel plusieurs ontologies peuvent être exploitées. L'alignement permet aussi d'échanger, d'un point de vue sémantique, les avis de nombreuses personnes (Bach *et al.,* 2004).

Bien que certains travaux sur les ontologies montrent la nécessité de l'utilisation d'une connaissance du domaine (Aleksovski *et al.,* 2006) dans certaines situations, plusieurs méthodes d'alignement d'ontologies qui n'exploitent pas de connaissances du domaine ont été développés. Les principales méthodes citées sont : ANCHOR-PROMPT (Noy *et al.,* 2001), IF-MAP (Kalfoglou *et al.,* 2003), ASCO (Bach *et al.,* 2004), GLUE (Doan *et al.,* 2004), QOM (Ehrig *et al.,* 2004a) et , OLA (Touzani, 2005, Euzenat *et al.,* 2004b). Ces principales méthodes[1] exploitent des ontologies en format des langages de balises (XML, RDF(S) et OWL-Lite[2]). En outre, la majorité de ces méthodes exploite des mesures de similarité qui couvrent plus ou moins toute la structure des ontologies à aligner. Ces méthodes exploitent d'une manière générale un seuil de stabilité, fourni par l'utilisateur, pour garantir l'arrêt du processus d'alignement. Cependant, ce seuil de stabilisation ne permet pas une propagation étendue pour le calcul de la similarité. La méthode OLA est la seule à présenter l'avantage de la prise en charge d'ontologies en format OWL-Lie. La méthode OLA exploite un seuil de stabilisation pour calculer l'alignement. La nouvelle méthode d'alignement proposée, **EDOLA**, implémente un nouvel algorithme automatique d'alignement d'ontologies OWL-Lite. À chaque couple d'entités appartenant à une même catégorie, l'algorithme d'alignement calcule les mesures de similarités. Il définit deux modèles de calcul de similarité (local et global), tout en remédiant au problème de la circularité et de l'intervention de l'utilisateur dans le processus d'alignement. Les résultats expérimentaux obtenus par EDOLA montrent une amélioration des métriques d'évaluation par rapport à OLA.

L'article est organisé comme suit. La deuxième section propose une étude comparative des principales méthodes d'alignement d'ontologies retenues. Dans la troisième

1. Dans ce papier nous proposons de mettre l'accent sur ces méthodes.
2. Le format OWL-Lite permet de prendre en charge tous les composants d'ontologies : concepts, relations, axiomes, etc.

section, la nouvelle méthode d'alignement d'ontologies OWL-Lite EDOLA est décrite. La quatrième section illustre une évaluation expérimentale. La conclusion et les travaux futurs font l'objet de la dernière section.

2. Étude Comparative des méthodes d'alignement

Les ontologies créées peuvent être décrites en plusieurs langages, *e.g.*, XML (Marsh, 2001), RDF(S) (Klyne *et al.*, 2004), DAML+OIL (Connolly *et al.,* 2001) et OWL (Smith *et al.,* 2004). Le but de ces langages est de représenter les ontologies dans un langage commun. Le langage OWL permet aussi le partage, l'import et l'export d'ontologies. Il est considéré comme le standard des ontologies pour le domaine du Web sémantique (Berners-Lee *et al.,* 2001). Pour ces raisons toute ontologie qui n'est pas décrite en OWL présente des inconvénients. L'alignement de deux ontologies revient à trouver une correspondance entre leurs entités qui sont sémantiquement similaires (Ehrig *et al.,* 2004b). D'une façon formelle, l'alignement est défini par la fonction map comme suit :

$$map : O \longrightarrow O^{'} tel\ que\ map(e_1) = e_1^{'} \quad \text{si } sim(e_1, e_1^{'}) > t,$$

où O et $O^{'}$ sont les deux ontologies à aligner, $\mathtt{t}$ désigne un seuil minimal de similarité appartenant à l'intervalle [0,1], $e_1 \in O$ et $e_1^{'} \in O^{'}$. Ce seuil indique le niveau minimum pour que deux entités soient similaires. Chaque entité e_i est alignée au plus à une seule entité $e_j^{'}$. Plusieurs critères ont été utilisés pour la comparaison des méthodes d'alignement, *e.g.*, le format en entrées, le format en sortie, les mesures de similarité et la qualité de l'alignement (Do *et al.,* 2002).

2.1. *Formats en entrée et en sortie*

Le type de données utilisé doit être précisé pour chaque méthode d'alignement. Les ontologies à aligner peuvent être représentée avec des langages à balises ou au format des graphes conceptuels. Les langages à balises sont XML, RDF(S), DAML+OIL et OWL. Les dictionnaires de synonymies ou de lexiques sont des informations supplémentaires pouvant parfois s'ajouter et qui sont nécessaires pour l'amélioration du rendu du processus d'alignement.

Le format et la structure du résultat de l'alignement sont précisés pour chaque méthode. Il faut préciser si l'alignement s'effectue entre les structures entières ou entre couples d'entités des deux ontologies. Le résultat pour la majorité des méthodes existantes est un fichier d'alignement (généralement en format XML), indiquant quelles sont les couples entités ontologiques qui correspondent. Toutes les méthodes d'alignement déterminent des correspondances entre les entités ontologiques en utilisant des mesures de similarité.

2.2. *Mesures de similarité*

La taxinomie suivante sont proposée pour la classification des différentes mesures de similarité (Rahm *et al.,* 2001) : **(i)** La méthode terminologique (T) : compare les labels des entités. Elle est décomposée en approches purement syntaxiques (TS) et celles utilisant un lexique (TL). L'approche syntaxique effectue la correspondance à travers les mesures de dissimilarité des chaînes (*e.g.*, EditDistance). Tandis que, l'approche lexicale effectue la correspondance à travers les relations lexicales (*e.g.*, synonymie, hyponymie, etc.) ; **(ii)** La méthode de comparaison des structures internes (I) : compare les structures internes des entités (*e.g.*, intervalle de valeur, cardinalité d'attributs, etc.) ; **(iii)** La méthode de comparaison des structures externes (S) : compare les relations d'entités avec d'autres. Elle est décomposée en méthodes de comparaison des entités au sein de leurs taxinomies (ST) et méthodes de comparaison des structures externes en tenant compte des cycles (SC) ; **(iv)** La méthode de comparaison des instances (E) : compare les extensions des entités, *i.e.*, elle compare l'ensemble des autres entités qui lui sont attachées (instances des classes) ; **(v)** La méthode sémantique (M) : compare les interprétations (ou plus exactement les modèles) des entités.

2.3. *Qualité de l'alignement*

Les mesures de *Précision*, *Rappel* et *Fallout* (Do *et al.,* 2002) ont été des métriques largement exploitées pour estimer la qualité des alignements obtenus. Le EON[3] "Evaluation of Ontology-based Tools" (EON, 2004, EON, 2006, Euzenat *et al.,* 2006) retient ces mesures pour l'évaluation de la qualité de l'alignement. L'objectif principal de ces mesures est l'automatisation du processus de comparaison des méthodes d'alignement ainsi que l'évaluation de la qualité des alignements produits. La première phase dans le processus d'évaluation de la qualité de l'alignement consiste à résoudre le problème manuellement. Le résultat obtenu manuellement est considéré comme l'alignement de référence. La comparaison du résultat de l'alignement de référence avec celui de l'appariement obtenu par la méthode d'alignement produit trois ensembles : N_{found}, $N_{expected}$ et $N_{correct}$. L'ensemble N_{found} représente les paires alignées avec la méthode d'alignement. L'ensemble $N_{expected}$ désigne l'ensemble des couples appariés dans l'alignement de référence. L'ensemble $N_{correct}$ est l'intersection des deux ensembles N_{found} et $N_{expected}$. Il représente l'ensemble des paires appartenant à la fois à l'alignement obtenu et l'alignement de référence. La *précision* est le rapport du nombre de paires pertinentes trouvées, *i.e.*, "$N_{correct}$", rapporté au nombre total de paires, *i.e.*, "N_{found}". Il renvoie ainsi, la partie des vraies correspondances parmi celles trouvées. Ainsi, la fonction *précision* est définie par : $précision = \frac{|N_{correct}|}{|N_{found}|}$. Le *rappel* est le rapport du nombre de paires pertinentes trouvées, "$N_{correct}$", rapporté au nombre total de paires pertinentes, "$N_{expected}$". Il spécifie ainsi, la part des vraies correspondances trouvées. La fonction *rappel* est dé-

3. http ://oaei.ontologymatching.org/2004/Contest/ et http ://km.aifb.uni-karlsruhe.de/ws/eon2006/

finie par : $rappel = \frac{|N_{correct}|}{|N_{expected}|}$. La mesure *Fallout* permet d'estimer le pourcentage d'erreurs obtenu au cours du processus d'alignement. Elle est définie par le rapport des paires erronées, "$(N_{found} - N_{correct})$", rapporté au nombre total des paires trouvées, "N_{found}", *i.e.*, $Fallout = \frac{|N_{found}|-|N_{correct}|}{|N_{found}|}$. La table 1 présente une revue récapitulative et transversale des principales méthodes d'alignement sus-discutées. La première entrée de la table 1 présente les formats d'ontologies pris en charge par chaque méthode d'alignement. Ces formats sont en majorité des langages de balises à l'exception de KIF et OCML. La deuxième entrée de la table 1 indique la nature du fichier résultat qui est soit un fichier XML ou un fichier RDF(S). La troisième entrée de la table 1 regroupe les différentes mesures de similarité exploitées au niveau de chaque méthode. La dernière entrée de la table 1 met en exergue les bornes des mesures de *précision* pour chaque méthode dans le cadre des tests réalisés par EON (EON, 2004). Ainsi, la méthode OLA présente un léger avantage comparativement à la méthode QOM. Les performances "qualitatives" de ces méthodes sont presque similaires, puisqu'elles prennent en considération toutes les caractéristiques de l'ontologie à savoir, la similarité terminologique, structurelle et extentionnelle d'entités d'ontologies. En outre, la qualité de l'alignement produit par OLA est meilleure. En effet, la valeur de la précision minimale et la valeur de la précision maximale sont plus élevées que celles fournies par QOM. À noter que, OLA propose une méthode de calcul de similarité qui résout le problème de circularité entre les concepts lors du processus d'alignement (Touzani, 2005, Euzenat *et al.,* 2004b). Le résultat de l'alignement se présente sous la forme d'un fichier RDF/XML.

	GLUE	**ANCHOR-PROMPT**	**QOM**	**OLA**	**IF-MAP**	**ASCO**
Entrée	XML	RDF(S)	RDF(S)	OWL-Lite	KIF, OCML RDF(S)	RDF(S)
Sortie	XML	RDF	RDF	RDF	RDF	RDF
Similarité	E	T, I S, ST	T, TS, I S, ST SC, E	T, TS, I S, ST SC, E	ST, E	T, TS TL, ST
Précision	[0,3-0,6]	[0,1-0,5]	[0,5-0,7]	[0,6-0,8]	-	-

Tableau 1. *Tableau comparatif des principales méthodes d'alignement*

Dans la majorité des principales méthodes d'alignement d'ontologies, la stabilisation de la mesure de similarité est exploitée. Cette mesure de stabilité est fournie par l'utilisateur à travers un seuil. Ce seuil permet la propagation de la similarité pour atteindre l'alignement optimal. Cette propagation risque de ne pas exploiter convenablement le voisinage des différentes entités ontologiques. De cette façon, la méthode d'alignement peut s'arrêter sans explorer d'avantage le voisinage. Cet arrêt est dû au fait que le traitement de deux voisins successifs n'apporte pas un gain inférieur au seuil précisé. De même, l'arrêt limite le traitement des entités intéressantes et risque de nuire au résultat de l'alignement obtenu. Ces inconvénients nous ont encouragé à proposer une nouvelle méthode d'alignement. Le principal avantage réside

dans le fait qu'elle élimine l'intervention de l'utilisateur en exploitant un voisinage plus étendu des entités à apparier. La section suivante introduit la nouvelle méthode d'alignement d'ontologies OWL-Lite développée que nous comparons ensuite avec la méthode OLA.

3. Nouvelle méthode d'alignement d'ontologies

La nouvelle méthode d'alignement d'ontologies, EDOLA, que nous proposons prend en entrée des ontologies décrites en format OWL-Lite. Les ontologies OWL-Lite à apparier sont transformées sous forme d'un graphe OWL-Graph que nous introduisons.

Le graphe OWL-Graph permet de représenter toutes les informations contenues dans l'ontologie OWL-Lite (Smith *et al.,* 2004). Les classes, les propriétés et les instances sont les noeuds du graphe proposé. Les noeuds du graphe OWL-Graph représentent les six types d'entités qui existent dans une ontologie OWL-Lite : les concepts, les instances des concepts, les types de données, les valeurs des types de données et les propriétés des classes (de nature objet et de nature type de données). Les relations entre les entités au niveau de l'ontologie OWL-Lite sont les arcs entre les noeuds du graphe. Les arcs qui existent dans le graphe OWL-Graph reflètent les relations sémantiques qui existent entre les entités d'une ontologie. Le graphe OWL-Graph permet de représenter quatre catégories de liens de *spécialisation*, d'*attribution*, d'*instantiation* et d'*équivalence*. La figure 1 présente un exemple deux ontologies représentées via deux graphes OWL-Graph distincts. La première ontologie indique qu'un enseignant encadre un étudiant qui réalise son mémoire. La seconde ontologie indique qu'un mémoire est réalisé par un étudiant qui est encadré par un enseignant. Les graphes OWL-Graph ainsi obtenus par le module de construction sont exploités par le module d'alignement d'ontologies OWL-Lite, EDOLA (Extended Diameter OWL-Lite Alignment). En effet, le module d'alignement effectue le parcours des deux ontologies représentées sous la forme de deux graphes OWL-Graph. Ce parcours permet de comparer les noeuds et les arcs des graphes pour déterminer les correspondances entre les différentes entités ontologiques en exploitant le diamètre des noeuds. Le diamètre d'un noeud est le nombre de noeuds le séparant de l'extrémité du graphe (les instances).

La nouvelle méthode d'alignement, EDOLA, est une approche se basant sur un modèle de calcul des similarités locale et globale. Ce modèle suit la structure du graphe OWL-Graph pour calculer les mesures de similarité entre les noeuds des deux ontologies. Le module d'alignement associe pour chaque catégorie de noeuds une fonction d'agrégation. La fonction d'agrégation prend en considération toutes les mesures de similarités entre les couples de noeuds voisins au couple de noeud à apparier. Ainsi, cette fonction exploite toute l'information descriptive de ce couple. La table 2 présente les notations utilisées dans les algorithmes développés. L'algorithme qui implémente la méthode d'alignement EDOLA prend en entrée deux ontologies à aligner sous forme de deux fichiers OWL-Lite et produit un résultat sous forme d'un fichier XML.

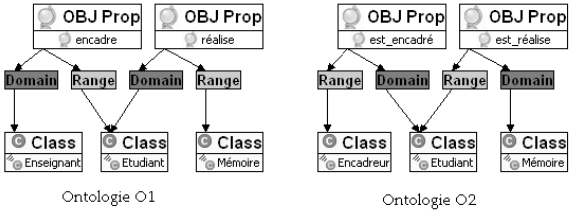

Figure 1. *Exemple de deux graphes OWL-Graph de deux ontologies*

- O_1,O_2 : les deux ontologies à aligner
- VS_T : vecteur de similarité terminologique
- VS : vecteur de similarité sémantique
- VD : vecteur des diamètres minimum respectifs à chaque couple de noeuds

Chaque noeud de l'ontologie présente parmi ses caractéristiques les champs suivants :
- type : le type du noeud
- diamètre : le diamètre du noeud

Chaque élément des vecteurs VS_T et VS se caractérise par les champs suivants :
- le noeud de l'ontologie O_1
- le noeud de l'ontologie O_2
- la valeur de similarité

Tableau 2. *Notations utilisées dans les algorithmes* PHASE1_SIMTERM *et* PHASE2_SIMSEM

EDOLA opère en deux étapes successives. La première étape, implémentée par le biais de la fonction PHASE1_SIMTERM, permet de calculer la similarité locale (terminologique). La deuxième étape, *c.f.* la fonction PHASE2_SIMSEM, permet de calculer la similarité globale, dite *sémantique*.

3.1. *Calcul de similarité locale*

Le calcul de la similarité locale s'effectue une seule fois pour chaque couple de noeuds. La mesure de similarité locale des couples d'entités est calculée par l'intermédiaire de l'algorithme 1 (*c.f.*, la fonction PHASE1_SIMTERM). Le calcul de la similarité locale (ou terminologique) est effectué entre les descripteurs d'entités comme

les noms, les commentaires, etc. La similarité terminologique est composée de la similarité syntaxique et la similarité lexicale. Ainsi, la similarité syntaxique est calculée par l'intermédiaire des fonctions de LEVEINSTEIN ou EditDistance (Euzenat *et al*., 2004a). Tandis que l'API de WORDNET (Miller, 1995) est exploitée pour le calcul de la similarité lexicale. La fonction PHASE1_SIMTERM permet de calculer les similarités terminologiques des couples de noeuds des deux ontologies. Elle prend en entrée les deux ontologies O_1 et O_2 à aligner, représentées sous la forme de deux graphes OWL-Graph, ainsi que la fonction de similarité terminologique à utiliser et donne en retour un vecteur de similarité terminologique de chaque couple de noeuds. La fonction ***CalculSimTerm*** (Algorithme 1, ligne 8) prend en entrée deux noeuds N_1 et N_2, et retourne une valeur de similarité. Cette fonction est assurée par l'une des méthodes de calcul de similarité suivante : la mesure de LEVENSHTEIN, la distance des sous-chaînes ou l'API de WORDNET. La similarité locale pour les différents couples d'entités est exploitée par la suite pour le calcul de la similarité globale. La section suivante décrit en détail le processus de calcul de la similarité globale.

```
1  Fonction : PHASE1_SIMTERM
   Données : (1) Deux ontologies O1 et O2 (2) Fonction de similarité terminologique
   Résultats : Vecteur de similarité locale VS_T
2  début
3      /* parcours des noeuds de l'ontologie O1 */
4      pour chaque (N1 ∈ O1) faire
5          /* parcours des noeuds de l'ontologie O2 */
6          pour chaque (N2 ∈ O2) faire
7              si N1.type=N2.type alors
8                  Sim_T = CalculSimTerm(N1,N2)
9                  /* Ajouter : 2 noeuds et la valeur de la similarité terminologique*/
10                 Ajouter((N1,N2,Sim_T),VS_T)
11     retourner(VS_T)
12 fin
```

Algorithme 1. PHASE1_SIMTERM

3.2. *Calcul de similarité globale*

Le calcul de la similarité globale, dite *sémantique*, se fait entre les ensembles de noeuds voisins par catégories. La fonction PHASE2_SIMSEM organise, par catégories, les noeuds adjacents au couple d'entités à apparier. Ensuite, il calcule la mesure de similarité entre chaque paires de même catégorie. Pour effectuer ce calcul, la mesure de similarité "*Match-Based similarity*" (Touzani, 2005) est utilisée :

$$MSim(E, E') = \frac{\sum_{(i,i') \in Paires(E,E')} Sim(i, i')}{Max(|E|, |E'|)} \qquad (1)$$

où E et E' représentent deux ensembles de noeuds de même catégorie. Cette fonction, requiert que les similarités locales des couples (i, i') soient déjà calculées, donne comme résultat les couples de l'ensemble $P = E \times E'$. Les couples (i, i'), intervenant dans le calcul, doivent présenter les meilleurs mesures de similarité. Pour les choisir, il existe deux approches : l'algorithme glouton (Touzani, 2005) et la programmation dynamique (Boddy, 1991). L'algorithme glouton effectue des choix locaux. En effet, lorsqu'il est confronté à un choix, il prend ce qui lui semble le meilleur pour avancer, et espère ensuite que la succession de choix locaux contribue à une solution optimale. Tandis que, la programmation dynamique essaie d'aboutir à une approche d'optimisation globale. Dans notre algorithme d'alignement, EDOLA, l'algorithme glouton est implémenté. En effet, l'algorithme glouton choisit un couple d'entités ayant la plus grande similarité et qui est supérieur ou égal au seuil fixé. Ensuite, il ôte les deux entités du couple de la table des similarités. L'algorithme continue la vérification pour chaque couple jusqu'à ce que il n'existe plus de couples ayant une mesure de similarité supérieure au seuil.

```
1  Fonction : PHASE2_SIMTERM
   Données : (1) Deux ontologies O1 et O2 (2) Vecteur de similarité terminologique
             VS_T (3) Poids de la similarité terminologique Π_L
   Résultats : Vecteur de similarité globale (sémantique) VS
2  début
3      /*calcul du diamètre minimal pour chaque couple de noeuds*/
4      pour chaque (e ∈ VS_T) faire
5          VD_i=min(e1_O1.diametre,e2_O2.diametre)
6      /*itérer jusqu'à atteindre le maximum des diamètres appartenant à VD*/
7      pour (it=1 ; it ≤ Max_{j∈[1,VD.taille]} VD_j ;it++) faire
8          /*parcourir le vecteur des similarités de l'itération précédente, le vecteur de
           similarité de la première itération est VS_T*/
9          pour (j=0 ; j < VS.taille ; j++) faire
10             /* vérifier numéro itération et diamètre minimum de noeuds à aligner*/
11             si it < VD_j alors
12                 Simvois = CalculSimVois(VS_j.N_O1, VS_j.N_O2)
13                 Sim = Π_L × VS_T(j) + Simvois
14                 VS_j=(N_O1,N_O2,Sim)
15     retourner(VS)
16 fin
```

Algorithme 2. PHASE2_SIMSEM

Afin de résoudre le problème des dépendances de similarité, la méthode du système d'équations à point fixe (Euzenat *et al.,* 2004b) est exploitée. Elle utilise une fonction quasi-linéaire qui attribue à chaque catégorie de noeuds un poids Π. Formellement, étant donné une catégorie de noeuds X et l'ensemble des relations impliqués N(X), la mesure de similarité globale $Sim_X : X \rightarrow [0, 1]$ est définie par :

$$Sim_X(x, x') = \sum_{F \in N(x)} \Pi_F^X Sim_Y(F(x), F(x')). \qquad (2)$$

La fonction est normalisée puisque $\sum(\Pi_F^X) = 1$. Dans l'approche d'alignement EDOLA, que nous proposons, les poids sont fixés par défaut pour chaque catégorie de noeuds. Ceci n'empêche pas que l'utilisateur peut assigner les poids qu'il souhaite. En utilisant l'équation (2), pour calculer la similarité globale des différentes catégories, un système d'équations linéaires est obtenu. Les variables de ce système sont les similarités des couples de noeuds déduite de l'équation (1). La résolution du système de l'équation (2), se fait par itérations. L'itération `0` de l'algorithme 2 (*c.f.* ligne 10) exploite les similarités terminologiques, déjà calculé par intermédiaire de l'algorithme 1. Ensuite, l'itération `1` de l'algorithme 2 utilise l'équation (2) pour calculer les similarités globales entre couples d'entités de même catégories. Les mesures de similarités des catégories intervenant dans le calcul de la similarité d'un couple sont issues de l'itération précédente. Ainsi, l'itération `j` fonctionne de la même manière que l'itération précédente. Le calcul de la similarité globale de chaque couple est basé sur les mesures de similarités calculées à l'itération `(j-1)`. Dans chaque itération, le nombre de candidats à aligner diminue en fonction du diamètre minimum du couple de noeud à apparier. L'exploration du diamètre de chaque noeud permet la propagation de la similarité à travers le voisinage. Le principe de cette propagation est expliqué dans ce qui suit.

3.3. *Propagation de la similarité à travers le voisinage*

La méthode, EDOLA, effectue une propagation de similarité nettement meilleure que celle de OLA. En effet, dans son processus d'alignement, tout le voisinage du couple d'entité à aligner est intégré dans le calcul de similarité. Par exemple, considérons la figure 1 qui présente deux ontologies O_1 et O_2. Étant donné le couple d'entités (Étudiant(O_1), Étudiant(O_2)), le calcul de la similarité inclut les entités voisines qui entrent en jeux. Le calcul de similarité du couple en question évoque dans cet exemple le type `objectProperty` et varie pour les deux algorithmes EDOLA et OLA. Ainsi, la table 3 présente les entités voisines du couple (Étudiant(O_1), Étudiant(O_2)) pour, respectivement, EDOLA et OLA. Ainsi, la méthode EDOLA intègre les mesures de similarité des couples d'entités (encadre(O_1), est_encadré(O_2)) et (réalise(O_1), est_réalisé(O_2)) dans le calcul de similarité du couple (Étudiant(O_1), Étudiant(O_2)), tandis que OLA, se limite à calculer la mesure de similarité entre (réalise(O_1), est_encadré(O_2)). Par conséquent, la mesure de similarité pour ce couple est mieux cernée avec EDOLA qu'avec OLA.

En outre, la méthode EDOLA, contrairement à OLA, ne se base pas sur la stabilité de la mesure de similarité, en utilisant un seuil ϵ défini par l'utilisateur. En effet, l'algorithme OLA effectue des itérations successives et dans chaque itération, les mesures de similarités des entités à aligner sont comparées à celles de l'itération précédente. Si

Entités voisines	**Étudiant(O_1)**	**Étudiant(O_2)**
Pour EDOLA	encadre, réalise	est_encadré, est_réalisé
Pour OLA	réalise	est_encadré

Tableau 3. *Table des entités voisines du couple (Étudiant(O_1), Étudiant(O_2))*

la variation est inférieure au seuil ϵ, les entités en question ne sont plus traitées dans les itérations qui suivent. Cependant, cette méthode risque de faire perdre des couples d'entités, dont les mesures de similarité peuvent augmenter la valeur de la similarité dans les itérations ultérieurs. Pour remédier à ce problème, EDOLA utilise la notion du diamètre, *i.e.*, la profondeur de l'entité dans le graphe OWL-Graph. Ainsi, la méthode EDOLA n'arrête pas d'itérer sur un couple d'entités qu'après avoir exploité toute sa structure avoisinante.

La mesure de similarité de chaque couple de noeuds varie d'une itération à une autre jusqu'elle converge. Le nombre d'itérations dans EDOLA est égal au minimum des maximums des diamètres des candidats à apparier. Dans chaque itération, l'algorithme 2, (*c.f.*, ligne 12), vérifie les candidats à aligner. Les couples de noeuds dont le diamètre minimum est inférieur au numéro de l'itération courante, ne seront pas traités. Cependant, le diamètre de chaque noeud dans le graphe doit être déterminé. Pour déterminer le diamètre d'un noeud, il faut considérer deux aspects. Le premier consiste à vérifier si le graphe est orienté ou non. Le deuxième consiste à tenir compte des relations circulaires. L'algorithme du calcul du diamètre utilise la représentation du graphe OWL-Graph de l'ontologie, et permet de déterminer les diamètres des noeuds existants. En outre, le graphe OWL-Graph considéré est un graphe non orienté. Cependant, il existe des catégories de noeuds pour lesquels un diamètre égal à zéro est donné. En effet, ces noeuds doivent être traités seulement dans l'itération 0 de l'algorithme 2 (*c.f.*, ligne 10), *i.e.*, dans l'itération de calcul de la similarité terminologique. Ces noeuds sont soit de nature type de données (string, non negative integer, etc.), ou valeur de données (une valeur numérique, une chaîne de caractères, etc.). La mesure de similarité de chaque couple de noeuds varie d'une itération à une autre pour prendre en charge les informations incorporées dans le voisinage. Le nombre d'itérations dans EDOLA est égal au maximum des minimums des diamètres des candidats à apparier. Dans la section suivante, une évaluation expérimentale de la méthode EDOLA est présentée.

4. Évaluation expérimentale de la méthode d'alignement d'ontologie EDOLA

L'évaluation expérimentale de la méthode d'alignement EDOLA a été menée sur deux aspects complémentaires. L'aspect "intra-méthode" se penchera à évaluer les performances, *i.e.*, le temps d'exécution, de la méthode *vs* la variation de la taille des ontologies à aligner et de la mesure de similarité utilisée. Le deuxième aspect,

Test	Caractéristiques de l'ontologie
101	La même ontologie de base
103	Les axiomes non reconnus sont remplacés par leur généralisation
205	Les noms des entités sont remplacés par leurs synonymes
222	La hiérarchie des classes est strictement réduite
225	Les restrictions de classes exprimées par des propriétés ont été supprimées
301	L'ontologie à comparer est réelle et semblable à l'ontologie de base
304	L'ontologie est aussi réelle et semblable à la l'ontologie de base

Tableau 4. *Ontologies de tests*

dit "inter-méthodes", permet de comparer les résultats qualitatifs obtenus par la méthode EDOLA *vs* les autres méthodes, *e.g.*, OLA. Dans le cadre des expérimentations menées, quelques tests fournis dans la base benchmark mise à la disposition de la communauté par la compétition EON (EON, 2004) sont utilisés. Ces tests sont décrit par la table 4 (EON, 2004). L'ontologie de base est constituée par un ensemble de références bibliographiques. Elle représente une version plus allégée en nombre d'entités ontologiques comparativement avec des ontologies réelles. Chaque cas de test de la base benchmark met en exergue une caractéristique de la deuxième ontologie à aligner avec la base de test. L'objectif de cette base de tests est de prendre en charge tous les aspects qui existent dans une ontologie OWL-Lite et qui pourraient avoir un impact considérable sur les métriques d'évaluation du résultat de l'alignement.

4.1. *L'aspect "intra-méthode"*

Dans ce qui suit, nous allons essayer de mesurer l'évolution des performances de la méthode EDOLA par rapport à l'augmentation de la composition structurelle de l'ontologie. Le tableau 5, présente les statistiques relevées quant à trois séries de tests qui ont été menés. En effet, la même ontologie a été utilisée, i.e., l'ontologie 101 décrite dans la table 4. Chaque test apporte un aspect incrémental de la composition structurelle de l'ontologie. Les tests effectués sont trois types de tests. Dans le TEST1, l'ontologie de référence est composée seulement de classes. Ainsi, elle est composée de 33 entités à aligner. Dans le TEST2, les 24 propriétés de nature objet sont ajoutées aux classes. Le nombre d'entités devient donc 57. Dans le TEST3, l'ontologie complète est utilisée, *i.e.*, l'ontologie est composée de 97 entités réparties comme suit : 33 classes, 24 propriétés de nature objet et 40 propriétés de nature type de données. D'après les résultats présentés dans la table 5, les performances du processus d'alignement dépend des deux aspects suivants : la taille des ontologies à aligner et le choix de la fonction

	TEST1	TEST2	TEST3
TE : Construction OWL Graph	3,450	4,700	6,780
TE : EDOLA (LEVENSHTEIN)	110,465	225,677	357,561
TE : EDOLA (WORDNET)	148,542	301,978	455,843

Tableau 5. *Temps d'exécution du construction OWL-Graph et* EDOLA *en secondes*

de similarité terminologique. En effet, le temps d'exécution[4] augmente considérablement quand le nombre d'entités à aligner accroît et inversement. Cette augmentation est plus considérable au niveau du module d'alignement qu'au niveau du module de construction du graphe OWL-Graph. Le choix de la fonction de similarité terminologique influe aussi sur le temps d'exécution du module d'alignement. En effet, l'utilisation d'une fonction simple, comme celle de LEVENSHTEIN, pour le calcul de la similarité terminologique réduit le temps d'exécution. Par contre, l'utilisation d'une fonction plus complexe comme le WORDNET augmente considérablement le temps d'exécution du processus d'alignement. Cette variation est due au temps consommé par l'algorithme d'alignement pour l'obtention de la valeur de similarité syntaxique ou lexicale. Ce temps qui est beaucoup plus important avec l'utilisation de WORDNET qu'avec une autre fonction de calcul de similarité syntaxique comme celle de LEVENSHTEIN. En effet, l'utilisation de l'API WORDNET nécessite des accès disque coûteux pour rechercher les synonymies.

4.2. *L'aspect "inter-méthodes"*

En se basant sur la qualité de l'alignement (mesure de *précision*), la méthode OLA présentait de meilleurs résultats (*c.f.*, tableau 1). De même, la méthode OLA exploite les ontologies au format OWL-Lite. Pour ces raisons, la méthode OLA servirait de méthode de référence dans la facette intra-méthode. Dans ce cadre, il importe de rappeler que la méthode EDOLA effectue une propagation de similarité sur tout le voisinage des entités. Elle exploite la notion de diamètre des entités ontologiques afin d'explorer la totalité de la structure de l'ontologie. L'alignement, produit par l'algorithme EDOLA à chaque test, est comparé à l'alignement de référence. Ainsi, les résultats des mesures de qualités sont calculées. La table 6 récapitule les résultats obtenus par les deux méthodes d'alignement EDOLA et OLA (EON, 2004). Les meilleurs résultats des valeurs de précisions de EDOLA sont obtenus lorsque les structures d'ontologies sont semblables ou identiques, *i.e.*, les tests 101, 103, 222 et 225. Ainsi, EDOLA obtient des valeurs de précision pour ces tests qui sont supérieures à 0,920. Ceci s'explique par le fait que l'approche EDOLA exploite plus efficacement les structures des

4. Les expérimentations sont réalisées sur une machine tournant sous le système d'exploitation Windows XP familial et dotée d'un processeur Pentium4 3,20 GHz, 512 Mo de RAM et 80 Go de disque.

		Précision		*Rappel*		*Fallout*	
Test	**Similarité**	**EDOLA**	**OLA**	**EDOLA**	**OLA**	**EDOLA**	**OLA**
101	LEVENSHTEIN	1,000	0,590	1,000	0,970	0,000	0,410
103	LEVENSHTEIN	0,989	0,550	0,989	0,901	0,011	0,450
205	WORDNET	0,505	0,490	0,505	0,802	0,495	0,510
222	WORDNET	0,927	0,550	0,967	0,901	0,073	0,450
225	WORDNET	0,969	0,590	0,969	0,967	0,031	0,410
301	WORDNET	0,643	0,493	0,770	0,607	0,357	0,507
304	WORDNET	0,627	0,439	0,710	0,618	0,373	0,561

Tableau 6. *Comparaison entre* EDOLA *et* OLA

entités à aligner. D'où, les entités qui ont presque la même structure sont correctement alignés. Les résultats des tests où la valeur de précision est moins bonne s'explique par deux aspects. Premièrement, l'algorithme EDOLA calcule les mesures de similarités des entités de même catégorie. Ceci induit que certains couples d'entités ne sont pas pris en considération par le processus d'alignement, d'où l'ensemble des paires appartenant à la fois à l'alignement obtenu et l'alignement de référence, $N_{Correct}$, est faible. Par conséquent, la valeur de précision est affaiblie. En outre, les couples qui ont été exclus du processus d'alignement peuvent aider à l'augmentation les mesures de similarités des couples d'entités voisines et par conséquent, augmenter le nombre de couples correctement alignés. Deuxièmement, l'algorithme EDOLA n'utilise pas dans son processus d'alignement une comparaison entre les libellés ou les commentaires des entités.

Afin d'évaluer les résultats de l'approche d'alignement proposée, la table 6 compare les résultats des deux algorithmes EDOLA et OLA. Les statistiques obtenues sont présentées dans la table 6. À partir des données présentées dans la table 6, la méthode d'alignement, EDOLA, se montre meilleure par rapport à la méthode OLA. En effet, la méthode d'alignement, EDOLA, fournit des mesures de qualités plus performantes sur presque la majorité des tests. Ces meilleurs résultats s'expliquent par les deux aspects suivants. Le premier est le fait que EDOLA effectue une propagation de similarité nettement meilleure que celle de OLA. Le deuxième aspect est que la méthode EDOLA, contrairement à OLA, ne se base pas sur la stabilité de la mesure de similarité en utilisant un seuil ϵ défini par l'utilisateur. La valeur par défaut de ce seuil est fixé à 0,01 dans OLA.

5. Conclusion

Dans ce papier, nous avons présenté une nouvelle méthode d'alignement d'ontologies OWL-Lite. La nouvelle méthode d'alignement, EDOLA, réalisée permet de rechercher les meilleurs couples à apparier en exploitant leurs graphes OWL-Graph

respectifs. Les résultats obtenus par le module d'alignement, EDOLA, sont satisfaisantes comparées aux résultats obtenus par d'autres méthodes d'alignement. En outre, la méthode proposée offre des meilleurs résultats sur la majorité des tests réalisés par rapport à la méthode OLA. Une comparaison des temps d'exécution des deux méthodes sera envisagée afin d'étudier le passage à l'échelle d'ontologies réelles et complexes.

Plusieurs améliorations sont possibles sur la méthode d'alignement, EDOLA, permettant de la rendre plus pertinente. Ces améliorations incluent : le calcul plus riche et plus complet de la similarité terminologique, le calcul de similarité inter-catégorie et l'alignement des ontologies plus complexes.

Remerciements

Le présent travail est partiellement soutenu par le projet d'action intégrée franco-tunisien PAI CMCU 05G1412 intitulé fouille de données et parallélisme. Nous remercions les relecteurs pour leurs remarques constructives.

6. 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*, Hong Kong, p. 182-197, 2006.

Bach T. L., Dien-Kuntz R., Gandon F., « On Ontology Matching Problems - for Building a Corporate Semantic Web in a Multi-Communities Organization », *Proceedings of ICEIS (4)*, Porto, Portugal, p. 236-243, 2004.

Berners-Lee T., Hendler J., Lassila O., « The semantic Web », *Scientific American*, 2001.

Boddy M., « Anytime problem solving using dynamic programming », *Proceedings of the Nineth National Conference on Artificial Intelligence*, Verlag, Anaheim, California, p. 738-743, 1991.

Charlet J., Bachimont B., Troncy R., « Ontologies pour le Web Sémantique », *Revue I3, numéro Hors Série «Web sémantique»*p. 43-63, 2004.

Connolly D., Harmelen F. V., Horrocks I., McGuinness D. L., Patel-Schneider P. F., Stein L. A., DAML+OIL : Reference Description, Technical report, W3C : Word Wide Web Consortuim, http ://www.w3.org/TR/2001/NOTE-daml+oil-reference-20011218, December, 2001.

Do H., Melnik S., Rahm E., « Comparison of schema matching evaluations », *Proceedings of the 2nd Int. Workshop on Web Databasess*, German Informatics Society, Erfurt, 2002.

Doan A. H., Madhavan J., Domingos P., Halevy A., *Handbook of Ontologies : International Handbooks on Information Systems*, Springer Verlag, Berlin, chapter 18 "Ontology Matching : a Machine Learning Approach", p. 385-404, 2004.

Ehrig M., Staab S., « QOM : Quick Ontology Mapping », *Proceedings of The 3rd ISWC, GI Jahrestagung (1)*, Hiroshima, Japon, p. 356-361, November, 2004a.

Ehrig M., Sure Y., « Ontology Mapping - An Integrated Approach », *in* , C. Bussler, , J. Davis, , D. Fensel, , R. Studer (eds), *Proceedings of the 1st ESWS*, vol. 3053, Springer Verlag, Hersounious, p. 76-91, 2004b.

EON W., « EON : Ontology Alignment Contest », *Proceedings of the 3rd Workshop Evaluation of Ontology-based Tools (EON)*, 2004.

EON W., « EON2006 : Evaluation of Ontologies for the Web », *Proceedings of the 4th International EON Workshop*, 2006.

Euzenat J., Bach T., Barrasa J., Bouquet P., Bo J. D., Dieng R., Ehrig M., R.Lara, Maynard D., Napoli A., Starmou G., Stuckenschmidt H., Shvaiko P., Tessaris S., Acker S. V., Zaihrayeu I., State of art on ontology alignment, Technical Report n° KWEB/2004/D2.2.3/v1.2, Knowledge WebConsortium, August, 2004a.

Euzenat J., Loup D., Touzani M., Valtchev P., « Ontology Alignement with OLA », *Proceedings of the 3rd International Workshop : Semantic Web Conference* EON, Hirochima, Japan, p. 341-371, November, 2004b.

Euzenat J., Mochol M., Shvaiko P., Stuckenschmidt H., Svab O., Svatek V., van Hage W. R., Yatskevich M., « Results of the Ontology Alignment Evaluation Initiative 2006 », *in* , P. Shvaiko, , J. Euzenat, , N. Noy, , H. Stuckenschmidt, , R. Benjamins, , M.Uschold (eds), *Proceedings of the 1st ESWC 2006 international workshop on ontology matching*, Athens, (GA US), 2006.

Gruber T., « A translation approach to portable ontology specifications », *Knowledge Acquisition*, vol. 5, n° 2, p. 199-220, 1993.

Kalfoglou Y., Schorlemmer M., « IF-Map : an ontology mapping method based on information flow theory », *Journal of data semantics*, vol. 1, n° 1, p. 98-127, 2003.

Klyne G., Carroll J. J., Resource Description Framework (RDF) : Concepts and Abstract Syntax, Technical report, W3C : Word Wide Web Consortuim, http ://www.w3.org/TR/2004/REC-rdf-concepts-20040210/, February, 2004.

Marsh J., XML Base, Technical report, W3C : Word Wide Web Consortuim, http ://www.w3.org/TR/2001/REC-xmlbase-20010627/, March, 2001.

Miller G. A., « WordNet : a Lexical Database for English », *Communications of the ACM*, vol. 38, n° 11, p. 39-41, November, 1995.

Noy N. F., Musen M. A., « Anchor-PROMPT : Using Non-Local Context for Semantic Matching », *Proceedings of the Workshop on Ontologies and Information Sharing at the Seventeenth International Joint Conference on Artificial Intelligence*, Seattle, WA, August, 2001.

Rahm E., Bernstein P., « A survey of approaches to automatic schema matching », *VLDB Journal*, vol. 10, n° 4, p. 334-350, 2001.

Smith M. K., Welty C., Mcguinness D. L., OWL : Ontology Web Language Guide, Technical report, W3C : Word Wide Web Consortium, http ://www.w3.org/TR/2004/REC-owl-guide-20040210/, February, 2004.

Touzani M., *« Alignement des ontologies* OWL-*Lite »*, Master's thesis, University of Montréal, 2005.

Welty C., Guarino N., « Supporting ontological analysis of taxonomic relationships », *Data and Knowledge Engineering*, vol. 1, n° 39, p. 51-74, 2001.