Utiliser des résultats d'alignement pour enrichir une ontologie

Fayçal Hamdi, Brigitte Safar, Chantal Reynaud

LRI, Université Paris-Sud 11, Bât. G, INRIA Saclay Ile-de-France (Equipe Leo)
2-4 rue Jacques Monod, F-91893 Orsay, France
{Faycal.Hamdi, safar, chantal.reynaud}@lri.fr,
http://www.lri.fr/~hamdi

Résumé. En établissant des relations entre des concepts issus de deux ontologies distinctes, les outils d'alignement peuvent être utilisés pour enrichir une des deux ontologies avec les concepts de l'autre. Cet article identifie des traitements complémentaires à l'alignement pour finaliser l'enrichissement et montre comment ils peuvent être spécifiés et mis en œuvre dans *TaxoMap Framework*. Une expérience d'enrichissement d'ontologie menée dans le cadre du projet ANR GeOnto [1] dans le domaine de la topographie est présentée.

1 Introduction

Les ontologies et les outils d'alignement d'ontologies sont des composants essentiels du Web sémantique puisqu'ils permettent l'intégration de sources dispersées dans un environnement distribué. En définissant les concepts associés à des domaines particuliers, les ontologies permettent à la fois de décrire le contenu des sources à intégrer et d'expliciter le vocabulaire utilisable dans des requêtes par des utilisateurs. Comme aucune ontologie globale ne peut couvrir l'ensemble des systèmes distribués, de multiples ontologies ont été développées indépendamment les unes des autres par des communautés différentes. Si des sources doivent être partagées, il est essentiel d'établir des correspondances sémantiques entre les ontologies qui les décrivent. Le rôle des outils d'alignement (Euzenat et Shvaiko, 2007) est ainsi de rechercher des mappings (ou appariements) entre les concepts d'ontologies distinctes, de façon à permettre la prise en compte conjointe de ressources décrites par des ontologies différentes.

Les relations établies entre les concepts issus de deux ontologies distinctes peuvent aussi être utilisées pour enrichir une des deux ontologies, dite ontologie cible, avec les concepts de l'autre, dite ontologie source. La tâche d'enrichissement d'ontologie est composée de deux phases (Ghezaiel et al., 2010) : l'identification du concept pertinent à introduire et son placement, avec les bonnes relations, au sein de l'ontologie cible.

L'objectif de cet article est d'identifier les traitements complémentaires à l'alignement utiles pour l'enrichissement et de montrer comment ils peuvent aisément être mis en œuvre au sein de *TaxoMap Framework*, un environnement permettant aux experts d'un domaine de

1. Ce travail est financé par l'Agence Nationale de la Recherche (ANR) au travers du projet ANR-07-MDCO-005 pour la création, la comparaison et l'exploitation d'ontologies géographiques hétérogènes (http://geonto.lri.fr/).

spécifier et d'effectuer des traitements sur des alignements (Hamdi et al., 2010a). Les traitements étant spécifiés de facon déclarative et générique en utilisant un langage de patterns basés sur un vocabulaire prédéfini, l'environnement est extensible et a priori applicable à n'importe quel traitement basé sur des résultats d'alignement comme le raffinement de mappings, l'enrichissement, la restructuration ou la fusion d'ontologies.

L'article est organisé comme suit. Dans la section suivante, nous présentons les résultats d'alignement, les mappings, à partir desquels s'effectue l'enrichissement. La section 3 décrit deux scénarios étudiés, les tâches qui en découlent et les patterns exprimant les traitements associés. La section 4 présente quelques travaux proches et la section 5, quelques perspectives.

2 Résultats d'alignement et exemples de traitements complémentaires

Soient deux ontologies $O_i = (C_i, H_i)$ décrites en OWL, où C_i est un ensemble de concepts caractérisés par leurs labels et H_i une hiérarchie de subsomption entre les nœuds correspondants aux concepts. Un alignement produit par *TaxoMap* est un ensemble de mises en correspondance, les mappings, établies entre chaque concept de l'ontologie source et un unique concept de l'ontologie cible. Les mappings sont exprimés par des relations d'équivalence (*isEq*), de subsomption (*isA* ou *isMoreGnl*) ou de proximité (*isClose*), auxquelles sont associées des mesures de similarité et une référence à la technique utilisée par *TaxoMap* pour les identifier (Hamdi et al., 2010b).

Suivant les contextes, tous les mappings générés n'ont pas la même fiabilité ni le même intérêt pour l'enrichissement. Ainsi, la technique t_1 établit des mappings d'équivalence entre deux concepts si leur label sont identiques. Ces mappings ne sont sûrs que si les ontologies alignées décrivent le même domaine d'application. En effet, quand les domaines sont proches et très focalisés, les risques d'ambiguïté dûs au phénomène de polysémie sont très limités, les mots ont le même sens et les concepts qui ont le même label peuvent être considérés comme équivalents pour l'alignement. Mais dès qu'une des ressources alignées est généraliste, les problèmes d'homonymie surgissent. Les éviter nécessite d'introduire une phase de vérification de la compatibilité des domaines des concepts manipulés pour ne pas mêler des sens différents.

Les mappings d'équivalence ou de proximité ne vont pas permettre d'introduire de nouveaux concepts dans la cible. Eventuellement, si le concept de l'ontologie source possède différents labels ou propriétés, ceux-ci pourront être importés dans l'ontologie cible. En revanche, les mappings faisant intervenir les relations de subsomption (*isA* ou *isMoreGnl*) sont susceptibles de conduire à de nombreux enrichissements. Ainsi, un concept c_s de la source intervenant dans un mapping de la forme <c_s isA c_t> pourrait être considéré, sous certaines conditions, comme un nouveau concept pertinent intégrable à la cible et directement placé comme une spécialisation de c_t dans celle-ci. Mais sous d'autres conditions, il devra être reconnu comme redondant et ne pas être pris en compte pour l'enrichissement.

Vérifier la compatibilité des domaines des concepts manipulés ou qualifier les concepts candidats sont autant de traitements complémentaires qui s'appuient sur les résultats d'alignement. D'autres interviennent pour faciliter le placement des concepts dans la cible, comme le regroupement des mappings liés à un même concept. L'identification et la mise en œuvre de ces traitements dépendent du contexte et des ontologies considérées. La section suivante présente deux exemples de scénarios d'enrichissement et les les traitements associés.

3 Scénarios d'enrichissement

Ces scénarios ont été identifiés dans le cadre du projet ANR $GeOnto$ dont un des buts est de construire une ontologie de concepts topographiques par enrichissement d'une première taxonomie de concepts, élaborée manuellement, à partir de spécifications de base de données (Abadie et Mustière, 2008). L'enrichissement est principalement réalisé à partir de deux sources.

La première source est constituée de plusieurs petites ontologies, chacune formée par un ensemble de concepts extraits de façon semi-automatique du thésaurus RAMEAU, ressource généraliste composée de 400 000 termes reliés entre eux par des relations qui peuvent en partie être traduites par des relations de subsomption. Ces ensembles ont été obtenus en exploitant automatiquement des récits de voyage dans lesquels sont repérées des entités nommées spatiales ("le gave de Pau") constituées de termes ("le gave") associés à des toponymes ("Pau") (Kergosien et al., 2009). L'idée est d'identifier, parmi les termes associés aux toponymes, ceux qui ont un caractère topographique et qui pourraient enrichir la taxonomie s'ils n'y sont pas déjà. Le problème est donc de distinguer, parmi les termes des entités nommées considérés comme candidats pour l'enrichissement, ceux qui sont des termes topographiques comme "gave" dans "le gave de Pau", de ceux qui n'en sont pas, comme "maire" dans "le maire de Pau".

La deuxième source est une ontologie construite automatiquement (Kamel et Aussenac-Gilles, 2009) en appliquant des techniques de traitement du langage naturel sur les spécifications ayant permis la réalisation manuelle de la taxonomie à enrichir. Le domaine de cette source est donc le même que celui de la taxonomie cible et les problèmes de validation de thématique ne se posent pas.

3.1 Validation de la thématique d'extraits de source généraliste

Pour juger de la compatibilité des domaines et écarter le risque de fausse homonymie éventuelle, chaque source issue de RAMEAU est alignée avec la cible et les mappings résultants sont analysés. Une source étant de très petite taille, elle sera jugée valide pour l'enrichissement si on peut trouver dans l'alignement au moins deux mappings tels que l'un soit un mapping d'équivalence entre un c_s et un c_t, et l'autre, un mapping jugé fiable, entre un c_{s2} différent de c_s et un c_{t2} différent de c_t. De plus, les deux concepts de la cible c_t et c_{t2} doivent être en relation (fils-père ou fils d'un même père) dans O_T. La notion de mapping *fiable* recouvre les mappings d'équivalence, d'inclusion par la technique t_2 ou de très grande proximité lexicale (t_5). Elle exclut en particulier les mappings issus d'une technique, t_7, qui produit des mappings dépendants de mappings d'équivalence déjà trouvés et qui, pour chaque concept c_s intervenant dans un mapping <c_s $isEq$ c_t>, introduit tous ses fils c_{sf} non encore alignés dans un mapping de subsomption <c_{sf} isA c_t>.

Le test de validité est exprimé dans *TaxoMap Framework* par le pattern suivant :

Partie contexte du Pattern :
$\exists x_1 \exists y_1\ \exists x_2 \exists y_2$ (***isEquivalent***$(x_1, y_1) \wedge mappingFiable(x_2, y_2)$
$\wedge$ ***conceptDifferent***$(x_1, x_2) \wedge$ ***conceptDifferent***$(y_1, y_2) \wedge$ ***isSubClassOf***(y_1, y_2, O_T))
Partie solution du Pattern :
$contexteValide$

Deux autres patterns dupliqués de celui-ci doivent être explicités, en remplaçant successivement la dernière condition du pattern, ***isSubClassOf***(y_1, y_2, O_T), par ***isParentOf***(y_1, y_2, O_T),

puis par *isSiblingOf*(y_1, y_2, O_T), pour exprimer toutes les relations structurelles acceptables entre y_1 et y_2 dans O_T. Si le contexte est validé, les différents mappings produits peuvent être pris en compte dans la suite des traitements d'enrichissements.

3.2 Enrichissement à partir d'une source focalisée et de même thématique que la cible

Le contexte de cet enrichissement est un peu particulier puisque les deux ontologies ont été construites à partir de la même spécification et sont donc vraiment très proches. La taxonomie cible manuelle comprend 600 concepts, l'ontologie construite automatiquement en comprend plus de 800. L'alignement produit 471 mappings d'équivalence, 160 autres mappings avec les techniques jugées pertinentes pour l'enrichissement, plus un certain nombre d'autres que nous ne jugeons pas assez sûrs pour être utilisés.

Parmi les 160 mappings a priori intéressants, nous avons plus particulièrement travaillé sur les mappings issus de la technique d'inclusion (t_2) qui crée un mapping <c_s isA c_{tmax}> si le label du concept de la cible c_{tmax} le plus similaire à c_s est inclus dans le label de celui-ci, et si tous les mots du label inclus sont des mots pleins (i.e. n'apparaissant pas derrière un déterminant) dans les deux labels.

En étudiant les 90 mappings produits par t_2, nous avons remarqué que dans beaucoup d'entre eux, le label du concept source c_s était constitué d'une conjonction de labels de concepts, tous déjà présents par ailleurs dans l'ontologie cible, comme par exemple le mapping <*Parc et Jardin* isA *Jardin*> où les deux concepts *Parc* et *Jardin* existent dans la cible.

Dans un contexte classique où les résultats d'alignement sont utilisés pour accéder à des documents indexés avec les concepts de la source en effectuant une interrogation à partir de ceux de la cible, ce mapping est important puisqu'il permet aussi de trouver les documents annotés uniquement par le concept <*Parc et Jardin*> quand on recherche les documents relatifs à des *Jardin*. En revanche, dans un contexte d'enrichissement, le concept <*Parc et Jardin*> bien que différent des concepts déjà présents dans la cible n'apporte rien de plus, et les mappings correspondants ne doivent pas être considérés.

L'élimination de ces mappings peut être réalisée par le pattern suivant dont la partie contexte vérifie qu'il existe un mapping de subsomption construit par la technique t_2 entre x et y, et que le label de x peut être exactement décomposé en deux composantes dont l'une est bien sûr le label de y et l'autre est exactement le label d'un autre concept z de la source.

Partie contexte du Pattern :

$\exists x \exists y\ (isAStrictInclusion(x, y) \wedge inclusionInLabel(x, "and", y)$
$\wedge\ \exists z\ (extractFromLabel(x, "and", y, L) \wedge ConceptInclus(x, L, z))$

Partie solution du Pattern :

$Delete_Mapping(x, y, _)$

4 Travaux proches

Devant les difficultés à construire une ontologie ex nihilo, des travaux de recherche portent aujourd'hui sur l'enrichissement automatique d'ontologies existantes. Une des directions de recherche concerne l'utilisation de la notion de pattern, directement issue des design patterns utilisés en génie logiciel et réintroduit dans le domaine de l'ingénierie ontologique par (Clark

et al., 2000). Ainsi, dans (Fernandez-Breis et al., 2010), des design patterns d'ontologie (ODP) sont utilisés pour enrichir l'ontologie de gènes (GO) en s'appuyant sur une analyse sémantique fine des labels des concepts et le fait que des régularités existent dans la façon de les nommer. Ces conventions de nommage permettent d'établir des mises en correspondance entre ces labels et des axiomes de l'ontologie, ce qui rend l'information sémantique explicite et permet ensuite de raisonner automatiquement dessus. Le langage OPPL (Ontology Preprocessor Language) est utilisé pour spécifier via des ODPs comment l'ontologie doit être modifiée et sous quelles conditions. Les travaux de (Scharffe et Fensel, 2008) sont très proches.

Différents travaux se sont déjà confrontés aux problèmes dûs à la polysémie, en particulier ceux proposant de valider des mappings en utilisant des ressources complémentaires comme WordNet. Cette ressource lexicale généraliste en langue anglaise regroupe les termes (noms, verbes, adjectifs et adverbes) en ensembles de synonymes appelés synsets et dénotant un concept donné. Les synsets sont reliés entre eux par des relations de généralisation/spécialisation et composant/composé. Ainsi Gracia et al. (2007) introduisent une phase d'étude des synsets associés dans WordNet aux termes qui composent un mapping. Deux termes X et Y sont considérés comme sémantiquement similaires s'il existe un synset S' de X qui soit aussi un synset de Y ou qui soit relié à un synset de Y par une suite de relations de généralisation/spécialisation dans WordNet. Un mapping <$X\ isA\ Y$> sera rejeté si S' n'existe pas ou s'il existe mais n'est pas sémantiquement similaire à l'un des ancêtres de X dans son ontologie d'origine. Une phase de désambiguïsation plus simple est effectuée dans (Mougin et al., 2006) qui ne valide que les mappings s'appuyant sur des synsets dont la définition ou les hypernymes contiennent des termes appartenant à une liste prédéfinie de mots-clés d'un domaine, en l'occurence biomédical. Dans un contexte d'enrichissement d'ontologies, Zablith et al. (2010) évaluent la pertinence de l'introduction d'une nouvelle relation < $c_s\ relation\ c_t$> en comparant le contexte du concept c_s dans son ontologie d'origine avec celui de l'ontologie cible O_T de façon à identifier les concepts partagés, i.e. de mêmes labels. Si des concepts partagés existent et qu'ils vérifient certaines relations structurelles avec les concepts intervenant dans la nouvelle relation à introduire, celle-ci est jugée pertinente pour l'enrichissement. Le système Romie (Elbyed, 2009) s'appuie sur la même idée de renforcement mutuel pour enrichir une ontologie de nouvelles relations autres que structurelles à partir de l'étude des propriétés vérifiées par les instances des concepts.

5 Conclusion

Nous avons montré dans cet article comment l'environnement *TaxoMap Framework* permet de spécifier différents traitements s'appuyant sur des résultats d'alignement et nécessaires pour l'enrichissement. Les patterns d'enrichissement présentés ont pu être construits facilement en réutilisant les primitives déjà écrites pour la tâche de raffinement et en n'introduisant que peu de nouvelles primitives, ce qui prouve l'extensibilité de l'approche.

Références

Abadie, N. et S. Mustière (2008). Constitution d'une taxonomie géographique à partir des spécifications de bases de données. In *Conf. int. de Géomatique et Analyse Spatiale, SAGEO*.

Clark, P., J. Thompson, et B. Porter (2000). Knowledge patterns. In *KR2000 : principles and Knowledge Representation and Reasoning*, pp. 591–600.

Elbyed, A. (2009). *ROMIE, une approche d'alignement d'ontologies à bases d'instances*. Thèse de doctorat, Institut National des Télécommunications, Evry.

Euzenat, J. et P. Shvaiko (2007). *Ontology Matching*. Heidelberg: Springer Verlag.

Fernandez-Breis, J., L. Iannone, I. Palmisano, A. Rector, et R. Stevens (2010). Enriching the gene ontology via the dissection of labels using the ontology pre-processor language. In *Knowledge Engineering and Management by the Masses*, Volume 6317, pp. 59–73.

Ghezaiel, L. B., C. C. Latiri, M. B. Ahmed, et N. Gouider-Khouja (2010). Enrichissement d'ontologie par une base générique minimale de règles associatives - application aux maladies neurologies : les dystonies. In *CORIA*, pp. 289–300.

Gracia, J., V. Lopez, M. D'Aquin, M. Sabou, E. Motta, et E. Mena (2007). Solving semantic ambiguity to improve semantic web based ontology matching. In *Workshop on Ontology Matching in Int. Semantic Web Conference*.

Hamdi, F., C. Reynaud, et B. Safar (2010a). Pattern-based mapping refinement. In *International Conference on Knowledge Engineering and Knowledge Management (EKAW)*.

Hamdi, F., B. Safar, N. Niraula, et C. Reynaud (2010b). Taxomap alignment and refinement modules. In *Ontology Alignment Evaluation Initiative (OAEI)*.

Kamel, M. et N. Aussenac-Gilles (2009). Ontology learning by analysing xml document structure and content. In *Knowledge Engineering and Ontology Development*, pp. 159–165.

Kergosien, E., M. Kamel, C. Sallaberry, M.-N. Bessagnet, N. Aussenac-Gilles, et M. Gaio (2009). Construction automatique d'ontologie et enrichissement à partir de ressources externes. In *Journées Francophones sur les Ontologies*, pp. 1–10.

Mougin, F., A. Burgun, et O. Bodenreider (2006). Using wordnet to improve the mapping of data elements to ulms for data sources integration. In *AMIA*, pp. 574–578.

Scharffe, F. et D. Fensel (2008). Correspondence patterns for ontology alignment. In *Proceedings of the 16th int. conf. on Knowledge Engineering: Practice and Patterns*, EKAW '08, Berlin, Heidelberg, pp. 83–92. Springer-Verlag.

Zablith, F., M. D'Aquin, M. Sabou, et E. Motta (2010). Using ontological contexts to assess the relevance of statements in ontology evolution. In *International Conference on Knowledge Engineering and Knowledge Management (EKAW)*.

Summary

With the relations generated between concepts coming from two distinct ontologies, the alignment tools can be used to enrich one of the ontology with the concepts of the other one. This paper focuses on tasks composing enrichment that completes alignment treatments. It shows their specification and how they are performed in the Taxomap Framework. An experiment in the settings of the ANR project GeOnto is presented in the topographic field.