L'approche TaxoMap Framework et son application au raffinement de mappings

F. Hamdi[1] C. Reynaud[1] B. Safar[1]

[1]LRI – Université Paris Sud 11, CNRS & INRIA Saclay Île-de-France

Parc Orsay Université – 4 rue Jacques Monod – 91893 Orsay (France)
faycal.hamdi@lri.fr

Résumé

La tâche d'alignement d'ontologies est particulièrement importante dans les systèmes d'intégration. Les outils actuels d'alignement ne sont pas performants sur tous les domaines ni quelles que soient les ontologies. La qualité de leurs résultats (mappings) pourrait souvent être améliorée si le processus d'alignement tenait davantage compte des spécificités des ontologies alignées. Nous proposons, dans cet article, un environnement, TaxoMap Framework, basé sur l'outil d'alignement TaxoMap. Cet environnement est conçu pour aider un expert du domaine à spécifier des traitements prenant appui sur des mappings produits, afin de les raffiner ou bien de fusionner, restructurer ou enrichir des ontologies. Une utilisation de cet environnement pour le raffinement de mappings est présentée dans le cadre du projet ANR GéOnto, dans le domaine de la topographie.

Mots Clef

Alignement d'ontologies, mises en correspondance, affinement de l'alignement.

Abstract

Ontology alignment is an important task for information integration. Current ontology matchers are not efficient for all application domains or ontologies. Very often the quality of the results can be improved by considering the specificities of the ontologies domain. In this paper, we propose an environment, called TaxoMap Framework, based on TaxoMap, an alignment tool, which helps an expert to specify treatments based on alignment results. The aim is to refine these results or to merge, restructure or enrich ontologies. We apply our approach to mapping refinement in the topographic field within the ANR project, GéOnto.

Keywords

Ontology alignment, mappings, alignment refinement.

1 Introduction

L'explosion du nombre de sources d'informations accessibles multiplie le besoin de techniques permettant l'intégration de ces sources. En définissant les concepts associés à des domaines particuliers, les ontologies sont un élément essentiel des systèmes d'intégration. La tâche d'alignement d'ontologies est particulièrement importante dans de tels systèmes car elle autorise la prise en compte conjointe de ressources décrites par des ontologies différentes. Les outils actuels d'alignement [4] ne sont pas performants sur tous les domaines ni quelles que soient les ontologies. Ils sont très bons dans certains cas, moins bons dans d'autres. La qualité de leurs résultats n'est pas toujours garantie et pourrait souvent être améliorée si le processus d'alignement tenait davantage compte des spécificités des ontologies alignées.

La prise en compte de ces particularités peut se faire de différentes façons : (1) lors du processus d'alignement lui-même ou (2) en raffinant les résultats générés par l'alignement, considérés comme préliminaires. Dans le premier cas, l'adaptation aux ontologies traitées peut passer par une modification des paramètres du processus d'alignement ou par la définition d'une combinaison particulière de systèmes d'alignement. Aucune différenciation n'est ainsi faite dans la façon dont sont traités les différents éléments des ontologies. A l'inverse, le raffinement de résultats d'alignement prolonge un traitement d'alignement appliqué de la même façon sur toutes les ontologies et le complète. Cette deuxième solution autorise une adaptation plus fine de l'alignement aux spécificités des ontologies traitées en permettant d'effectuer des raffinements différenciés suivant les résultats générés. Nous l'avons retenue en l'étendant de façon à pouvoir considérer, non seulement l'amélioration de la qualité d'un alignement mais également d'autres tâches telles que la fusion, la restructuration d'ontologies ou leur enrichissement. Toutes ces tâches s'appuient sur des résultats d'alignement et sont spécifiques aux caractéristiques des ontologies traitées, par exemple, à la façon dont elles sont structurées ou à la façon dont les labels de leurs concepts sont construits. Elles doivent être réalisées en interaction avec l'expert.

Il n'existe pas aujourd'hui de logiciels permettant de spécifier aisément des traitements particuliers à appliquer à un alignement, c'est pourquoi nous proposons l'environnement TaxoMap Framework, qui permet ces

spécifications en se basant sur l'outil d'alignement TaxoMap [14][7].
Nos contributions, dans ce papier, portent sur la conception de cet environnement, sur la définition d'un premier ensemble de primitives d'aide à la spécification des traitements, sur la présentation d'une utilisation de l'environnement pour le raffinement de mappings dans le domaine de la topographie.
L'article est organisé comme suit. Dans la section suivante, nous présentons le contexte de travail, en particulier l'outil d'alignement d'ontologies TaxoMap et les objectifs visés par la conception de TaxoMap Framework. La section 3 présente l'approche adoptée dans TaxoMap Framework et la section 4 décrit l'utilisation de cette approche pour le raffinement de mappings appliquée au domaine de la topographie dans le cadre du projet ANR GéOnto [5]. La section 5 présente quelques travaux proches. Enfin nous concluons et donnons quelques perspectives en section 6.

2 Cadre de travail

TaxoMap Framework s'appuie sur l'outil d'alignement d'ontologies, TaxoMap [14][7]. Nous décrivons l'outil en section 2.1 et les objectifs visés par l'approche en section 2.2.

2.1 TaxoMap

TaxoMap a été conçu pour aligner des ontologies $\mathcal{O} = (\mathcal{C},\mathcal{K})$ dans lesquelles les concepts de $\mathcal{C}$ sont seulement définis par leurs labels[1] et les relations de subsomption qu'ils entretiennent avec les autres concepts au sein de la hiérarchie de subsomption $\mathcal{K}$. Le processus d'alignement est un processus orienté qui cherche à relier chaque concept d'une ontologie source $\mathcal{O}_S = (\mathcal{C}_S,\mathcal{K}_S)$ à un unique concept d'une ontologie cible $\mathcal{O}_C = (\mathcal{C}_C,\mathcal{K}_C)$. Les relations de mise en correspondance, appelées mappings, sont soit des relations d'équivalence (*isEq*), soit des relations de subsomption (*isA*), soit des relations de proximité (*isClose*), auxquelles sont associées des mesures de similarité.
Pour identifier ces correspondances, TaxoMap met en œuvre des techniques qui exploitent toute la richesse des labels des concepts et sont donc particulièrement bien adaptées en présence de descriptions fines de domaines, ce qui se traduit au niveau des concepts par (1) des labels étendus correspondant à des expressions composées de plusieurs mots, (2) des labels de concepts généraux inclus dans des labels de concepts plus spécifiques.
Les différentes techniques s'appuient sur l'utilisation de l'analyseur morpho-syntaxique *TreeTagger* [15] et sur une mesure de similarité appliquée aux labels des concepts vus comme des ensembles de tri-grammes [12]. L'analyseur permet un paramétrage en fonction de la langue, une lemmatisation et une catégorisation des mots qui composent les labels. Chacun des mots d'un label est ainsi étiqueté par sa catégorie morpho-syntaxique (nom, adjectif, adverbe, verbe, préposition, article, pronom, conjonction), puis lemmatisé, i.e. mis sous sa forme canonique, l'infinitif pour les verbes et le masculin singulier pour les noms et adjectifs. Ainsi, pour le label « hôtel de montagne isolé situé dans le parc national », l'analyse de *TreeTagger* donne :

```
hôtel       NOM     hôtel
de          PRP     de
montagne    NOM     montagne
isolé       ADJ     isolé
situé       VER:pper    situer
dans        PRP     dans
le          DET:ART     le
parc        NOM     parc
national    ADJ     national
```

Une fois les différents mots d'un label étiquetés par *TreeTagger*, ils sont répartis en deux classes, *mot plein* ou *autre*, en fonction de leur catégorie et de leur position relative dans le label. A priori, tous les noms sont des *mots pleins* sauf s'ils sont placés derrière une préposition et tous les autres mots sont classés comme *autre*. Dans l'exemple précédent, « hôtel » est le seul mot plein, puisque les deux autres noms du label, « montagne » et « parc », sont tous les deux placés derrière une préposition. Cette répartition, entre *mot plein* ou *autre*, est ensuite utilisée pour prendre en compte l'importance relative des mots dans les labels et donner plus de poids aux *mots pleins* dans le calcul de similarité entre concepts.
Etant donné un concept c_S de l'ontologie source $\mathcal{O}_S$ que l'on cherche à aligner avec un concept de l'ontologie cible $\mathcal{O}_C$, la mesure de similarité permet d'identifier l'ensemble des concepts de $\mathcal{C}_C$ qui seront candidats au mapping avec c_S. Le choix du concept le plus pertinent parmi l'ensemble de ces candidats repose sur un ensemble de techniques variées, totalement automatiques, majoritairement terminologiques mais également structurelles [6]. Elles sont appliquées séquentiellement de façon à rendre le processus de génération de mappings le plus efficace possible. Une proposition de mapping résulte de l'application d'une technique donnée et d'une seule. Chaque concept de $\mathcal{C}_S$ ne peut être aligné qu'avec au plus un concept de $\mathcal{C}_C$. En revanche les concepts de $\mathcal{C}_C$ peuvent intervenir dans plusieurs propositions d'alignement.

2.2 Objectifs de TaxoMap FrameWork

De nombreux outils d'alignement d'ontologies ont été développés ces dernières années mais, comme le montrent les résultats des compétitions OAEI (Ontology Alignment Evaluation Initiative) [9] organisées chaque année depuis

[1] Le terme « label », correspondant à la terminologie OWL, sera utilisé dans cet article en tant que synonyme d' «étiquette ».

2004, au niveau international, dans le domaine de l'alignement d'ontologies [3][1], aucun outil n'atteint une précision et un rappel de 100 %, même si les résultats obtenus par certains de ces outils sont très bons. Ce constat concerne également TaxoMap. Nous l'avons observé au travers des résultats obtenus lors de cette compétition ces deux dernières années [7][6] mais également dans le cadre de notre participation au projet ANR GéOnto [5]. Ce projet vise la construction d'une ontologie topographique à partir de différents documents du domaine géographique, et en s'appuyant sur l'application de techniques d'alignement. Les tests effectués sur les taxonomies mises à disposition par le COGIT de l'IGN, partenaire du projet, ont montré que TaxoMap fournissait dans ce contexte de très bons résultats (précision de 92,3[2] %) mais que ces derniers pouvaient encore être améliorés.

Une étude des améliorations souhaitées par les experts a montré que celles-ci étaient souvent spécifiques aux ontologies alignées. Pour ne pas faire de TaxoMap un outil uniquement dédié à l'alignement de taxonomies topographiques et dont la qualité des résultats ne serait absolument pas garantie lors de l'alignement d'autres ontologies, nous proposons de mettre à la disposition des experts du projet, un environnement leur permettant de spécifier eux-mêmes les traitements souhaités. Cet environnement sera utilisable pour améliorer la qualité d'un alignement fourni par TaxoMap, mais également pour tout autre traitement prenant appui sur les résultats d'un alignement entre ontologies, telles des traitements de fusion, de restructuration ou d'enrichissement d'ontologies.

3 L'approche Taxomap Framework

L'approche TaxoMap Framework a été conçue pour répondre aux objectifs décrits en section 2.2. Nous décrivons l'approche et un schéma représentant l'architecture de cet environnement respectivement en section 3.1. et 3.2. Cet environnement permet la spécification de traitements à partir de primitives pré-définies. Celles-ci sont présentées en section 3.3.

3.1 Présentation de l'approche

Une caractéristique importante de l'approche est de permettre une spécification déclarative de traitements basés sur des résultats d'alignement particuliers et concernant des ontologies particulières, à l'aide d'un ensemble de primitives de base génériques et prédéfinies. Les traitements pouvant être spécifiés sont fonction des caractéristiques des ontologies concernées et de la tâche visée (raffinement de mappings, fusion d'ontologies, restructuration, enrichissement), ils sont donc associés à des modules de spécifications indépendants, un pour chaque tâche, ayant chacun leur propre ensemble de primitives de spécification. L'approche est extensible en étant a priori applicable à tout traitement prenant appui sur les résultats d'un alignement.

Cette approche doit permettre de raffiner les résultats d'alignement produits par TaxoMap. Il doit être possible, par exemple, de spécifier que le mapping « isA » généré entre « Chemin et sentier côtier » et « Sentier », conformément à la figure 1, doit être remplacé par un mapping de même type mais entre « Chemin et sentier côtier » et « Chemin ». En effet, « Sentier » est défini comme une sorte de « Chemin » dans $\mathcal{O}_C$ et le terme « chemin » est lui-même utilisé dans le label de « Chemin et sentier côtier ». L'expert préfèrerait donc établir une mise en correspondance directement entre « Chemin et sentier côtier » et « Chemin ».

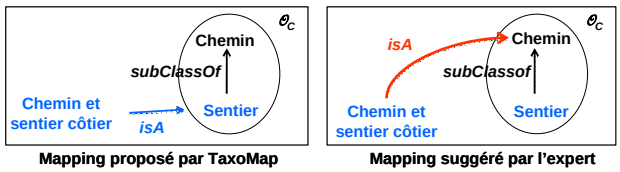


Fig. 1. Exemple de traitement à spécifier

La spécification des traitements doit pouvoir être exprimée de la façon la plus générique possible. Ainsi, celle du traitement illustré figure 1 ne devra pas faire référence directement aux concepts dénotés par « Chemin », « Sentier » et « Chemin et sentier côtier ». Pour aider l'expert à expliciter les conditions d'application des traitements qu'il souhaite mettre en œuvre, nous proposons de mettre à sa disposition un ensemble de primitives génériques prédéfinies. Ces primitives permettent de représenter les différentes conditions qui peuvent être testées sur les concepts intervenant dans un mapping construit par TaxoMap. C'est en analysant les résultats de l'alignement et en s'appuyant sur les primitives proposées que l'expert sera en mesure d'identifier des « familles » de mappings nécessitant un même raffinement puis de spécifier le traitement qu'il souhaite appliquer à chaque ensemble de cas identifié. La spécification sera ainsi déclarée de façon générique puis instanciée sur les résultats de l'alignement et les ontologies concernées pour exécuter les traitements correspondants.

L'approche doit également permettre d'autres traitements tels que la restructuration d'une ontologie $\mathcal{O}'$ construite à partir de $\mathcal{O}_S$ et de $\mathcal{O}_C$, et des alignements générés par TaxoMap entre ces deux ontologies. Ainsi, elle doit permettre d'expliciter un traitement décidant, par exemple, quels mappings « isA » doivent être transformés

[2] La précision est le ratio entre le nombre de mappings corrects trouvés et le nombre total de mappings trouvés. Le rappel est le ratio entre le nombre de mappings corrects trouvés et le nombre total de mappings corrects. Le rappel ne peut pas être calculé ici car les experts n'ont pas explicité tous les mappings corrects.

en relations « subClassOf », et accompagnant cette transformation de l'importation dans $\mathcal{O}'$ des concepts liés.

3.2. Architecture de Taxomap Framework

La figure 2 présente l'environnement de spécification mettant en oeuvre l'approche TaxoMap Framework. Cet environnement comporte trois parties : une partie « contrôleur », une partie « connaissances » et une partie « traitements ».

La partie « connaissances » regroupe l'ensemble des connaissances sur lesquelles les traitements à spécifier peuvent porter. Elle comprend ainsi les ontologies alignées par TaxoMap $\mathcal{O}_S$ et $\mathcal{O}_C$ et l'alignement généré correspondant (Base de mappings). Selon les traitements effectués, on peut y trouver également l'ontologie $\mathcal{O}_F$ issue de la fusion entre $\mathcal{O}_S$ et $\mathcal{O}_C$ réalisée en exploitant la base de mappings ou l'ontologie $\mathcal{O}'_F$ correspondant à une version restructurée ou enrichie de $\mathcal{O}_F$.

La partie « traitements » regroupe l'outil d'alignement TaxoMap et l'ensemble des modules associés aux différentes tâches à réaliser. TaxoMap enchaîne a priori 9 techniques, qui peuvent être ou non mises en œuvre lors d'une session particulière et dont l'ordre d'exécution est paramétrable. Les modules associés aux tâches permettent de spécifier des traitements particuliers qu'un expert souhaite mettre en œuvre sur des ontologies particulières, mais également d'exécuter ces traitements. Des modules supplémentaires peuvent facilement être ajoutés à condition de leur associer des primitives de spécification adaptées (pouvant être reprises de primitives proposées dans d'autres modules).

Le « contrôleur » permet de gérer l'ensemble des traitements possibles à l'aide de cet environnement, c'est-à-dire la spécification des traitements et leur exécution, l'accès aux données utiles et le stockage des résultats obtenus.

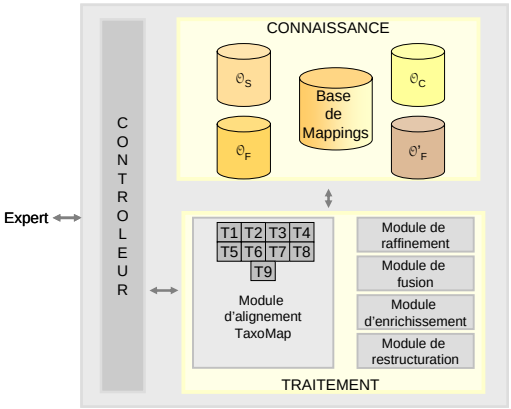


Fig. 2. Architecture de TaxoMap Framework

3.3 Primitives de Taxomap Framework

La spécification des traitements à mettre en œuvre sur les résultats de l'alignement doit être aisée. L'ensemble des éléments utiles à leur spécification est ainsi mis à disposition de l'expert sous la forme de primitives. Ces ensembles de primitives diffèrent selon la tâche visée (affinement de mappings, fusion d'ontologies, enrichissement, ..).

La spécification d'un traitement comporte deux parties : une partie « condition » devant être satisfaite pour que le traitement soit exécutable, et une partie « action » énonçant les processus à réaliser lorsque la partie « condition » est satisfaite.

La partie condition s'exprime au travers d'un ensemble de primitives, identifiées comme nécessaires pour traduire les spécifications des traitements proposés par les experts. Ces primitives vont permettre de tester (1) la technique employée pour identifier le mapping considéré, (2) des contraintes structurelles portant sur les éléments mis en correspondance, par exemple, le fait qu'ils soient liés par une relation de subsomption à des concepts vérifiant ou pas certaines propriétés, ou (3) des contraintes terminologiques, par exemple le fait que des labels de concepts soient inclus dans d'autres labels de concepts. Ces conditions sont représentées à l'aide de trois sortes de prédicats :

- les prédicats portant sur le type de techniques appliquées dans l'identification d'un mapping par TaxoMap. En testant l'existence dans la base de mappings d'une relation de correspondance particulière générée par une technique donnée, ces prédicats testent également implicitement l'ensemble des conditions d'application de cette technique. L'expert n'a donc pas besoin de les connaître précisément ni de les respécifier. Ainsi la primitive « isAInclusionStricte(X,Y) » teste l'existence d'un mapping « isA » généré entre deux concepts X et Y par la technique t_2. Elle valide en même temps implicitement les conditions d'application de t_2, c'est-à-dire le fait qu'un des labels de Y est inclus dans un des labels de X, sans apparaître derrière un déterminant, et que le concept Y est le concept de $\mathcal{O}_C$ qui a la plus forte similarité avec le concept X.

TaxoMap comportant 9 techniques, il y aura 9 prédicats de ce type. Plus formellement, étant donnés l'ensemble $\mathcal{R}_{\mathcal{M}}$ des relations de correspondance utilisées par TaxoMap, $\mathcal{R}_{\mathcal{M}}$ = {isEq, isA, isClose}, l'ensemble $\mathcal{T}$ des techniques mises en œuvre, $\mathcal{T}$ ={t_1, t_2, t_3, t_4, t_5, t_6, t_7, t_8, t_9}, la table T_M stockant les mappings générés sous la forme de 4-uplet (x, y, r, t) où x ∈ $\mathcal{C}_S$, y ∈ $\mathcal{C}_C$, r ∈ $\mathcal{R}_{\mathcal{M}}$, t ∈ $\mathcal{T}$, les couples de variables (X, Y) qui pourront instancier ces primitives prendront leurs valeurs dans l'ensemble {(x, y) | (x, y, r, t) ∈ T_M}. Les primitives utiles à la tâche de raffinement de mappings sont définies comme suit :

- isEquivalent(X,Y) est vraie ssi ∃ (X, Y,isEq, t_1) ∈ T_M

- isAInclusionStricte(X,Y) est vraie ssi $\exists$ (X,Y,isA,t_2)$\in$ T_M
- isCloseInclusionStricte(X,Y) est vraie ssi $\exists$ (X, Y, isClose, t_3) $\in$ T_M
- isCloseInclusionRelachée(X,Y) est vraie ssi $\exists$ (X, Y, isClose, t_4) $\in$ T_M
- isCloseRelatif(X,Y) est vraie ssi $\exists$(X, Y, isClose, t_5) $\in$ T_M
- isARelatif (X,Y) est vraie ssi $\exists$(X, Y, isA, t_6) $\in$ T_M
- isAapresEq(X,Y) est vraie ssi $\exists$(X, Y, isA, t_7) $\in$ T_M
- isCloseFilsCommuns(X,Y) est vraie ssi $\exists$(X, Y, isClose, t_8) $\in$ T_M
- isAPèrePartagé(X,Y) est vraie ssi $\exists$ (X, Y, isA, t_9) $\in$ T_M

Ces différentes primitives seront présentées à l'expert par le biais d'une interface, avec des commentaires explicitant leurs conditions de validation ainsi que des exemples et des contre-exemples d'utilisation (cf. Fig.3.).

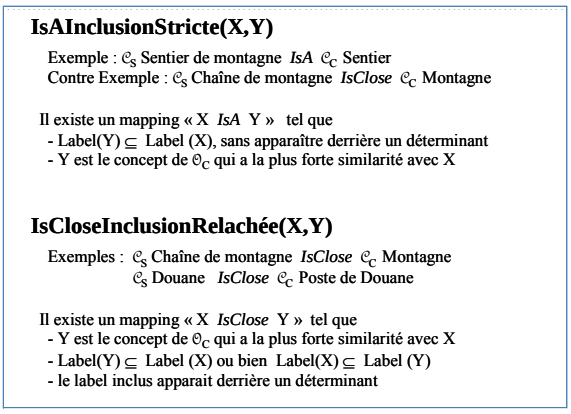


Fig. 3. Illustration de l'interface de présentation des primitives

• les prédicats exprimant des relations structurelles entre concepts X et Y d'une même ontologie $\mathcal{O} = (\mathcal{C},\mathcal{K})$, en remarquant que les instances de variables intervenant dans ces prédicats seront contraintes, soit directement parce qu'elles instancient les prédicats précédents, i.e. portant sur le type de techniques appliquées, soit indirectement par le fait de devoir être en relation avec d'autres instances.

- estSousClasseDe(X,Y,$\mathcal{O}$) est vrai $\Leftrightarrow$ subClassOf(X,Y) $\in$ $\mathcal{K}$
- estPèreDe(X,Y,$\mathcal{O}$) est vrai $\Leftrightarrow$ subClassOf(Y,X) $\in$ $\mathcal{K}$
- profondeurMax(X,$\mathcal{O}$,n) est vrai si la longueur du plus long chemin menant de X à la racine de $\mathcal{K}$ est inférieure ou égale à n.

• les prédicats exprimant des relations terminologiques entre labels de concepts :

- inclusionLabelStricte(X,Y) est défini de la façon suivante :

Pour chaque label L_1 de X
  Pour chaque label L_2 de Y
    Si $L_1 \subseteq$ *MotsPleins*(L_2,L_1) alors retourner Vrai
  FinPour
FinPour

où X et Y $\in \mathcal{C}_S \cup \mathcal{C}_C$ et *MotsPleins(L_2,L_1)* est une fonction qui calcule l'ensemble des termes de L_2 considérés comme des mots pleins dans sa comparaison avec L_1.

- conceptsDifferents(X,Y) est vrai $\Leftrightarrow$ ID(X) $\neq$ ID(Y)

avec ID(X) l'identifiant du concept X.

- inclusionDansLabel(X,Y) est vrai ssi $\exists$ un label L_1 de Y / X $\subset L_1$, où X $\in$ {« et, « ou »} et Y $\in \mathcal{C}_S \cup \mathcal{C}_C$.

Les actions décrivent les procédures à exécuter. Nous avons identifié un premier ensemble d'actions. Elles sont représentées à l'aide des trois procédures suivantes :

- ajout_Mapping(X,Y,R) qui a pour effet d'ajouter un tuple à la table T_M qui devient $T_M \cup$ {(X, Y, R, t)} où R et t sont fixées dans la partie condition du traitement, par l'instanciation du prédicat identifiant la technique considérée.
- suppression_Mapping(X,_,Y) qui a pour effet de supprimer un tuple dans la table T_M qui devient T_M - {(X, Y, _ , _)}
- ajout_Relation(X,Y,R) qui correspond à l'ajout d'une relation R reliant X et Y avec R $\in \mathcal{R}_{\mathcal{O}} \cup$ {subClassOf}.

Toutes ces primitives doivent pouvoir être sélectionnées par un expert dans TaxoMap Framework via une interface graphique appropriée. Notons que l'approche repose sur l'utilisation de TaxoMap en tant qu'outil d'alignement mais qu'elle pourrait reposer sur un autre outil à condition de définir les primitives associées à cet outil. La méthode est donc a priori reproductible. D'autres prédicats exprimant des relations structurelles ou terminologiques entre concepts seront très probablement introduits pour le traitement des tâches d'enrichissement ou de restructuration.

4 Application au raffinement de mapping

Le module de raffinement de mappings est le premier module de TaxoMap Framework, réalisé dans le cadre du projet ANR GéOnto [5]. Nous décrivons le cadre applicatif puis nous présentons les spécifications des traitements d'affinement de mappings demandés par les experts du COGIT de l'IGN, partenaires du projet.

4.1 Cadre applicatif

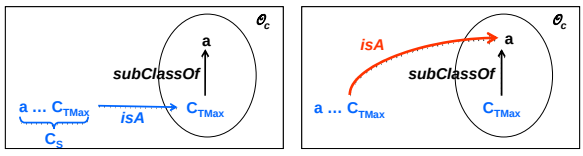
L'un des buts du projet GéOnto est de construire une ontologie de concepts topographiques, la plus complète possible, par enrichissement d'une première taxonomie de termes, Topo-Cogit, déjà réalisée par le COGIT. L'enrichissement doit être effectué, entre autres, par l'alignement de cette première ontologie avec d'autres ontologies du domaine. Ainsi, au sein du projet, d'autres

partenaires élaborent une ontologie à partir des spécifications de bases de données topographiques de l'IGN et à partir de récits de voyage de la médiathèque de Pau [8,10]. Le processus d'enrichissement doit pouvoir être automatisé, et réutilisé dans le futur sur d'autres ontologies du domaine. Ce processus s'appuyant sur les résultats d'alignement, ceux-ci doivent être les plus précis possible, pour minimiser les interventions des experts.
Les premiers tests ont été effectués en utilisant une deuxième ontologie, Carto-Cogit, construite manuellement à partir des spécifications d'une base de données de l'IGN. Dans ces tests, l'objectif est d'aligner les 495 concepts de l'ontologie source Carto-Cogit, avec l'un des 600 concepts de l'ontologie cible à enrichir, Topo-Cogit. Lors de ces tests, 326 mappings ont été identifiés par TaxoMap et présentés aux experts suivant les techniques qui avaient permis de les obtenir. 25 mappings (précision 92,33%) ont été jugés incorrects et pour certains, les experts ont énoncé des mappings alternatifs. Les traitements de raffinement de mappings proposés ci-dessous visent ainsi à obtenir ces mappings alternatifs.

4.2 Spécifications d'affinement de mappings

Nous présentons successivement les différentes demandes de modification et, pour chacune d'elles, les spécifications de traitement telles qu'elles peuvent être exprimées dans l'environnement TaxoMap Framework.

• Cas 1 : La première amélioration est celle présentée en exemple en section 3.1 (cf. Fig. 1). En toute généralité, elle concerne les mappings reliant par une relation de subsomption « isA » un concept c_S de l'ontologie source $\mathcal{O}_S$ à un concept c_{TMax} de l'ontologie cible $\mathcal{O}_C$, tels qu'un des labels de c_{TMax} est inclus dans le label de c_S. Si un des labels du concept a qui subsume c_{TMax} dans $\mathcal{O}_C$ est aussi inclus dans le label de c_S. (cf. Fig. 4), l'expert préfère rattacher c_S à a, le concept le plus général de $\mathcal{O}_C$.

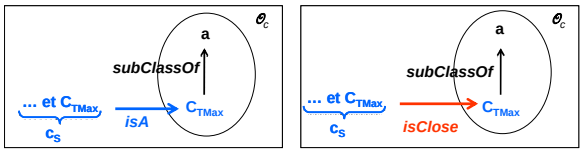


Mapping proposé par TaxoMap — Mapping suggéré par l'expert

Fig. 4. : Illustration du traitement 1 demandé par l'expert

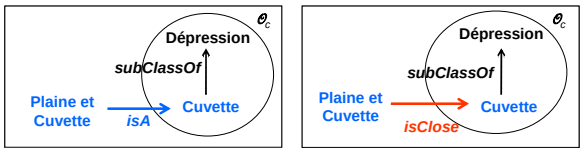
La spécification du traitement correspondant est :
Conditions d'application :
∃X ∃Y isAInclusionStricte (X,Y)
∧ ∃Z estSousClasseDe (Y,Z,$\mathcal{O}_C$)
∧ inclusionlabelStricte (Z,X))
Actions : suppression_Mapping (X,_,Y)
∧ ajout_Mapping (X,Z, isA)

• Cas 2 : Cette deuxième amélioration porte sur les mêmes types de mapping que le précédent, i.e. des mappings
c_S de l'ontologie source $\mathcal{O}_S$ à un concept c_{TMax} de l'ontologie cible $\mathcal{O}_C$, tels qu'un des labels de c_{TMax} est inclus dans le label de c_S. Si aucun des labels du concept a qui subsume c_{TMax} dans $\mathcal{O}_C$ n'est inclus dans le label de c_S mais qu'au contraire celui-ci contient l'un des connecteurs « et » ou « ou », l'expert considère que c_S n'est pas une spécialisation de c_{TMax} mais plutôt un généralisant de celui-ci, ce que nous traduisons par la relation de proximité « isClose » (cf. Fig. 5). Un mapping correspondant à ce cas est donné figure 6.



Mapping proposé par TaxoMap — Mapping suggéré par l'expert

Fig. 5. : Illustration du traitement 2 demandé par l'expert

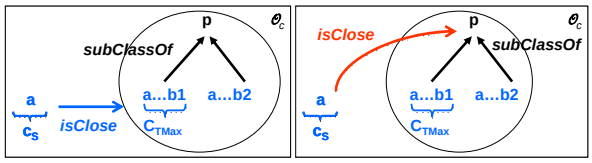


Mapping proposé par TaxoMap — Mapping suggéré par l'expert

Fig. 6. : Modification de mapping - cas 2

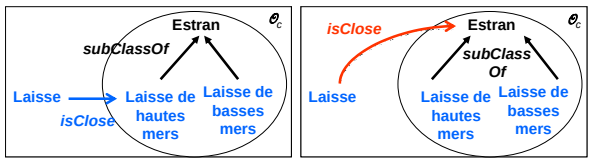
La spécification du traitement associé au cas 2 est la suivante (elle sera dupliquée pour la prise en compte du connecteur « ou ») :
Conditions d'application :
∃X ∃Y isAInclusionStricte (X,Y)
∧ inclusionDansLabel (« et », X)
∧ ∃Z estSousClasseDe (Y,Z,$\mathcal{O}_C$)
∧ ¬ inclusionLabelStricte(Z,X))
Actions : suppression_Mapping (X,_,Y)
∧ ajout_Mapping (X,Y, isClose)

• Cas 3 : Ce cas concerne les mappings reliant par une relation de proximité « isClose » un concept c_S de l'ontologie source $\mathcal{O}_S$ à un concept c_{TMax} de l'ontologie cible $\mathcal{O}_C$, tels qu'un des labels de c_S est inclus dans le label de c_{TMax}. S'il existe un autre concept de $\mathcal{O}_C$ dont un des labels contient aussi c_S et que cet autre concept a le même père p dans $\mathcal{O}_C$ que c_{TMax}, l'expert préfère rattacher c_S à ce père p (cf. Fig. 7). Un mapping correspondant à ce cas est donné figure 8.



Mapping proposé par TaxoMap — Mapping suggéré par l'expert

Fig. 7. : Illustration du traitement 3



Mapping proposé par TaxoMap — Mapping suggéré par l'expert

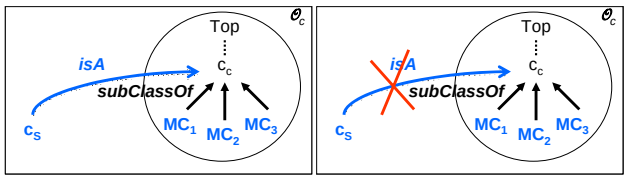
Fig. 8. : Modification de mapping - cas 3

La spécification du traitement associé au cas 3 est :
Conditions d'application :

∃X ∃Y (isCloseInclusionStricte (X,Y)
∧ ∃Z ∃P (estPèreDe(P,Y,$\mathcal{O}_C$)
∧ estPèreDe(P,Z,$\mathcal{O}_C$)
∧ conceptsDifferents(Y,P)
∧ inclusionLabelStricte(X,Z)))

Actions : suppression_Mapping (X,_,Y)
∧ ajout_Mapping (X,P, isClose)

• Cas 4 : Ce cas concerne les mappings reliant par une relation de subsomption « isA » un concept c_S de l'ontologie source $\mathcal{O}_S$ à un concept c_c de l'ontologie cible $\mathcal{O}_C$, tel que c_c est le père dans $\mathcal{O}_C$ d'au moins deux des concepts de $\mathcal{O}_C$ qui ont la similarité la plus forte avec c_S. Si ce concept c_c est trop général, i.e. positionné trop haut dans la hiérarchie $\mathcal{H}$, l'expert souhaite ne pas prendre en compte le mapping (cf. Fig. 9).



Mapping proposé par TaxoMap — Modification demandée par l'expert

Fig. 9. : Illustration du traitement 4 demandé par l'expert

La spécification du traitement correspondant est la suivante :
Conditions d'application :

∃X ∃Y (isAPèrePartagé (X,Y)
∧ profondeurMax (Y,$\mathcal{O}_C$, 2)

Action : suppression_Mapping (X,_,Y)

5 Travaux proches

Beaucoup d'outils d'alignement existant aujourd'hui génèrent de bons résultats dans certains cas et de moins bons dans d'autres. Ce constat oriente les recherches dans trois directions principales [16] : le choix de l'outil d'alignement le plus adapté, la combinaison d'outils d'alignement la plus appropriée, le problème du réglage des paramètres (seuils, coefficient de formules, etc.) utilisés au sein des outils d'alignement mis en œuvre.

Nos travaux sont issus du même constat mais ont été développés dans une direction différente, celle du raffinement de l'alignement, étendue dans un second temps à l'aide à la spécification de traitements basés sur l'alignement. Ils peuvent alors être rapprochés de ceux développés dans le cadre du système d'alignement COMA++ [2]. Ce système a pour objectif de construire des outils d'alignement puissants par combinaison d'outils existants puis de raffiner les résultats d'alignement obtenus considérés comme préliminaires. Le processus de raffinement est totalement automatique. Il consiste à réappliquer le processus d'alignement de COMA++ sur des groupes d'éléments dont la proximité a été établie par un premier traitement appliqué sur les ontologies dans leur globalité. Le raffinement de l'alignement peut aussi être vu comme une adaptation des solutions d'alignement au contexte d'une application. Ainsi le système eTuner [11] adapte un alignement en recherchant de façon totalement automatique les valeurs les plus appropriées des paramètres des systèmes d'alignement qu'il met en œuvre. Enfin, nous rapprocherons notre travail de PROMPT Suite intégrant IPROMPT, un outil de fusion d'ontologies [13], et d'autres outils de gestion d'ontologies multiples tels qu'un outil d'alignement ANCHORPROMPT, de gestion de versions, de comparaison, de traduction, au sein d'un même environnement. Ces outils sont interactifs et semi-automatiques. En matière de fusion, par exemple, le système fait des suggestions. L'expert peut retenir l'une d'elles ou spécifier une opération à exécuter. Le système exécute alors l'opération, calcule les changements qui en découlent, fait d'autres suggestions, détecte des inconsistances éventuelles.

Tous les systèmes combinant plusieurs systèmes d'alignement sont très modulaires. La possibilité de définir la stratégie de combinaison ou d'adapter automatiquement des paramètres les rend adaptables à un nouveau domaine d'application. Cette modularité et cette adaptabilité sont des points forts qui caractérisent également notre approche. Les traitements spécifiables dans TaxoMap Framework sont en effet modulaires et conçus pour intégrer les caractéristiques très particulières des ontologies traitées. Cela va même au-delà des possibilités des outils précédemment cités.

En revanche, TaxoMap Framework se distingue des outils existants (tels COMA++, eTuner ou PROMPT-Suite) en considérant que les performances d'un outil d'alignement

mettant en œuvre des algorithmes généraux d'alignement sont nécessairement limitées (même si les valeurs des paramètres sont optimales). Certaines améliorations ne peuvent être obtenues qu'après prise en compte des particularités des ontologies alignées, ce qui suppose des améliorations différentes selon les ontologies. La définition de telles améliorations nécessite de bien connaître les ontologies alignées. Ce processus ne peut donc être automatique ; seul un expert du domaine est compétent pour le faire. Comme dans PROMPT-Suite, nous proposons un environnement interactif pour aider l'expert à réaliser cette tâche, mais nous le faisons intervenir différemment. Nous lui permettons de définir des traitements particuliers génériques. Dans PROMPT-Suite, ceci n'est pas possible. Les traitements sont tous pré-définis.

6 Conclusion et perspectives

TaxoMap Framework est un environnement de spécification de traitements qui s'appuie sur les résultats d'alignement générés par TaxoMap. Nous avons présenté l'approche mise en œuvre au sein de ce système, son architecture, puis un premier ensemble de primitives pré-définies permettant à un expert du domaine de spécifier facilement les traitements qu'il souhaiterait appliquer sur un alignement. Nous avons présenté le module d'aide à l'affinement de mappings que nous avons conçu en nous appuyant sur les résultats d'expérimentations réalisées dans le cadre du projet ANR GéOnto.

La conception de TaxoMap Framework est adaptée à la spécification d'autres traitements tels que la fusion, la restructuration et l'enrichissement d'ontologies qui, comme le raffinement de mappings, exploitent un alignement. Les modules correspondant sont en cours de réalisation de même que l'interface graphique qui permettra aux experts de facilement sélectionner les primitives utiles.

Remerciements

Cette recherche est financée par l'Agence Nationale de la Recherche à travers le projet GéOnto (ANR-07-MDCO-005, http://geonto.lri.fr/).

Bibliographie

[1] C. Caraciolo, J. Euzenat, L. Hollink, R. Ichise, A. Isaac, V. Malaisé, C. Meilike, J. Pane, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, *Results of the ontology alignment initiative 2008*, in P. Shvaiko, J. Euzenat, F. Giunchiglia, H. Stuckenschmidt (Eds), Proceedings 3rd ISWC workshop on Ontology Matching, Karlsruhe (DE), pp 73-119, 2008.

[2] H.-H. Do, E. Rahm, *Matching large schemas: Approaches and Evaluation*, Information Systems 32 (2007), 857-885.

[3] J. Euzenat, A. Isaac, C. Meilike, P. Shvaiko, H. Stuckenschmidt, O. Svab, V. Svatek, W. van Hage, M. Yatskevich, *Results of the ontology alignment initiative 2007*. In Proceedings of the workshop on Ontology Matching at ISWC/ASWC, 2007.

[4] J. Euzenat, P. Shvaiko, *Ontology Matching*, Springer-Verlag, Heidelberg (DE), 2007.

[5] GéOnto : http://geonto.lri.fr/

[6] F. Hamdi, B. Safar, N. Niraula, C. Reynaud, TaxoMap in the OAEI 2009 alignment contest. – Int. Workshop on Ontology Matching, 2009.

[7] F. Hamdi, H. Zargayouna, B. Safar, C. Reynaud, *TaxoMap in the OAEI 2008 Alignment Evaluation Initiative (OAEI) 2008 Campaign* – Int. Workshop on Ontology Matching, 2008.

[8] M. Kamel, N. Aussenac-Gilles, *Ontology Learning by Analysing XML Document Structure and Content*, Knowledge Engineering and Ontology Development (KEOD), Oct. 2009, Madère, Portugal, 2009.

[9] http://www.ontologymatching.org/

[10] E. Kergosien, M. Kamel, C. Sallaberry, M.-N. Bessagnet, N. Aussenac, M. Gaio, *Construction automatique d'ontologie et enrichissement à partir de ressources externes*, 3ème Journées Francophones sur les Ontologies (JFO'2009), Poitiers, 3-4 déc. 2009.

[11] Y. Lee, M. Sayyadian, A. Doan, A.S. Rosenthal, *eTuner: tuning schema matching software using synthetic scenarios*, The VLDB Journal (2007) 16:97-122.

[12] D. Lin, *An Information-Theoretic definition of Similarity*, in proc. of the International Conference on machine Learning – ICML-98, Madison, pp. 296-304, 1998.

[13] N. F. Noy, M.A. Musen, *The PROMPT Suite: Interactive Tools For Ontology Merging And Mapping*, IJHCS, 59(6), pp. 983-1024, 2003.

[14] C. Reynaud, B. Safar, *Techniques structurelles d'alignement pour portails Web*, Revue RNTI W-3, Fouille du Web, ISBN : 978.2.85428.793.6, Cépaduès, 2007.

[15] H. Schmid, *Probabilistic Part-of-Speech TaggingUsing Decision Trees*, International Conference on New Methods in Language Processing, 1994.

[16] P. Shvaiko, J. Euzenat, *Ten Challenges for Ontology Matching*, in proc. Of the 7th International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE), 2008.