SPIM-AlignmentGUI - un logiciel d'aide à la réalisation d'alignements entre ontologies

Laurent Mazuel et Jean Charlet

INSERM UMR_S 872, Eq. 20
15, rue de l'École de Médecine, 75006 Paris
{Laurent.Mazuel,Jean.Charlet}@spim.jussieu.fr

Résumé : Dans le cadre de son travail sur les ontologies médicales et leurs alignements possibles, l'INSERM UMR_S 872, Éq. 20 a développé son propre outils d'aide à la réalisation d'alignements. Ce logiciel a pour objectif l'aide à la saisie manuelle d'alignements (basés sur l'équivalence uniquement) entre deux ontologies (format OWL, SKOS et DOE). Ce processus peut être aidé par l'utilisation d'algorithmes automatiques.

1 Introduction

À l'INSERM UMR_S 872, Éq. 20, nous avons développé un certain nombre d'ontologies pour des applications, souvent d'aide au codage médical, mais aussi pour des motivations de modélisation liées à des études d'usage (Charlet *et al.*, 2008). Dans ce contexte, les applications médicales diverses qui, pour la plupart encore une fois, indexent des données médicales liées à des patients, ne peuvent prétendre à l'avenir contribuer à l'indexation de données de santé pour des études épidémiologiques que si elles sont alignées avec les ontologies de référence médicale, tel que la SNOMED v3.5. Cet alignement est donc un passage obligé du développement de ces ontologies.[1]

Néanmoins, dans le cadre de notre travail, nous n'avons pas besoin d'un formalisme et de mécanismes très évolués tirés des paradigmes d'alignements actuels (tel que que l'on peut en trouver dans (Euzenat & Shvaiko, 2007; Stuckenschmidt *et al.*, 2004)). Ainsi, notre besoin se limite à :

- La gestion de modèles de connaissances en SKOS et en OWL (qui sont les deux formats standards que nous utilisons), mais aussi d'ontologies au format DOE[2], qui est une extension de OWL utilisant les spécialisations des « rdfs :label » pour définir les informations terminologiques.
- L'expression d'*équivalences* entre les concepts de deux ontologies (et non d'autres types de relations tels que la subsomption ou le chevauchement).[3]

[1] La problématique de recherche associée à l'utilisation d'ontologies médicales dans notre équipe est expliquée plus en détails dans (Mazuel & Charlet, 2009).

[2] The Differential Ontology Editor, http://homepages.cwi.nl/~troncy/DOE/

[3] A noter que nos modèles de connaissances sont tous issus du Web sémantique. Une *équivalence* cor-

Pour ce faire, nous avons d'abord privilégié la recherche d'un système existant pouvant fournir ces fonctionnalités. Néanmoins, nous n'avons pas été en mesure d'en trouver :

- Le plugin Prompt[4] pour Protégé n'est plus applicable sur la dernière version et ne gère pas le SKOS (4 ans depuis la dernière version).
- Le logiciel NeOn Toolkit[5] associé au plugin d'alignement de l'INRIA Grenoble[6] est proche de notre besoin, mais le SKOS n'est toujours pas géré. D'autre part des problèmes de compatibilité rendent le plugin actuellement difficilement utilisable[7].
- Le logiciel WSMT[8] du projet WSMX[9] propose un outil de recherche d'alignements. Néanmoins ce projet ne tourne actuellement que sous environnement Windows, et nous n'avons pas été capable de le démarrer, malgré la documentation d'installation. D'autre part, le format SKOS n'est pas précisé comme étant géré (2 ans depuis la dernière version).

Dans ces trois situations, le format DOE ne pouvait être géré (étant non standard).

Devant les besoins limités de notre service et les difficultés rencontrées à trouver une solution extérieure existante, le développement d'un outil spécifique simplifié pouvait être envisagé.

L'objectif de cet article est de présenter l'outil que nous avons développé à cette occasion. Cet outils permet actuellement :

- De visualiser sous forme hiérarchique avec des codes de couleurs l'état de l'alignement d'une ontologie avec une autre.
- D'éditer ces équivalences manuellement, de les sauver, de les charger et de proposer des méthodes simples pour fusionner plusieurs fichiers d'alignements obtenus de différents moyens (par exemple pour ajouter les améliorations apportées par une recherche automatique à un alignement manuel déjà commencé).
- De lancer des algorithmes automatiques (telle que la distance de Levenshtein ou la distance de Stoilos (2005)) pour proposer automatiquement certaines équivalences.
- La gestion des formalismes OWL, SKOS et DOE.

2 Aperçu des fonctionnalités

2.1 Présentation de l'interface

Un aperçu de l'interface graphique est donné sur la figure 1. L'interface se compose de deux panneaux :

- le panneau de gauche affiche l'arborescence d'une des deux ontologies ;

---

respondra donc à un couple d'URI avec un concept de chaque ontologie. Ceci nous permet sans difficultés de stocker des alignements entre des modèles différents, tel qu'un alignement entre un modèle SKOS et un modèle OWL.

[4] `http://protege.stanford.edu/plugins/prompt/prompt.html`

[5] `http://www.neon-toolkit.org/`

[6] `http://alignapi.gforge.inria.fr/`

[7] A date d'écriture de cet article, une mauvaise gestion des namespaces empêche le chargement d'ontologies ayant une définition « xml :base ».

[8] `http://sourceforge.net/projects/wsmt/`

[9] `http://www.wsmx.org/`

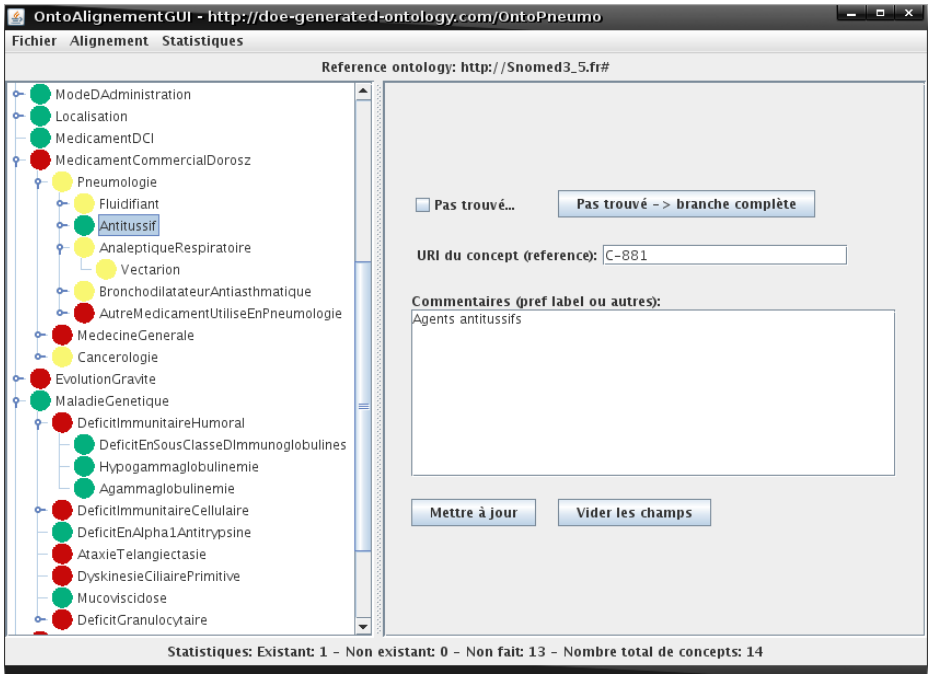

FIG. 1 – Aperçu de l'interface du logiciel.

– le panneau de droite affiche les informations associées à l'alignement d'un nœud sélectionné dans la partie gauche.

A noter que l'interface actuelle ne permet de visualiser qu'une seule ontologie, bien que les deux ontologies soit chargées en mémoire.[10] Les alignements sont donc notés par l'utilisateur à partir des nœuds de l'ontologie chargée dans le panneau gauche. Nous parlerons dans la suite « d'ontologie locale » et de « concept local » pour désigner l'ontologie et ses concepts affichés par l'interface et « d'ontologie de référence » pour désigner l'ontologie chargée en arrière-plan.

Chaque nœud de l'arbre de l'ontologie locale est associé à un code de couleur en fonction de la saisie de l'utilisateur dans la partie droite :[11]

– le vert pour signaler qu'une équivalence existe. Les références à cette équivalence sont alors notées dans la colonne de droite ;
– le rouge lorsque l'utilisateur estime qu'il n'existe pas d'équivalent dans l'ontologie de référence du concept sélectionné dans l'interface ;
– le jaune lorsque le concept n'a actuellement pas été étudié.

[10] Ceci pour des raisons historiques qui devraient disparaître dans une future version. En effet, la première version ne permettait que l'alignement entre une ontologie quelconque et la SNOMED v3.5 et l'affichage de la SNOMED était alors inutile.

[11] Afin de simplifier les explications dans les sections suivantes de cet article, nous réutiliserons le code de couleur présenté ici pour designer l'état d'alignement d'un concept de l'ontologie locale.

D'autre part, la barre d'état située en bas de l'interface permet d'avoir des informations sur la répartition de ces trois types de nœuds dans la descendance du nœud sélectionné. Cela permet de retrouver aisément dans la hiérarchie les noeuds qui n'ont pas encore été étudiés, sans avoir à « déplier » toute l'arborescence. Par exemple, sur la figure 1, la barre d'état nous apprend qu'il existe 13 concepts non étudiés dans la descendance du nœud « Antitussif ».

2.2 Gestion des fichiers d'alignement

Dans certaines situations, nous pouvons obtenir une proposition d'alignement par différents moyens (algorithme automatique, travail d'une autre équipe, etc.) et vouloir l'intégrer dans un alignement pré-existant. Nous utilisons ainsi un système simple et général (*i.e.* qui ne permet pas une gestion au cas par cas) composé de plusieurs stratégie d'agrégation.

Soit un alignement actuellement en cours dans l'interface. Au moment du chargement d'un nouveau fichier d'alignement, nous proposons différentes stratégies pour le chargement de ce nouveau fichier :

- Tout écraser : toute information du nouvel alignement sera prioritaire sur toute annotation existante (*i.e.* les concepts « rouge » ou « vert » seront toujours écrasés).
- Ne rien écraser : à l'inverse, si un concept local est « rouge » ou « vert », alors l'information en provenance du fichier sera ignorée.
- Écraser les concepts « verts » uniquement : dans ce cas, les concepts rouges de l'interface seront toujours sauvegardés, quelque soit le contenu du nouveau fichier. Conceptuellement parlant, cela signifie que si l'utilisateur a noté localement qu'il n'existait pas d'équivalences possibles pour un concept (*i.e.* annotation « rouge »), alors il faut lui faire confiance et les alignements associés proposés par le fichier en chargement ne seront pas lus.
- Écraser les concepts « rouges » uniquement : à l'inverse, ici nous ne faisons pas confiance aux annotations « rouges » de l'utilisateur, mais nous faisons confiance à ses propositions d'équivalences manuelles (annotation « vertes »).
- Écrasement sélectif : dans ce cas particulier, tous les concepts de l'interface sont écrasés, sauf un cas particulier : il est n'est pas autorisé d'écraser une équivalence manuelle (concept local « vert ») par une annotation « rouge » du fichier. Conceptuellement parlant, nous donnes toujours raison aux nouveaux alignements, sauf lorsqu'ils annotent qu'une équivalence est impossible alors que l'utilisateur en a pointé une.

2.3 Alignement automatique

Actuellement, notre interface permet de lancer 2 algorithmes d'alignements automatiques morpho-syntaxiques :

- un alignement basé sur la distance de Levenshtein ;
- un alignement basé sur la distance de Stoilos (Stoilos *et al.*, 2005).

Ces algorithmes sont lancés sur l'ensemble des termes des deux ontologies.[12] Pour chaque concept de l'ontologie locale, une recherche est faite sur l'ensemble des concepts de l'ontologie de référence. Si une équivalence dépassant un seuil acceptable est trouvée, alors le concept de l'ontologie de référence est assigné au concept de l'ontologie local sur l'interface (affichage « vert »). Dans le cas où plusieurs équivalents sont trouvés, un retour arbitraire est actuellement effectué.

3 Perspectives d'amélioration

Nous envisageons plusieurs pistes d'améliorations pour ce logiciel :

- L'utilisation d'un panneau droit pour afficher l'ontologie de référence. Afficher les deux ontologies en parallèle pourra permettre de désigner les alignements de manière graphique.
- La gestion et l'affichage des différents résultats pour un couple d'équivalence (afin d'éviter entre autres le choix arbitraire qui est actuellement fait en résultat des algorithmes automatiques). D'autre part, cette gestion des différents résultats permettrait la résolution des conflits au cas par cas lors de l'ouverture de plusieurs fichiers d'alignement entre un même couple d'ontologies.
- Utiliser la librairie de l'INRIA Grenoble comme modèle d'alignement, de chargement et de sauvegarde. La gestion de SKOS est en effet annoncée pour la version 4 de cette librairie et l'effort de codage nécessaire pour intégrer la gestion du format DOE semble acceptable.
- A terme, en fonction de l'évolution des besoins de l'équipe, proposer la gestion de relations plus complexes que l'équivalence, tel que la subsomption ou le chevauchement.

Références

CHARLET J., BANEYX A., STEICHEN O., ALECU I., DANIEL C., BOUSQUET C. & JAULENT M.-C. (2008). Utiliser et construire des ontologies en médecine : Le primat de la terminologie. *Techniques et Sciences Informatiques. À paraître*.

EUZENAT J. & SHVAIKO P. (2007). *Ontology matching*. Heidelberg (DE) : Springer-Verlag.

MAZUEL L. & CHARLET J. (2009). Alignement entre des ontologies de domaine et la Snomed : trois études de cas. In F. GANDON, Ed., *Actes des 20 ^{es} Journées Ingénierie des Connaissances*, p. 1–12, Hammamet, Tunisie. *À paraître*.

STOILOS G., STAMOU G. & KOLLIAS S. (2005). A string metric for ontology alignment. *Lecture notes in computer science*, **3729**, 624.

STUCKENSCHMIDT H., HARMELEN F. V., SERAFINI L., UQUET P. B. & GIUNCHIGLIA F. (2004). Using c-owl for the alignment and merging of medical ontologies. In *In First International Workshop on formal Biomedical Knowledge Representation. Collocated with KR 2004*.

[12] A noter que l'exécution de ces algorithmes tient compte de toutes les informations terminologiques fournies par le format SKOS et le format DOE (notamment les synonymes).