# Sémantique et Réutilisation d'ontologie générique

Sylvie Després*, Sylvie Szulman**

* LIPN, UMR7030 Univ. Paris 13
sylvie.despres@lipn.univ-paris13.fr
http://www-lipn.univ-paris13.fr/∼despres
** LIPN, UMR7030 Univ. Paris 13
ss@lipn.univ-paris13.fr
http://www-lipn.univ-paris13.fr/∼szulman

**Résumé.** Dans ce papier, nous enrichissons la méthode Terminae de construction d'ontologie à partir de textes en proposant une semi-automatisation de la construction du modèle conceptuel. Nous présentons un algorithme permettant la conceptualisation d'un terme en s'appuyant sur les informations linguistiques contenues dans l'ontologie générique de référence.

## 1 Introduction

Cet article présente une extension à la méthode de construction d'ontologie à partir de textes Terminae Aussenac-Gilles et al. (2008). Lors de la création d'une nouvelle ontologie, nous proposons de réutiliser une ontologie générique de référence afin de faciliter la phase de conceptualisation des termes d'un corpus. Une ontologie générique de référence (traduction du terme core ontology) couvre un domaine composite (par exemple le droit) comportant de nombreux sous domaines (droit public, privé, européen, etc.). A ce titre, une telle ontologie constitue un cadre unifié pour la construction d'ontologies de domaine composite puisqu'elle décrit les concepts communs à l'ensemble des sous-domaines.

La réutilisation constitue actuellement un point central de l'ingénierie des ontologies soulevant des questions complexes. De nombreux travaux sont en cours dans ce domaine Euzenat et al. (2004), Noy (2004b), Shvaiko et Euzenat (2005), Predoiu et al. (2005), Bach (2006), Safar et al. (2007). Toutefois, peu de travaux exploitent la distinction entre les différents types d'ontologies et leur articulation. En effet, parmi les concepts de l'ontologie générique, certains jouent un rôle de pivot entre les ontologies des sous-domaines et permettent d'ancrer l'ontologie en cours de construction. Le processus d'alignement proposé exploite également des informations lexicales et sémantiques de l'ontologie de référence.

Dans le paragraphe 2, nous situons notre approche de la réutilisation d'ontologies et son intégration dans la méthode Terminae. Le paragraphe 3 détaille l'algorithme d'alignement sémantique. Puis quelques exemples illustrent les premières expérimentations faites dans le domaine juridique. Enfin, nous concluons en discutant les apports et les limites de la méthode adoptée.

# 2 Méthode

## 2.1 Positionnement par rapport à alignement

Lors de la construction d'une nouvelle ontologie relative à un domaine de connaissances pour supporter une activité particulière, deux stratégies sont envisageables : la construction *ex nihilo* ou une réutilisation des ressources existantes. L'existence d'ontologies génériques disponibles, accessibles *via* les moteurs de recherche, conduit à définir des méthodes pour leur réutilisation. Les travaux relatifs à l'intégration d'ontologies cités *supra* concernent l'élaboration de méthodes d'alignement semi-automatique ou automatique. Elles conduisent à la définition de mesures de similarité lexicales ou structurelles, dites sémantiques. Si les résultats obtenus sont indéniables, ils sont relativement peu nombreux dans le domaine de la construction d'ontologies à partir de textes. Une des raisons vient sans doute de la difficulté à maîtriser le passage du niveau linguistique au conceptuel. Certaines des techniques utilisées pour l'alignement ont recours à l'exploitation structurelle des entités ou à des informations lexicales. Dans ce travail, l'alignement sémantique utilise des informations linguistiques contenues dans la définition associée à chacune des entités considérées. L'alignement est réalisé au fil de l'élaboration avec l'ontologie générique de référence et il est dirigé de l'ontologie en cours de construction vers l'ontologie générique de référence.

## 2.2 Intégration dans Terminae

Les concepts de l'ontologie en cours de construction sont définis et organisés à partir des connaissances exprimées dans les textes et en fonction des besoins de l'application. L'alignement intervient dans l'étape de conceptualisation qui se décompose en deux phases, l'amorçage et la consolidation. La phase d'amorçage consiste à identifier, dénoter et définir les concepts terminologiques (en lien avec le corpus) du domaine puis à les organiser dans des hiérarchies locales. La dénotation et la définition en langue naturelle du concept terminologique sont élaborées à partir des occurrences du terme étudié. Le repérage de propriétés structurelles et fonctionnelles liant ces concepts est obtenu à l'aide de patrons lexico-syntaxiques. Les propriétés structurelles servent de support à leur organisation hiérarchique et les propriétés fonctionnelles qui sont propres au domaine permettent d'établir des liens autres que hiérarchiques. La phase de consolidation a pour objectif de relier les hiérarchies locales obtenues et d'enrichir le modèle. Trois processus y contribuent : la généralisation, la spécialisation et le regroupement. La généralisation selon un axe ascendant permet de déterminer les nouveaux concepts ancêtres et de définir des concepts plus abstraits ou de faire référence à des catégories de plus haut niveau dans la hiérarchie. La spécialisation intervient pour chaque concept existant afin de s'assurer que ses sous-concepts ont bien été définis. Le regroupement qui permet de créer de nouveaux concepts partageant des propriétés identiques peut conduire à la définition de concepts non terminologiques.

# 3 Semi-automatisation du processus d'alignement sémantique

Les difficultés rencontrées, lors de la construction d'une ontologie à partir de textes utilisant un processus d'alignement, portent sur la mise en correspondance des concepts terminolo-

giques en cours de conceptualisation avec les entités de l'ontologie générique. Cet appariement nécessite une comparaison lexicale et sémantique de ces entités. L'aspect lexical vient accentuer les problèmes associés à la désignation des concepts terminologiques. L'aspect sémantique conduit à comparer le sens des entités de l'ontologie et des concepts terminologiques en s'appuyant à la fois sur les définitions établies à partir des occurrences des termes dans le corpus et les commentaires décrivant les entités de l'ontologie.

L'objectif poursuivi est de semi-automatiser cette mise en correspondance en exploitant la langue utilisée pour décrire les entités (dénotation de concepts, commentaires associés) de l'ontologie générique, ce qui suppose une similitude avec le vocabulaire utilisé dans le corpus. L'évaluation de la ressemblance entre les entités de deux ontologies conduit à utiliser des techniques qui permettent de comparer les concepts en mesurant la proximité lexicale Euzenat et Shvaiko (2007) des termes qui les dénotent, des propriétés structurelles et fonctionnelles qui les lient, de leur voisinage et de leurs extensions. La comparaison des propriétés peut être obtenue en comparant l'intersection et l'union des ensembles auxquelles elles appartiennent Staab et Maedche (2001), les domaine et codomaine des relations peuvent également être utilisés Cullot et al. (2003), Noy (2004a), ou encore les propriétés définissant les relations entre les concepts telles que la symétrie et la transitivité peuvent également être étudiées Ehrig et Sure (2004). Dans ce travail, la mise en correspondance repose sur une comparaison des chaînes de caractères des formes canoniques des entités (dénotation de concepts, de rôles et commentaires, les termes extraits du corpus).

## 3.1 L'algorithme d'alignement sémantique

On suppose que les deux ontologies sont exprimées dans le même langage et que l'ontologie générique est commentée en langue naturelle. On dispose de trois listes constituées à partir des entités de l'ontologie : la liste des dénotations des concepts (LCO) ; la liste des dénotations des rôles (LRO) ; la liste des termes utilisés dans les commentaires (LCom). La dénotation du concept terminologique est désignée par CT.

**Pseudo-Algorithme**

```
1  procedure ancrage(CT){
2    //  recherche de c dans LCO tel que Sim(CT,c)
3    if ( sim(CT,c)){ //Sim(CT, c) similitude lexicale de CT et c
4      if ( sem(CT,c)){//Sem(CT,c) similitude sémantique de CT et c
5        if ( feuille (c)){
6          // création ancrage
7        } else {
8          // décision externe pour ancrage de CT
9          // vers c ou un de ses descendants
10       }
11     }
12   } else {//  recherche dans les commentaires
13     //  recherche dans les composants du syntagme nominal
14   }
15 }//fin procedure
```

```
1  //recherche dans les commentaires
2  ensCom = { t tels que CT  ∈ LComm }
3    if ( ensCom !=  ∅) {
4    //  exploitation des commentaires
5      for each t in ensCom do {
6      //  lier CT et t
7        ancrage(t);
8        }
9    }
10
11 //  recherche dans les composants du syntagme nominal
12 ensMots = les syntagmes nominaux inclus dans CT
13   if( ensMots != ∅){
14     for each e in ensMots do {
15       ancrage(e);
16     }
17   }
```

Si aucun ancrage n'est trouvé, le processus ci-dessus est réitéré sur un hyperonyme jusqu'à rencontrer un terme dénotant un concept de l'ontologie générique. Au pire, l'itération se poursuit jusqu'aux concepts d'ancrage de l'ontologie de haut niveau tels que processus, objet abstrait, etc.

## 3.2 Illustration

Le domaine juridique sert de support à cette présentation et l'ontologie à construire relève du droit européen. L'ontologie générique LKIF-Core [1] notée LKIF (Legal Knowledge Interchange) exprimée en anglais est utilisée pour l'alignement. Le corpus est constitué de directives, en langue anglaise, relatives au droit des travailleurs.

Nous présentons un exemple de mise en oeuvre de ce pseudo-algorithme dans les cas suivants :
Terme : *directive*
Le concept terminologique directive correspond à un concept de l'ontologie générique qui a pour :
- commentaire associé : *Examples are European Union directive, a legislative act of the European Union and Directives, used by United States Government agencies (particularly the Department of Defense) to convey policies, responsibilities, and procedures*.
- fils les concepts : : *Proclamation* et *Legal-Document*.
Un ancrage est établi entre le concept dénoté par DIRECTIVE *via* la fiche terminologique qui permet de faire le lien avec le texte.

Terme : *employee's rights*
Il s'agit d'un terme composé qui ne dénote pas de concept de l'ontologie générique. Une recherche est entreprise sur les termes composant "employee's rights". On débute par "rights"

[1] http ://www.estrellaproject.org/lkif-core/doc/index.html

qui est la tête du syntagme nominal.
Terme : *rights*
L'étude montre que le terme lemmatisé (forme canonique) *right* dénote un concept RIGHT dans l'ontologie générique. En revanche, il ne s'agit pas d'un concept feuille, par conséquent nous explorons ses fils. Fils de RIGHT : PERMISSE, OBLIGATIVE, LIBERTY, LIABILITY. L'étude des commentaires associés à chacun des fils de RIGHT nous conduit à retenir le concept LIBERTY RIGHT auquel le commentaire suivant est associé : *When, for the benefit of a person, this person is both permitted to perform and to omit an action - that is, when the action is facultative - we can say that he or she has a liberty right with regard to that action*.
Un concept dénoté par le terme EMPLOYEES RIGHTS est créé dans l'ontologie en construction comme un fils de LIBERTY RIGHT. Un lien de subsomption est créé entre les deux concepts.

Terme : *employee*
L'étude du terme *employee* montre qu'il n'existe pas de concept correspondant dans l'ontologie générique. Néanmoins, on trouve une définition dans la directive où il apparaît : *' employee ' shall mean any person who, in the Member State concerned, is protected as an employee under national employment law*. Nous recherchons alors le terme dénotant le concept PERSON dans l'ontologie générique LKIF. Il est présent et a pour fils NATURAL_PERSON avec ce commentaire : *A natural person is a human being perceptible through the senses and subject to physical laws, as opposed to an artificial person, i.e., an organization that the law treats for some purposes as if it were a person distinct from its members or owner*.
Un concept dénoté par le terme *employee* est créé dans l'ontologie en construction comme un fils de NATURAL_PERSON. Un ancrage sous forme d'un lien de subsomption est créé entre les deux concepts.

Les termes de LRO dénotant les noms des rôles dans l'ontologie générique sont comparés aux relations lexicales extraites du corpus. Seuls quelques termes de LRO apparaissent dans le corpus comme *observe* qui intervient dans la définition de NATURAL_PERSON. L'étude des rôles fait apparaître au moins deux difficultés : au niveau conceptuel, la restriction des rôles de l'ontologie générique, au niveau de la normalisation, la mise en correspondance des relations lexicales avec les termes de LRO.

# 4 Conclusion et perspectives

Dans ce papier, nous avons proposé un algorithme d'alignement sémantique prenant en compte les informations linguistiques et sémantiques contenues dans une ontologie générique de référence (dénotation des concepts, des rôles et les commentaires associés). L'avantage de cette approche semi-automatique est d'affranchir l'ingénieur de la connaissance d'une exploration systématique de l'ontologie générique et de lui permettre de juger du sens après une comparaison lexicale des entités étudiées. Néanmoins, les premières expérimentations soulèvent des difficultés relatives à la représentation des entités contenues dans les ressources exploitées comme l'interprétation de la sémantique des commentaires. Des mesures plus précises relatives à l'amélioration apportée par cette approche sont en cours d'expérimentation.

## Références

Aussenac-Gilles, N., S. Despres, et S. Szulman (2008). *Bridging the Gap between Text and Knowledge : Selected Contributions to Ontology learning from Text*, Chapter The Terminae Method and Platform for Ontology Engineering from Texts, pp. A paraitre. IOS Press.

Bach, T.-L. (2006). *Web sémantique multi points de vue*. Ph. D. thesis, Université de Nice - Sophia Antipolis.

Cullot, N., F. Jouanot, et Yetongon (2003). Une méthode de réconciliation sémantique pour l'extraction des connaissances. In *RSTI-EGC 2003*, pp. 481–493.

Ehrig, M. et Y. Sure (2004). Ontology mapping - an integrated approach. In *Proceedings 1st ESWS Hersounisous (GR)*, Volume Volume 3053 of *Lecture Notes in Computer Science*, pp. 76–91.

Euzenat, J., T. L. Bach, J. Barrasa, P. Bouquet, J. D. Bo, R. Dieng, M. Ehrig, M. Hauswirth, M. Jarrar, R. Lara, D. Maynard, A. Napoli, G. Stamou, H. Stuckenschmidt, P. Shvaisko, S. Tessaris, S. V. Acker, et I. Zaihrayeu (2004). D2.2.3 : State of the art on ontology alignment. Technical report.

Euzenat, J. et P. Shvaiko (2007). *Ontology matching*. Springer-Verlag.

Noy, F. N. (2004a). *Tools for Mapping and Merging Ontology Handbook of ontologies*, Chapter Tools for Mapping and Merging Ontology, pp. 65–384. Springer-Verlag.

Noy, N. F. (2004b). Semantic integration : A survey of ontology-based approaches. *SIGMOD Record 33*(4).

Predoiu, L., C. Feier, F. Scharffe, J. de Bruijn, F. Martin-Recuerda, D. Manov, et M. Ehrig (2005). State of the art survey on ontology merging and aligning v2. Technical report.

Safar, B., C. Reynaud, et F. Calvier (2007). Techniques d'alignement d'ontologies basées sur la structure d'une ressource complémentaire. In *JFO2007*, Sousse Tunisie.

Shvaiko, P. et J. Euzenat (2005). A survey of schema-based matching approaches. *Journal on Data Semantics 4*, 146–171.

Staab, S. et A. Maedche (2001). Ontology learning for the semantic web. *IEEE Intelligent Systems, Special Issue on the Semantic Web 16(2)*.

## Summary

In this paper, the Terminae method which helps a knowledge engineer to build an ontology from texts, is improved. A term conceptualization algorithm is described. It uses linguistical information extracted from a reference core ontology