DAFOE : une plateforme multiméthodes et multimodèles pour construire des ontologies de domaine*

S. Szulman[1] J. Charlet[2,3] N. Aussenac-Gilles[4] A. Nazarenko[1]
Nathalie Hernandez[4] Nadia Nadah[5] Éric Sardet[6] Jean Delahousse[7]
Valery Teguiak[6]

[1] LIPN - UMR 7030, Université Paris 13 - CNRS - France
[2] INSERM UMR_S 872, Eq. 20, Paris
[3] Université Pierre et Marie Curie ; AP-HP, Paris
[4] CNRS/IRIT et Université de Toulouse
[5] Heudiasyc CNRS/UMR 6599, Université de Technologie de Compiègne
[6] LISI-ENSMA et CRITT-Informatique, Poitiers
[7] MONDECA, Paris

Sylvie.Szulman@lipn.univ-paris13.fr, Jean.Charlet@spim.jussieu.fr

Résumé

*La construction d'ontologie à partir de textes fait l'objet d'études depuis plusieurs années dans le domaine de l'ingénierie des ontologies. Un cadre méthodologique en quatre étapes (constitution d'un corpus de documents, analyse linguistique du corpus, conceptualisation, opérationnalisation de l'ontologie) est commun à la plupart des méthodes de construction d'ontologies à partir de textes. S'il existe plusieurs plateformes de traitement automatique de la langue (TAL) permettant d'analyser automatiquement les corpus et de les annoter tant du point de vue syntaxique que statistique, il n'existe actuellement aucune procédure généralement acceptée, ni a fortiori aucun ensemble cohérent d'outils supports, permettant de concevoir de façon progressive, explicite et traçable une ontologie de domaine à partir d'un ensemble de ressources informationnelles relevant de ce domaine. Le but de cet article est de présenter les propositions développées, au sein du projet ANR Da*FOE *4app, pour favoriser l'émergence d'un tel ensemble d'outils.*

Mots Clef

Ontologie, Ingénierie des connaissances, métamodélisation , TAL.

1 La plateforme DAFOE

Depuis son émergence, au début des années 1990, dans les recherches en modélisation de connaissances, la notion d'ontologie s'est rapidement diffusée dans un grand nombre de domaines de recherche en informatique. Compte tenu du caractère très prometteur de cette notion, de nombreux travaux ont visé à permettre son utilisation dans des domaines aussi divers que le traitement automatique de la langue naturelle, la recherche d'information, le commerce électronique, le web sémantique, la spécification des composants logiciels et l'intégration de système d'information.

L'efficacité de toutes ces approches présuppose néanmoins l'existence d'une ontologie de domaine susceptible d'être développée, ou d'être mise en œuvre, au sein de l'application cible. Or la conception d'une telle ontologie s'avère particulièrement difficile, surtout si l'on souhaite qu'elle fasse l'objet de consensus dans une communauté assez large. Un moyen très largement utilisé pour atteindre cet objectif est de partir d'éléments préexistants dans le domaine : corpus textuels, taxonomies, normes ou fragments d'ontologie préexistants, et de les exploiter comme base pour définir progressivement l'ontologie du domaine.

La construction d'ontologie à partir de textes fait l'objet d'études depuis plusieurs années dans le domaine de l'ingénierie des ontologies. Un cadre méthodologique en quatre étapes (constitution d'un corpus de documents, analyse linguistique du corpus, conceptualisation, opérationnalisation de l'ontologie) est commun à la plupart des méthodes de construction d'ontologies à partir de textes (TERMINAE[1] [1], Text2Onto [3]). Ces méthodes sont implémentées dans des outils qui se distinguent par leur approche de

*DaFOE4App <http ://dafoe4app.fr> est un projet ANR TLOG 010 qui vise à construire une plateforme de construction d'ontologie, DaFOE.

[1] http://www-lipn.univ-paris13.fr/~szulman/logi/index.html

la phase de conceptualisation plus ou moins automatique [4]. Cependant s'il existe des outils largement utilisés, tels que Protégé, pour représenter formellement une ontologie supposée déjà conçue, et s'il existe également plusieurs plateformes de traitement automatique de la langue (TAL) permettant d'analyser automatiquement les corpus et de les annoter tant du point de vue syntaxique que statistique, il n'existe actuellement aucune procédure généralement acceptée, ni a fortiori aucun ensemble cohérent d'outils supports, permettant de concevoir de façon progressive, explicite et traçable une ontologie de domaine à partir d'un ensemble de ressources informationnelles relevant de ce domaine. C'est ce que nous proposons dans la plateforme DaFOE pour laquelle nous espérons que de nombreux greffons viendront l'enrichir.

Un cadre méthodologique a été élaboré durant la définition de la plateforme. Il a été utilisé de deux façons, à savoir comme cadre permettant d'avoir une description commune des processus mis en jeu en même temps que modèle évoluant pour être à même de tenir compte des desiderata de tous les partenaires. Ainsi, la plateforme a différents niveaux d'entrées, correspondant aux différentes ressources, et différents niveaux de sortie correspondant à des produits de plus en plus élaborés (1) des réseaux terminologiques s'organisant durant l'analyse des données, (2) un niveau termino-conceptuel où les concepts sont organisés et (3) un niveau où l'ontologie est formalisée [2] (*cf.* fig. 1).

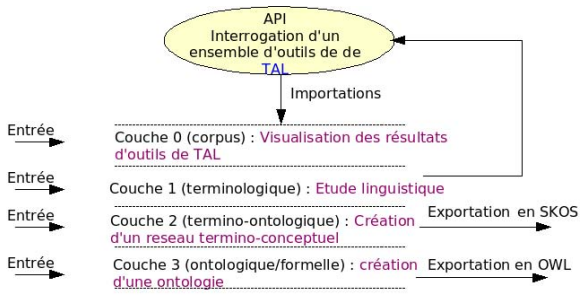

FIG. 1 – Couches du modèle de données.

Pour valider l'architecture de méta-modélisation proposée, un premier prototype est en cours de réalisation par le LISI. C'est ce prototype dont nous ferons la démonstration.

2 Démonstration

Durant la démonstration, la plateforme DaFOE sera chargée d'un corpus textuel sur le marketing sportif (plus de 20 000 mots, couche 0) et des résultats du traitement de ce corpus par un outil de TAL (couche 1).

Nous montrerons comment l'on se sert de DaFOE pour a) traiter les termes de la 1re couche en vue d'organiser un réseau de termino-concepts au niveau de la 2e couche, b) travailler sur ce résultat pour construire l'ontologie de la 3e couche.

Durant cette élaboration, nous montrerons en particulier comment les objets manipulés (termes, relations, termino-concepts, relations termino-conceptuelles, concepts, relation conceptuelles, etc.) sont liées les uns aux autres dans un vaste réseau qui permet de remonter des concepts de l'ontologie au mots que le tracent dans le corpus (*cf.* fig. 2).

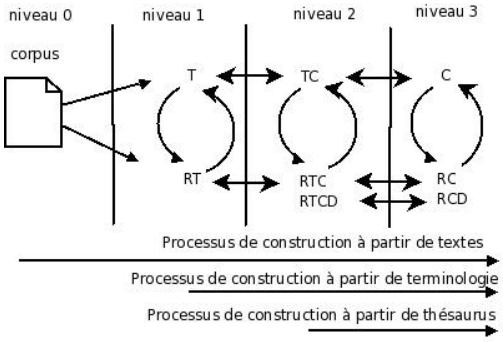

FIG. 2 – Liens entre les objets manipulés par la plateforme.

3 Remerciements

Les auteur remercient l'ensemble des membres du projet DaFOE 4app pour le travail réalisé qui a pu être synthétisé dans cet article.

Références

[1] N. Aussenac-Gilles, B. Biébow, and S. Szulman. Revisiting ontology design : a methodology based on corpus analysis. In R. Dieng and O. Corby, editors, *Knowledge Engineering and Knowledge Management : Methods, Models, and Tools. Proc. of the 12th International Conference, (EKAW'2000)*, LNAI 1937, pages 172–188. Springer-Verlag, 2000.

[2] J. Charlet, S. Szulman, G. Pierra, N. Nadah, H. V. Teguiak, N. Aussenac-Gilles, and A. Nazarenko. Dafoe : A multimodel and multimethod platform for building domain ontologies. In D. Benslimane, editor, *2e Journées Francophones sur les Ontologies*, Lyon, France, novembre 2008. ACM.

[3] P. Cimiano and J. Volker. Text2onto - a framework for ontology learning and data-driven change discovery. In A. Montoyo, R. Munoz, and E. Metais, editors, *Proceedings of the 10th International Conference on Applications of Natural Language to Information Systems (NLDB)*, volume 3513 of *Lecture Notes in Computer Science*, pages 227–238, Alicante, Spain, JUN 2005. Springer.

[4] T. Mondary, S. Despres, A. Nazarenko, and S. Szulman. Construction d'ontologies à partir de textes : la phase de conceptualisation. In Y. Prié, editor, *19es Journées Francophones d'Ingénierie des Connaissances (IC)*, pages 87–98, 18-20 Juin 2008.