

ITM, une infrastructure sémantique pour la maintenance du thésaurus multilingue Eurovoc

Thomas Francart, Charles Teissèdre

1. Introduction

Intelligent Topic Manager (ITM) est un outil propriétaire dédié à la gestion de connaissances. Dans le cas d'étude retenu pour la démonstration, l'outil sert de support logiciel pour maintenir et faire évoluer Eurovoc¹, un thésaurus multilingue (disponible en 25 langues), couvrant les différents domaines d'activité de l'Union Européenne (politique, géographie, agriculture, etc.). Ce thésaurus est utilisé par les services documentaires des administrations européennes et nationales et permet d'indexer des documents en les arrimant à un vocabulaire destiné à servir de norme. Sa constitution et sa maintenance soulèvent des difficultés, qui tiennent autant à son évolution constante, qu'à l'édition d'un vocabulaire commun à différentes langues, ce qui renvoie au problème de la traduction, dont ici le paradigme retenu (qui ne va pas de soi) est que les concepts sont partagés par tous et que les termes qui les désignent sont spécifiques à chaque langue. Les fonctionnalités d'ITM utilisées dans ce contexte peuvent être réparties en quatre grandes catégories, chacune répondant à des besoins spécifiques pour la maintenance du thésaurus dans un contexte collaboratif et partagé.

2. Un moteur de stockage "sémantique"

ITM s'appuie sur un moteur de stockage de données dites "sémantiques", qui recouvrent une superposition de modèles de représentation des connaissances. Dans le cadre de la gestion d'un thésaurus, on distingue : (1) le vocabulaire proprement dit (les concepts, les termes qui leurs sont associés, les synonymes, les relations

¹ <http://europa.eu/eurovoc/>

hiérarchiques ou transversales entre concepts, etc.) ; (2) le modèle du thésaurus, qui s'apparente ici à une transposition étendue de la norme SKOS², où les notions de termes et de concepts sont distinguées, afin d'une part de permettre de définir des relations structurelles au niveau des concepts et d'autre part de définir des relations sémantiques entre les termes attachés aux concepts (traduction, abréviation, terme préférentiel pour l'affichage, synonyme, etc.) ; (3) le modèle qui décrit les opérateurs permettant de construire formellement le modèle du thésaurus (classes, types d'attributs, cardinalités, etc.) ; (4) le modèle de réseau sémantique sur lequel repose les objets de modélisation : ce modèle propriétaire est un modèle formel proche, mais étendu, des Topics Maps³, qui manipule des topics, des associations, des rôles, des attributs et des métadonnées.

Pour la maintenance d'Eurovoc, un moteur d'inférence est utilisé pour valider des contraintes d'intégrité sur le thésaurus – contraintes paramétrables qui vérifient que les connaissances produites se conforment à différentes règles éditoriales (par exemple, que la traduction d'un terme a bien une langue cible différente de la langue source). Le moteur de stockage d'ITM fournit également une gestion de métadonnées administratives non-fonctionnelles, afin de gérer la traçabilité de chaque valeur d'attribut : date de création et de dernière modification, utilisateurs qui ont créé et modifié pour la dernière fois une valeur. Une piste d'audit complète associée à ces métadonnées permet de suivre l'utilisation de l'application ("*qui a fait quoi quand ?*"). Cette gestion implique par exemple de ne pas supprimer physiquement les informations qu'un contributeur a supprimées, mais seulement de les "marquer" comme telles, la suppression effective n'intervenant qu'à la validation par un administrateur.

3. Des fonctions dédiées à la gestion de thésaurus et d'ontologies

A un niveau plus fonctionnel, la démonstration du logiciel s'attache à illustrer les fonctionnalités dédiées à la maintenance et à l'évolution d'une base de connaissances et de sa modélisation, qui recouvre ce que la communauté du Web Sémantique désigne sous le terme d'"ontologie", soit, pour Eurovoc, le modèle du thésaurus, sa structure et les éléments

² <http://www.w3.org/TR/2009/REC-skos-reference-20090818/>

³ <http://www.topicmaps.org/xtm/index.html>

qu'elle permet d'articuler : les notions de hiérarchie arborescente ou de réseaux de concepts, les notions de termes associés, de termes équivalents, de définition, de synonymes, etc. ITM propose différentes interfaces dédiées à la maintenance d'un thésaurus : navigation hiérarchique dans des arborescences de concepts, interfaces de recherche croisant plusieurs critères, possibilité d'éditer, d'ajouter, de supprimer, de fusionner, de déprécier des termes, ou encore la possibilité d'associer plusieurs concepts pères à un concept donné (poly-hiérarchie).

De manière générale, l'ergonomie des interfaces s'efforce de rendre transparent le formalisme des connaissances manipulées en mettant en avant les fonctionnalités utiles aux utilisateurs finaux (dans le cadre d'Eurovoc, des terminologues et des traducteurs), qui n'ont ainsi pas à se soucier des modèles de représentation sous-jacents.

4. Un environnement collaboratif

ITM distingue plusieurs profils d'utilisateurs qui correspondent à différents droits. En particulier, pour Eurovoc, un profil d'utilisateur spécifique a été créé : celui de "contributeur". Un contributeur effectue des modifications dans le cadre des tâches qui lui sont affectées par des administrateurs - modifications qui doivent alors être validées par ces derniers pour intégrer la version stable du vocabulaire. L'administration des droits joue ainsi essentiellement sur différentes déclinaisons de trois paramètres : l'étendue de la visibilité sur les connaissances, les droits d'édition des connaissances visibles, et les accès aux fonctionnalités d'administration (boutons visibles ou masqués).

ITM dispose de fonctionnalités de validation des modifications des contributeurs qui s'appuient sur les capacités de traçabilité de son moteur de stockage. Cela se traduit par des écrans de revue et de validation des modifications par les administrateurs. Cela permet de conserver, à tout instant, une version du thésaurus stable, et de n'y inclure progressivement que des modifications validées. Ce workflow permet aux éditeurs d'Eurovoc de contrôler les versions successives du thésaurus publié.

5. Des capacités d'intégration et de synchronisation des données

ITM propose différents outils pour faciliter son intégration dans des systèmes d'informations plus vastes (portails intranet, chaînes de traitements de l'information), en exposant notamment une API ainsi que des web services pour manipuler ses fonctionnalités (création, interrogation, génération de rapports, etc.). Il permet une synchronisation des données avec d'autres agents logiciels notamment par des outils dédiés à l'importation et l'exportation des connaissances dans différents formats (RDF, OWL, SKOS, Excel, XML, XTM). Pour Eurovoc, un export du thésaurus en SKOS est réalisé vers un portail Web de diffusion.

Enfin, l'application permet de générer des tableaux Excel, selon des modèles paramétrables, destinés à servir de support pour la revue des comités qui assurent le suivi des évolutions du vocabulaire.

L'utilisation d'un outil basé sur la sémantisation des données, ainsi qu'un export en SKOS, vont permettre l'interconnection avec d'autres vocabulaires (Agrovoc, GEMET, etc.). La création de correspondances entre ces vocabulaires et leur publication en RDF seront un élément important dans le paysage des données sémantiques interreliées sur le web, les "linked data".

A propos des auteurs

Francart Thomas

Mondeca

Directeur technique

3, cité Nollez 75 018 Paris

thomas.francart@mondeca.com

<http://mondeca.wordpress.com>

Teissèdre Charles

Mondeca

Ingénieur de recherche

3, cité Nollez 75 018 Paris

charles.teissedre@mondeca.com