Enrichissement d'Ontologies dans le Secteur de l'Eau Douce en Environnement Internet Distribué et Multilingue

Lylia Abrouk, Mathieu Lafourcade

▶ **To cite this version:**

Lylia Abrouk, Mathieu Lafourcade. Enrichissement d'Ontologies dans le Secteur de l'Eau Douce en Environnement Internet Distribué et Multilingue. EGC: Extraction et Gestion des Connaissances, Jan 2006, Villeneuve d'Ascq, France. pp.709-710. lirmm-00102789

HAL Id: lirmm-00102789

https://hal-lirmm.ccsd.cnrs.fr/lirmm-00102789

Submitted on 2 Oct 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Enrichissement d'ontologies dans le secteur de l'eau douce en environnement Internet distribué et multilingue

Lylia Abrouk*,** Mathieu Lafourcade*

*LIRMM, 161 rue Ada, Montpellier
{abrouk,lafourcade}@lirmm.fr
**SEMIDE, 2229 route des crêtes, Valbonne
l.abrouk@semide.org

1 Introduction

Notre travail s'inscrit dans le contexte du projet européen SEMIDE (Système euro méditerranéen d'information sur les savoir-faire dans le domaine de l'eau). Le SEMIDE vise à développer une ontologie spécifique aux connaissances dans le domaine de l'eau. Ce travail s'est basé dans un premier temps sur un thésaurus du domaine de l'eau, or les ressources d'informations ne cessent de s'accroître de sources hétérogènes dans les formats, mais aussi dans le vocabulaire employé (agences de l'eau, ministères,...) engendrant une ontologie insuffisante et peu structurée. Cette ontologie doit pouvoir s'enrichir au fur et à mesure que de nouveaux documents apparaissent, mais également rester cohérente.

Nous nous intéressons à deux grandes parties : lŠannotation des ressources et l'enrichissement de l'ontologie globale définie par la communauté du SEMIDE. Ces deux grandes parties ne sont pas indépendantes étant donné que l'enrichissement de l'ontologie est fonction des nouvelles ressources et des concepts obtenus lors de l'annotation. La suite de cet article traitera la deuxième partie.

Notre hypothèse est qu'il serait intéressant de rajouter des relations ontologiques (est-un, partie-de, etc.) à l'ontologie du SEMIDE. Celle-ci prendrait donc la forme d'un pseudo-réseau sémantique ou les noeuds seraient des acceptions. Cependant, nous ne concevons la mise en place d'un tel réseau sémantique que via une automatisation poussée. La validation de certaines occurrences de relations entre acceptions pouvant être éventuellement l'objet d'un travail manuel d'un expert. Cette automatisation peut être envisagée à partir de deux types de sources : des corpus monolingues d'un même domaine technique, et des collections de bi (ou tri)-textes (textes traductions l'un de l'autres). Ce faisant, les occurrences de relations doivent d'abord être identifiées dans les parties monolingues avant d'être *migrées* dans la partie interlingue.

Nous attaquons le problème de l'enrichissement ontologique selon deux biais. La premier, via l'exploitation de paires de textes traduits, est la mise en correspondance directe de terme identifiés contre traduction mutuelle. Une acception (un sens de mot) peut être artificiellement créée, mais le problème des doublons potentiels et de l'identification et élimination n'est pas directement résolu. La seconde approche, à partir de corpus monolingue, consiste pour des termes cibles, à extraire le plus grand nombre des relations qu'ils peuvent entretenir avec d'autres mots. Les termes cibles sont identifiés comme tels via des méthodes classique de

fréquences et de cooccurrences. Les informations obtenues sont projetées sur l'ontologie monolingue, celle-ci servant également de filtre et de support quant à l'identification des acception concernées. Le processus est itératif à la fois sur les corpus et sur l'ontologie, les informations récurrentes étant progressivement recopiées dans la partie interlingue. Inversement, les relations de la partie interlingue sont progressivement recopiées vers la partie monolingue ainsi les informations extraites d'un corpus d'une langue donnée peuvent participer à l'affinement des informations dans d'autres langues.

2 Extraction de nouvelles relations - patrons d'extraction

Notre travail a consisté dans un premier temps à analyser des documents du Semide afin d'extraire des mots clés qui définiront nos règles d'extraction, cette analyse a donné une liste d'hypothèses d'extraction de relations entre les termes que nous définissons dans ce qui suit.
Hypothèse 1 : Si l'expression A est un B où A appartient à l'ontologie du Semide alors B est une spécialisation de A dans l'ontologie. Si par ailleurs, B appartient à l'ontologie globale alors B est une généralisation de A.
Hypothèse 2 : Si l'expression C qui a la forme suivante : A de B où A appartient à l'ontologie du Semide alors C est une spécialisation de A dans l'ontologie. Si, par ailleurs, C appartient à l'ontologie globale alors A est une généralisation de C.
Hypothèse 3 : Si l'expression C qui a la forme suivante : A B où A appartient à l'ontologie du Semide alors C est une spécialisation de A dans l'ontologie.
Si par ailleurs, C appartient à l'ontologie globale alors A est une généralisation de C.
Hypothèse 4 : Si on a l'expression C avec la forme suivante A non B où A appartient à l'ontologie du Semide alors C est une spécialisation de A dans l'ontologie. Et si C appartient à l'ontologie globale alors A est une généralisation de C.
Les quelques patrons d'extraction présentés ci-dessus ne sont qu'indicatifs de la méthode employée. D'autres patrons sont utilisés, en particulier pour extraire des relations d'autres natures. Par exemple, la relation de méronymie (partie de) est extraite des corpus afin de structurer l'ontologie, et de déterminer le plus finement possible les cas de doublons. Les doublons sont des termes identifiés comme des concepts synonymes et doivent être représentés comme tels dans l'ontologie.

Summary

The description of resources inside a community (or domain) must follow a controlled vocabulary. This is precisely a set of terms defined by a working group in order to tag contents and describe documents. Our problem at hand is slightly different from classical issues in controlled vocabulary as we focus ourselves on relations that may exist between concepts. Still, our resource description is based on ontology. The ontology is the backbone of a controlled and organized vocabulary and corresponds to the formalization of explicit relations created between terms of the vocabulary. Our work sticks to two main directions which are the resources annotations and the global ontology enhancement as defined by the SEMIDE community. The EMWIS (SEMIDE) is an organization viewed as a tool for exchanging information and knowledge on water between countries of the Euro-Mediterranean Partnership.