

# 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 : l'annotation 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 où 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. Le 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