Une méthode flexible de fusion de références

Fatiha Saïs*, Rallou Thomopoulos*,**

*LIRMM (CNRS & Univ. Montpellier II), 161 rue Ada, F-34392 Montpellier cedex 5
**INRA, UMR1208, 2 place P. Viala, F-34060 Montpellier cedex 1
Fatiha.Sais@lirmm.fr, rallou@supagro.inra.fr

Résumé. Dans cet article, nous nous intéressons au problème de fusion de références qui se pose une fois que les réconciliations entre références sont calculées. Il s'agit d'une tâche ayant comme objectif de fusionner les descriptions de références qui réfèrent à la même entité du monde réel pour en obtenir une seule représentation. Afin de pallier le problème d'incertitude dans les valeurs associées aux attributs, nous avons choisi de représenter les résultats de la fusion de références dans un formalisme fondé sur les ensembles flous. Nous indiquons comment les degrés de confiance sont calculés. Nous distinguons trois modes possibles de fusion. Enfin nous en proposons une représentation en RDF-Flou ainsi que son interrogation.

1 Introduction

La réconciliation et la fusion de données sont des problèmes majeurs pour l'intégration de données provenant de plusieurs sources. Ces problèmes sont liés à l'hétérogénéité syntaxique et sémantique du contenu des sources de données. La réconciliation consiste à décider si deux descriptions provenant de sources distinctes réfèrent ou non à la même entité du monde réel (e.g. la même personne, le même article, le même gène). La fusion consiste alors, à partir des descriptions réconciliées, à obtenir une seule représentation.

L'hétérogénéité des schémas est une des causes premières de la disparité de description des données entre sources (voir synthèse Rahm et Bernstein (2001); Shvaiko et Euzenat (2005)). Une autre cause d'hétérogénéité est due aux variations entre les descriptions des instances elles-mêmes. En effet, lors de l'intégration de données provenant de différentes sources, des représentations différentes peuvent référer la même entité du monde réel car des vocabulaires et des référentiels différents sont utilisés pour décrire les données. C'est dans ce cadre d'hétérogénéité liée aux données que nous nous situons dans cet article.

Comme les données sont créées de manière autonome et proviennent de différentes sources, nous ne pouvons pas faire l'hypothèse de l'identifiant unique. C'est pour cette raison que nous utilisons le terme *référence* d'une donnée au lieu d'*identifiant*. Nous parlerons alors des problèmes de réconciliation de références et de fusion de références.

Trouver les réconciliations entre références est néanmoins insuffisant pour obtenir une interrogation uniforme offrant à l'utilisateur des réponses non redondantes. Il est nécessaire d'avoir une méthode de fusion de descriptions des références réconciliées. C'est à ce problème que nous nous intéressons dans cet article. Les principales difficultés sont liées aux conflits