

Cohérence des données de bases RDF en évolution constante

Pierre Maillot*,** Thomas Raimbault* David Genest**

*De Vinci Technology Lab, ESILV, 92 916 Paris La Défense Cedex
{pierre.maillot, thomas.raimbault}@devinci.fr

**LERIA, Université d'Angers, 2 Boulevard Lavoisier, 49 045 Angers
genest@info.univ-angers.fr

Résumé. Le maintien de la qualité et de la fiabilité de bases de connaissances RDF du Web Sémantique est un problème courant. De nombreuses propositions pour l'intégration de « bonnes » données ont été faites, se basant soit sur les ontologies de ces bases, soit sur des méta-données additionnelles. Dans cet article, nous proposons une approche originale, basée exclusivement sur l'étude des données de la base. Le principe est de déterminer si les modifications apportées par la mise à jour candidate rendent la partie ciblée de la base plus similaire – selon certains critères – à d'autres parties existantes dans la base. La mise à jour est considérée cohérente avec cette base et peut être appliquée.

1 Introduction

Le *Web Sémantique* a été lancé en 2001 par le W3C¹ pour promouvoir le partage et la création de données structurées sur le Web en proposant des recommandations pour la description de données (RDF), d'ontologies (RDFS, OWL), et des méthodes et outils associés (SPARQL, ...) pour gérer les connaissances. Actuellement le Web Sémantique correspond à des centaines de bases RDF communautaires (e.g. DBPedia² ou Yago³). Ces bases contiennent des données de sources variées, résultant soit de contributions d'experts dans des domaines spécifiques, soit du moissonnage de pages Web ou de fichiers, soit enfin et surtout de contributions issues de la production participative. Les méthodes de moissonnage et de production participative sont hélas sources d'incohérences résultantes d'erreurs humaines directes ou indirectes.

Pour contrer l'apparition de ces incohérences, de nombreuses méthodes ont été proposées pour maintenir la qualité et la fiabilité des données au sein des bases RDF. Par exemple, Mendes et al. (2012) utilise des méta-données telles que la provenance ou l'historique des éditions précédentes, tandis que Jacobi et al. (2011) utilise les ontologies pour évaluer une valeur de confiance. Bonatti et al. (2011) combine les deux approches. Enfin, un état de l'art de l'évaluation de la qualité de des données RDF est fait dans Zaveri et al. (2013). On constatera néanmoins que les ontologies souffrent de leurs difficultés à s'adapter aux évolutions dans la description des données et que les approches d'intégration basées sur des méta-données requièrent l'utilisation de méthodes ad-hoc pour ces méta-données additionnelles.

1. <http://www.w3.org/standards/semanticweb/>

2. <http://dbpedia.org>

3. <http://www.mpi-inf.mpg.de/yago-naga/yago/>