WebLab-PROV : la gestion de la provenance dans la plateforme WebLab

Clément Caron*,**, Bernd Amann*
Camelia Constantin*, Patrick Giroux**

*LIP6
prenom.nom@lip6.fr

**EADS-Cassidian
prenom.nom@cassidian.com

Résumé. Dans une démarche de gestion de la qualité au sein de la plateforme de media-mining WebLab, nous présentons un modèle de provenance permettant aux fournisseurs de services de définir les dépendances entre les données en sortie et les données en entrée. Ce modèle utilise une version étendue du standard *XPath*, et *XQuery* afin de parcourir les noeuds XML des entrées et sorties des services.

Nous présentons également l'implémentation de ce modèle de provenance au sein de la plateforme WebLab, montrant tous les formats de stockage des informations, ainsi que le fonctionnement pas à pas de notre outil dans un cas d'utilisation typique.

1 Introduction

Avec le développement d'internet, la quantité de données non-structurées a augmenté de manière exponentielle. L'évolution des informations vers l'ère numérique vient avec le besoin de collecter ces données et de les structurer sémantiquement. Le domaine du media-mining (fouille de média) répond à ce besoin avec de nombreux outils, allant des collecteurs de sites Web, aux extracteurs d'entités nommées. À cause de la richesse de l'offre des fonctionnalités et des outils qui les réalisent, il n'est pas toujours aisé pour un utilisateur de déterminer quel composant choisir, à quel moment l'utiliser et s'il est le plus adapté à son cas d'utilisation. De plus, les utilisateurs ont besoin d'une solution pour composer simplement plusieurs composants, afin de créer leur propre application de media-mining.

C'est dans ce but que la plateforme Open-Source WebLab fut créée. Cette plateforme propose de combiner différents outils (ou services) de media-mining, afin de récolter des informations à partir de données non-structurées (fichiers PDF, sites Web, Vidéos, etc.). Différents standards sont utilisés au sein du projet, mais les données sont principalement stockées en XML et RDF, et requêtées à l'aide de SPARQL et XQuery. La plateforme possède une architecture distribuée, et se compose de différents services Web, suivant le type de media à traiter, et le contexte d'utilisation (politique, sportif, militaire).

Dans une démarche de gestion de la qualité, ainsi que pour faciliter la composition des services de media-mining, la problématique de la provenance des données nous est apparue.