

Modèle d'acquisition de données du LOD et génération automatique de requêtes

Céline Alec*, Chantal Reynaud-Delaître*, Brigitte Safar*

*LRI, Univ. Paris-Sud, CNRS, Université Paris-Saclay, Orsay, F-91405
prenom.nom@lri.fr

Résumé. Le LOD (Linked Open Data) représente aujourd'hui une source prometteuse pour de très nombreuses applications du Web Sémantique à condition toutefois de développer des techniques d'acquisition adaptées. Le travail présenté dans ce papier traite de ce problème. L'objectif est de peupler une ontologie OWL avec des assertions de propriétés en tenant compte de l'existence, dans le LOD, de propriétés multiples, équivalentes, multi-valuées, de propriétés absentes mais pouvant toutefois être calculées, de propriétés sans valeurs. Les correspondances à établir étant complexes, nous proposons un modèle pour les spécifier. Nous définissons également un modèle pour spécifier des chemins d'accès alternatifs à des propriétés en cas de valeurs manquantes. Enfin, les valeurs de propriétés auxquelles le LOD donne accès après application des modèles de correspondance et d'accès peuvent nécessiter des traitements complémentaires de façon à satisfaire les contraintes de l'ontologie à peupler. Nous proposons des mécanismes adaptés. Puis, nous montrons comment les différents modèles proposés sont utilisés pour générer automatiquement des requêtes SPARQL et ainsi faciliter l'interrogation. Cette génération automatique est déroulée sur deux exemples.

1 Introduction

Un des faits marquants de l'évolution de l'ingénierie des connaissances est d'avoir diversifié les sources de connaissances utilisées dans les systèmes de traitement d'information intelligents, permettant ainsi de tirer profit de leur complémentarité (Aussenac-Gilles et al., 2014). Le web des données liées est aujourd'hui une source de connaissances utilisable, tout comme les documents numériques exploités en complément d'experts spécialistes du domaine, les données dont sont extraites des connaissances, les modèles, ressources, thesaurus, représentations structurées comme les ontologies. Les données liées réfèrent à un style de publication et d'interconnexion de données structurées sur le web, basé sur le modèle RDF. Elles ont l'avantage de fournir un mécanisme d'accès unique et normalisé au lieu de s'appuyer sur différents formats d'interface et de résultat. Les sources de données peuvent être ainsi facilement explorées par les moteurs de recherche. Par définition, elles ont des liens avec d'autres sources de données. Le nombre de données publiées selon ce principe croît rapidement (on parle aujourd'hui de dizaines de milliards de triplets publiés sur Internet). Cette masse de données représente