Vers l'intégration multidimensionnelle d'Open Data dans les entrepôts de données

Alain Berro *, Imen Megdiche *, Olivier Teste *

*IRIT-Université de Toulouse
118 Route de Narbonne, 31062 Toulouse
{Alain.Berro, Imen.Megdiche, Olivier.Teste}@irit.fr

Résumé. L'émergence de nombreuses sources d'Open Data poussent plusieurs communautés de recherche ainsi que des entreprises à développer des outils permettant leur exploitation. En particulier, les données statistiques présentes dans les Open Data peuvent constituer des informations utiles aux analyses décisionnelles. Toutefois les Open Data très hétérogènes et disséminés en plusieurs morceaux de données sur le web, rendent difficile leur intégration au sein d'un entrepôt de données. Les travaux actuels sur l'intégration des Open Data proposent des processus d'intégration basés sur des Linked Open Data, dont la mise en place n'est pas automatisée. Dans cet article, nous proposons un processus visant à automatiser l'entreposage multidimensionnel des Open Data. Notre démarche repose sur la transformation des Open Data en un graphe générique et enrichi favorisant leur intégration. Ce graphe sert de support pour la définition semi-automatique et incrémentale du schéma multidimensionnel d'entreposage.

1 Introduction

Les données issues du web posent plusieurs défis pour l'informatique décisionnelle : ces données sont très riches en informations mais d'une complexité qui rend difficile leur intégration automatique dans les entrepôts de données (Ravat et al., 2010). Dans ce cadre, nous nous intéressons plus particulièrement à l'entreposage des open data qui sont un type de données du web en pleine croissance et qui ont la spécificité de contenir un nombre important d'informations utiles pour les décideurs notamment des données statistiques.

L'Open Data (ou données ouvertes) sont définies comme étant des données disponibles sous licence libre destinées à la réutilisation et à la redistribution par n'importe quelle personne (Coletta et al., 2012) (Mazón et al., 2012) (Eberius et al., 2012). Cependant elles ont d'autres propriétés telles qu'une importante hétérogénéité en format (XSL, CSV, RDF, TXT, PDF,...), en structure et en sémantique et couvrent plusieurs domaines (politique, santé, commerce ...).

Nous illustrons quelques problèmes des open data à travers les deux sources de données (datasets) de la Figure 1. Ces sources en format excel sont extraites du site data.gouv.fr. Elles montrent des statistiques sur les accidents de travail en France. Le premier dataset [1] décrit l'évolution des accidents de travail et de trajet pour le personnel civil y compris la gendarmerie

1. http ://www.data.gouv.fr/DataSet/30382535.