Dataforum : Faciliter l'échange, la découverte et la valorisation des données à l'aide de technologies sémantiques

Yoan Chabot*, Patrick Grohan**
Gilles Le Calvez***
Christèle Tarnec**

*Orange Labs Belfort,
yoan.chabot@orange.com
**Orange Labs Sophia Antipolis
***Orange Labs Lannion

Résumé. Seulement 10% des données disponibles en entreprise sont réellement utilisées tandis que les 90% restants ("dark data") restent inexploitées (Veritas, 2016). Dans une économie où la donnée est le nouveau pétrole, des outils facilitant la découverte et la compréhension de gisements de données représentent une opportunité importante. Cette démonstration présente Dataforum, une plateforme d'échange ciblant les organisations désireuses de partager de grands gisements de données d'une manière simple, rapide et fiable.

1 Introduction

Un des leviers de la création de services innovants est l'agrégation et l'analyse de données hétérogènes provenant de sources internes ou externes à l'entreprise (De Filippi, 2013). Cependant, l'ouverture de grandes masses de jeux de données entre organisations soulèvent de nombreux problèmes :

— Les données libérées découlent de connaissances métiers qui ne sont pas nécessairement possédées par leurs consommateurs. Cela complexifie la compréhension des données (étape critique, notamment pour la création de modèles par les data scientists).
— Le grand volume de données complexifie l'identification de jeux de données pertinents pour un besoin donné (une tâche à laquelle s'attaque Google Goods (Halevy et al., 2016)).
— De nombreux cas d'utilisations nécessitent de casser les barrières entre des silos de données. Cette tâche est rendue complexe par de multiples formes d'hétérogénéité des données (Pluempitiwiriyawej et Hammer, 2000).
— L'ouverture des données à des tiers constitue une évolution importante des usages. Les entreprises sont encore réticentes quant à ce partage par crainte des transgressions des consommateurs de la donnée (p. ex. l'appropriation d'un jeu de données sans accord préalable). Symétriquement, le consommateur attend des garanties quant à la qualité, la crédibilité et la traçabilité des données.

Ce papier présente Dataforum, un écosystème ouvert et auditable offrant des solutions aux challenges soulevés précédemment. Ce système, développé par Orange Labs, s'appuie sur les

technologies du Web sémantique, du traitement automatique du texte et de l'apprentissage automatique.

2 Plateforme Dataforum

L'approche choisie est d'utiliser un système de gestion de métadonnées n'impactant pas les jeux de données et les systèmes les hébergeant. L'élément central de Dataforum est un modèle de métadonnées, appelé sem4DS (SEMantics for DataSet, section 2.1), décrivant les jeux de données de manière précise et compréhensible. Afin d'assister les producteurs dans la description de leurs données, Dataforum introduit une chaîne d'algorithmes (section 2.2) suggérant des métadonnées. Enfin, grâce à des outils de recherche et de recommandation (section 2.3) les utilisateurs peuvent découvrir de nouvelles données pertinentes pour leurs besoins.

2.1 Modèle de métadonnées

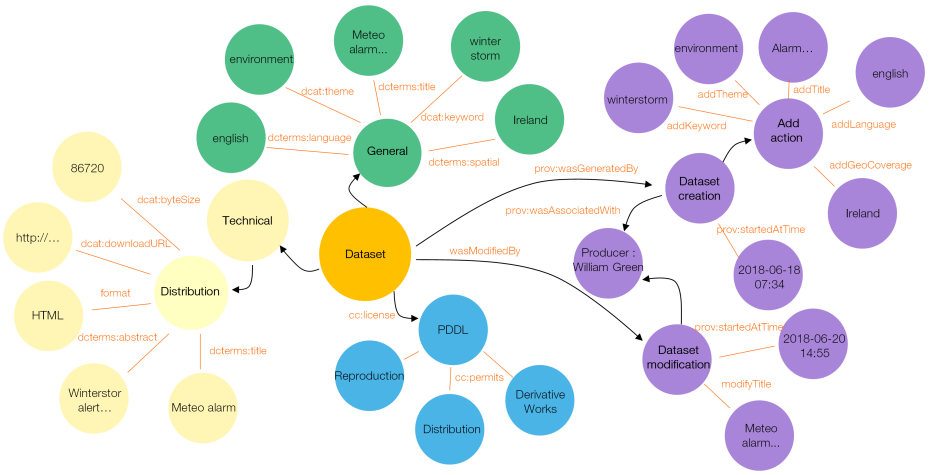

FIG. 1 – *Description d'un jeu de données à l'aide de sem4DS (descripteurs génériques en vert, techniques en jaune clair, provenance en violet et conditions d'utilisation en bleu)*

sem4DS est une agrégation de plusieurs modèles sémantiques reconnus et d'extensions développées pour les besoins de la plateforme : DCMI et DCAT pour décrire le contenu des jeux de données, PROV-O et DataId (Brümmer et al., 2014) pour spécifier la provenance des données et CCREL (Abelson et al., 2008) pour modéliser les licences. sem4DS est organisé autour de deux concepts principaux : le jeu de données et ses distributions (c.-à-d. les fichiers mis à disposition pour un jeu de données). Le modèle est découpé selon quatre catégories :

— Les descripteurs génériques décrivant le contenu des jeux de données (p. ex. titre, mots-clefs, langue, thème, couverture géographique). Ils permettent d'améliorer la recherche et la compréhension des jeux de données via une description formelle et précise de leur signification.

— Les descripteurs techniques fournissant des informations sur le support du jeu de données (p. ex. taille du fichier et format, URL de téléchargement).
— Les descripteurs de provenance spécifiant d'où proviennent les données, les personnes et les organisations contribuant au partage des données ainsi que les actions réalisées (copie, modification, etc.). Ces descripteurs permettent de connaître l'historique d'un jeu de données et de suivre ses utilisations sur la plateforme.
— Les descripteurs de conditions d'utilisation modélisant les licences apposées sur les jeux de données (c.-à-d. autorisations, interdictions et exigences que doivent respecter les consommateurs).

2.2 Sémantisation des jeux de données

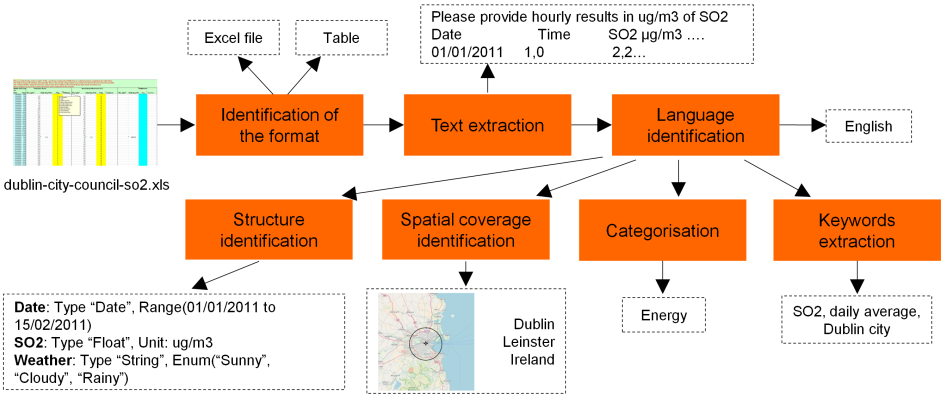

FIG. 2 – *Chaîne de sémantisation pour la caractérisation automatique de jeux de données ("Structure identification" et "Spatial coverage identification" sont des travaux en cours)*

Pour référencer un grand nombre de jeux de données et proposer des dispositifs de recommandation, il est nécessaire de disposer d'outils de caractérisation automatique adaptés à la très grande hétérogénéité des données en termes de contenu et de contenant. La plateforme propose une chaîne d'algorithmes (voir figure 2), activable à la demande par le producteur de données, générant automatiquement des métadonnées pour décrire le contenu d'un ensemble de données (langues, thèmes, mots clés, couverture géographique) et ses contenants (formats et taille des fichiers). Les traitements sémantiques se basent sur des techniques à base de trigrammes pour l'identification de la langue et à base de modèles de Markov associés à des traitements linguistiques et des techniques de recherche d'informations pour les autres métadonnées sémantiques. La chaine permet de traiter sept langues : anglais, français, néerlandais, espagnol, slovaque, roumain et allemand.

2.3 Recherche et recommandation

Le moteur de recherche de Dataforum permet à des non-initiés de sélectionner des critères de recherche (mots-clefs, thème, licence, langage, etc.) afin de générer automatiquement des

requêtes SPARQL. L'utilisateur obtient les résultats de sa recherche sous forme de "cartes" synthétiques et peut obtenir de plus amples informations via une vue graphe (présentée dans la figure 1).
La richesse du modèle sem4DS permet également de fournir des suggestions aux utilisateurs en fonction des jeux de données récemment consultés. Dataforum utilise pour cela Reperio, un moteur de recommandation développé à Orange Labs (Meyer, 2012). Dans la version actuelle de la plateforme, les mots-clefs et les thèmes sont utilisés pour mesurer les similarités entre jeux de données. Grâce à une visualisation du graphe de similarité, l'utilisateur peut naviguer de manière intuitive dans les jeux de données de proche en proche (noeuds = jeux de données, arcs = relations de similarité).

3 Perspectives

Le système présenté est un prototype de recherche en cours de déploiement pour répondre aux besoins des propriétaires de données et des data scientists de l'entreprise Orange.

Références

Abelson, H., B. Adida, M. Linksvayer, et N. Yergler (2008). ccREL : The Creative Commons Rights Expression Language.

Brümmer, M., C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, et S. Hellmann (2014). DataID : Towards Semantically Rich Metadata For Complex Datasets. In *Proceedings of the 10th International Conference on Semantic Systems - SEM '14*, New York, USA, pp. 84–91. ACM Press.

De Filippi, P. (2013). Une charte éthique pour le Big Data. *Documentaliste-Sciences de l'Information*, 8–9.

Halevy, A., F. Korn, N. F. Noy, C. Olston, N. Polyzotis, S. Roy, et S. E. Whang (2016). Goods : Organizing Google's Datasets. In *Proceedings of the 2016 International Conference on Management of Data - SIGMOD '16*, pp. 795–806.

Meyer, F. (2012). Recommender systems in industrial contexts. *arXiv preprint arXiv :1203.4487*, 170.

Pluempitiwiriyawej, C. et J. Hammer (2000). A classification scheme for semantic and schematic heterogeneities in XML data sources. *Technical report TR00-004 2000*(September).

Veritas (2016). Identify the value, risk and cost of your data. Technical report.

Summary

Only 10% of the available data is actually used while the remaining 90%, called "dark data", remains unused (Veritas, 2016). In an economy where data is the new oil, tools to facilitate the discovery and understanding of datasets represent an important opportunity. This demonstration presents Dataforum, an exchange platform targeting organizations that want to share datasets in a simple, fast and reliable way on a large scale.