

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 (Pluempitiwiriawej 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