UMR IRISA

OFFRE DE THÈSE

www.irisa.fr

Sujet de thèse •

Gestion de la qualité de données RDF issues de l'intégration de données hétérogènes

Résumé •

Contexte :

On se place dans le cadre de l'intégration automatique de données hétérogènes, provenant par exemple du Web (*open data*), ou produites par différentes entités et différents systèmes d'une entreprise, afin de pouvoir analyser les données de façon globale.

Les travaux à effectuer interviennent dans le cadre d'une collaboration avec l'entreprise SEMSOFT [1] développant une suite logicielle pour l'intégration de données.

L'objectif final des travaux est l'élaboration et la création d'une chaîne complète de traitement de gestion de données RDF d'éventuellement gros volume (*big data*), incluant l'extraction, l'intégration et la restitution des données aux utilisateurs. Le sujet de thèse proposé concerne plus particulièrement la gestion de la qualité des données dans ce contexte.

Sujet de la thèse :

Les données à exploiter présentent généralement des problèmes de qualité : elles peuvent par exemple être incomplètes, inexactes ou obsolètes [2]. Des données de mauvaise qualité peuvent être à l'origine de décisions non fondées potentiellement néfastes pour une entreprise. Il est ainsi aujourd'hui bien admis que la gestion de la qualité de données est un aspect incontournable de la gestion des données.

Cette gestion soulève un ensemble de questions, auquel des réponses à la fois méthodologiques et technologiques doivent être apportées, à savoir : Comment qualifier et mesurer la qualité des données ? Comment la représenter dans le système ? Comment réparer les données (pour les données réparables) ? Comment tenir compte des données restées imparfaites au moment de l'exploitation des données, par exemple au moment de l'interrogation et de l'analyse des données ?

Le sujet de thèse concerne l'étude de la gestion de la qualité des données dans le cadre de l'intégration de données RDF hétérogènes. Les travaux peuvent être abordés sous deux angles complémentaires :

- un angle théorique consistant en la définition de modèles, métriques et méthodes pour la gestion de la qualité de données RDF (les aspects énumérés ci-dessus mettant en jeu des notions graduelles, on pourra utiliser de la théorie des ensembles flous comme fondement théorique aux approches à définir), et
- un angle pratique consistant en l'implantation d'un logiciel indépendant et/ou intégré à la suite logicielle AGGREGO développée au sein de la société SEMSOFT.

Le candidat idéal devra avoir suivi un cursus informatique, être curieux, indépendant, dynamique, et devra être intéressé par des aspects théoriques et pratiques de l'informatique. Il sera intégré à l'équipe Shaman de l'IRISA, physiquement localisée à Lannion, sur la superbe Côte de granit rose bretonne.

Dpt scientifique •	D7- Gestion des données et de la connaissance Data and Knowledge Management
Equipe projet •	Shaman (DKM Department of IRISA) http://www-shaman.irisa.fr/
Directeur de thèse •	Olivier PIVERT, PR
Encadrant(s) •	Olivier PIVERT (PR IRISA/DKM/Shaman) Virginie THION (MCF IRISA/DKM/Shaman)
Contact(s) :	Pour candidater sur cette thèse, veuillez envoyer votre CV, une lettre de motivation et des lettres de recommandation à : Olivier Pivert <Olivier.Pivert@enssat.fr> and Virginie Thion <Virginie.Thion@irisa.fr>
Début des travaux •	Octobre 2014
Lieu •	IRISA – Lannion (http://www.ville-lannion.fr/en/accueil_en.html), Pink Granite Coast of Brittany, France. IRISA – Antenne de Lannion (http://www.ville-lannion.fr/en/accueil_fr.html), Côte de granit rose, Bretagne, France.
Bibliographie •	[1] Web Site of the SEMSOFT company. semsoft-corp.com [2] C. Batini and M. Scannapieco. Data Quality: *Concepts, Methodologies and Techniques*.
Mots clés •	Data Quality Management, Heterogeous data, RDF.