# *Fusion de données liées exploitant la sémantique de l'ontologie et fournissant une évaluation de la qualité des données.*

Contacts : Fatiha.sais@lri.fr, rallou.thomopoulos@supagro.inra.fr

L'initiative « Linked Open Data cloud (LOD) », consistant à publier des données RDF sur le Web et à les lier les unes aux autres, est aujourd'hui un phénomène mondial qui fait émerger de nombreuses applications innovantes. Le LOD contenait plus de 31 milliards de triplets RDF en 2011 et 503 millions de liens entre les données décrites par ces triplets. Etablir des liens exprimant la relation d'identité (owl:sameAs) entre données est essentiel pour exploiter au mieux toute la richesse du Web de données. Pour ce faire, toute paire de données pour laquelle un lien *owl :sameAs* aura été déclaré doit être fusionnée afin d'obtenir une représentation unifiée.

Lorsque l'on s'intéresse au problème de la fusion de données, la principale difficulté concerne les conflits dans les valeurs des propriétés, c'est-à-dire, plusieurs valeurs possibles pour une même propriété. Ces conflits sont principalement dus à l'hétérogénéité des données où différents vocabulaires et conventions sont utilisées pour décrire les données. Une mauvaise qualité des données (la fraicheur des données, les erreurs et les informations incomplètes) peut contribuer à l'amplification des conflits entre valeurs.

Il existe dans la littérature quelques approches de fusion de données (voir [J. Bleiholder and F. Naumann, 2008] pour un état de l'art) développées dans le domaine des bases de données relationnelles où les conflits sont gérés au moment de l'interrogation en utilisant des opérateurs prédéfinis, tels que Max, Min, la valeur de la source la plus fiable, la valeur la plus récente, etc. Dans [F. Saïs et R. Thomopoulos 2008], nous avons développé une méthode de fusion de données RDF où toutes les valeurs possibles d'une propriété sont conservées et stockées de façon ordonnées en fonction d'un degré de confiance associé à chaque valeur qui est calculé en utilisant différents critères, tels que la fiabilité des sources, la fréquence des valeurs, l'âge de la valeur, etc. Le formalise des ensembles flous a été utilisé pour représenter les données fusionnées. Les limites de cette approche résident dans le fait que tous les critères de choix de valeurs sont utilisés avec le même niveau d'importance et dans le fait que ni les contraintes du schéma (e.g., les cardinalités des propriétés) ni les mappings, éventuellement complexes, entre éléments de schémas (e.g., [s1.adresse = (s2.rue, s2.ville, s2.CP, s2.pays)]) ne sont pris en compte.

Le but de ce stage est d'étudier les aspects théoriques et pratiques d'une approche de fusion de données où : (i) il est possible de choisir et de combiner plusieurs critères (e.g., âge, fréquence, fiabilité des sources, des fonctions dépendantes du domaine) de choix de valeurs, (ii) les contraintes du schéma sont vérifiées dans les données fusionnées et (iii) des informations sur la provenance des données telles que les sources d'origine mais aussi les mappings appliquées sur les schémas des sources de données sont prises en compte. De plus, la richesse du Web de données devra être exploitées en navigant dans le graphe des liens *owl :sameAs* connus, pour avoir des information plus précises (dates de naissance sur *dbpedia*) et parfois plus fiables (coordonnées géographiques sur *geonames*).

Le stage s'inscrit dans le cadre du projet ANR Qualinca[1] dont le but est d'évaluer et de représenter la qualité des données de bases documentaires gérées par l'ABES (Agence Bibliographique de l'Enseignement Supérieur) et par l'INA (Institut National de l'Audiovisuel).
Les résultats de ce stage devront être validés sur les données bibliographiques de l'ABES et de l'INA.

Compétences requises (recommandées) :

- Web Sémantique, Ontologies, RDF/OWL, logique et raisonnement,
- Bases de données, Logique floue, Java/XML.

Co-encadrement : Fatiha Saïs et Rallou Thomopoulos
Lieu du stage : LRI – Université Paris Sud (Orsay)
Rémunération : oui. Durée du stage : 6 mois (03/2014 – 09/2014)

[1] http://www.lirmm.fr/qualinca/?q=en/objectifs