UMR IRISA

PHD OFFER

www.irisa.fr

Sujet de thèse •
PhD Subject Title:

Data Quality Management of heterogeneous RDF data

Résumé •
Abstract:

Context:

The subject of the thesis deals with the general problem of integrating and then analyzing heterogeneous data, for instance stemming from internal data sources of a company and external sources like *open data* published on the Web.

We treat this subject in cooperation with the SEMSOFT company [1], which produces a software chain for heterogeneous data integration.

Our final goal is to define a complete chain for the management of RDF (big) data, for extraction, integration, and reporting of data. The thesis focuses on data quality management throughout this process.

PhD Subject:

Data quality problems such as duplicates, inaccuracies, outdated data, contradictory beliefs, missing or incomplete data are omnipresent in data sources and widespread in every governmental, industrial, commercial and personal information systems. Recent studies showed that loss of quality can lead to considerable (or even disastrous) financial consequences. Data quality management [2] has then become a vital task of data management for every company.

Managing data quality implies dealing with methodological and technological problems like “How to define and measure quality?”, “How to design and store quality in an information system?”, “How to improve data quality (how to repair data if possible)?”, “How to deal with low quality data when using data (eg. when querying and analysing data)?”

In this PhD subject, we will deal with data quality management when integrating RDF heterogeneous data, tackling two complementary facets:

- A theoretical facet for defining model, metrics and method for RDF data quality management (possibly with approaches based on fuzzy set theory) , and
- A practical facet for the implementation of a software, which could be a standalone one or a plug-in for the AGGREGO software of the SEMSOFT company.

A perfect applicant should have strong background in Computer Sciences, and should be inquisitive, autonomous, dynamic and interested both in theoretical and practical aspects of the subject.

The student will be part of the Shaman team of the IRISA, located at Lannion, on the beautiful Pink Granite Coast of Brittany in France, which is a “paradise for nature lovers”.

Dpt scientifique •
Scientific department:

D7- Data and Knowledge Management

Equipe projet •
Research team:

Shaman (DKM Department of IRISA)
http://www-shaman.irisa.fr/

Directeur de thèse •
PhD Director:

Olivier PIVERT, PR

Encadrant(s) • PhD supervisors:	Olivier PIVERT (PR IRISA/DKM/SHAMAN) Virginie THION (MCF IRISA/DKM/SHAMAN)
Contact(s) :	To apply for this position please send CV, cover letter and recommendations to: Olivier Pivert <Olivier.Pivert@enssat.fr> and Virginie Thion <Virginie.Thion@irisa.fr>
Début des travaux • Work start date:	October 2014
Lieu • Place	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 • References:	[1] Web Site of the SEMSOFT company. semsoft-corp.com [2] C. Batini and M. Scannapieco. Data Quality: *Concepts, Methodologies and Techniques*.
Mots clés • Keywords	Data Management, Quality, Heterogeous data.