Please convert this file to PDF format before uploading it to http://www.u-psud.fr/csc/

Discipline

Computer Science

Doctoral School

Ecole Doctorale Informatique Paris-Sud (ED 427)
Computer Science Doctoral School of Paris-Sud (ED 427)

Thesis subject title
Knowledge-based Data Fusion combined with Data Quality Assessment

- **Laboratory name: LRI-Laboratoire de Recherche en Informatique**
- **Laboratory web site: http://www.lri.fr**

- **PhD supervisor (contact person)**
  - **Name: Fatiha SAIS**
  - **Position: Associate Professor**
  - **Email: Fatiha.Sais@lri.fr**
  - **Phone number:+33 (0)1 69 15 68 42**

- **Thesis proposal (max 1500 words)**

Nowadays, the *Web of Data* is one of the most important fields of the *Semantic Web*. It is increasingly used and acknowledged in a variety of application areas. It can be described as a structured way to store data and to interconnect them with meaningful correspondences. These connections among data objects are extremely useful, as they allow different and often heterogeneous data to be explored and queried by applications, thus expanding the data space.

The *Linked Open Data* (LOD) project, conceived in 2007, is a fundamental initiative in this direction. It supports the aggregation and interconnection of numerous data that are already available on the Web. Rapidly growing, it now contains over 58 billion RDF triples linked with over 719 million RDF links, composing an enormous area of knowledge. Within the setting of LOD, it is often the case that objects, possibly coming from different sources, represent the same real-world entity. In order to maintain the usability of the LOD, it is critical to detect this kind of relations between objects and to attempt to obtain one single object describing the real-world entity. Naturally, data is constantly evolving, new data are generated and creating links between objects becomes more an more complex. It is, thus necessary to use automatic procedures to this end.

Data linking, also known as *data reconciliation*, is the process where two object descriptions are examined in order to determine whether they refer to the same real-world entity, and if so, to link them together.

Then, data fusion encompasses the effort to acquire a single homogenized object by merging the conflicting information of the linked individual objects. The objects marked with the *owl:sameAs* may contain different, conflicting or inconsistent values in their properties. For each property, the conflict of its different values must be compromised and the most appropriate value must be chosen. The task of data fusion consists in merging the individual representations by resolving these conflicts and generating one single final object containing the entirety of the real-world entity's information. The data fusion is an essential step towards avoiding redundancy, grouping together the best quality information and giving consistent answers to the users, in the linked data environment.

In this PhD project we will explore and study the problem of data fusion. Clearly, in the setting of linked data, there exist various different reconciled datasets whose objects are linked with semantic *owl:sameAs* relations. We attempt to merge the often conflicting information of these reconciled objects in order to obtain unified representations that only contain the best quality information. During data fusion process, the main difficulty resides in the conflicts between property values, i.e. choose one value or one set of values from the possible values of one property. These conflicts are mainly due to the heterogeneity and to the low quality of the data (freshness, errors, information incompleteness, difference in granularity level). The data fusion approach will allow: (i) to select and combine different criteria (e.g., age, frequency, source reliability, domain-dependent functions) of value choice, (ii) ensure that the domain constraints are satisfied and (iii) take into account informations on data provenance.

In this project we will also study how domain knowledge usually represented in an ontology can be used to better determine the best property values. Indeed, domain knowledge as property functionality, generalisation/specialisation, disjunction and synonymy relations can be used to solve some of the value conflicts. Furthermore, more dataset-dependent knowledge can also be exploited in data fusion. This knowledge may express constraints on property values, semantic dependencies and functional dependencies between properties. For example, one can exploit the knowledge on the fact that in bibliographic datasets, the value of the property *birthDate* of an author should be lower than the *publicationDate* value of one of his publications. In addition, to keep track of information concerning the data fusion process, i.e. the reasons and the knowledge that conducted to a fusion decision, in this PhD we will also study and design a data fusion provenance model. Thus, a data fusion explanation step become straightforward. Finally, we will study an extension of the developed data fusion approach to fuse descriptions that do not refer to the same real world object but to the same abstract object. For example, fuse the two book descriptions representing two different editions or translations of the same art of work (book).

For the evaluation and the validation of the PhD results, real RDF data available on the LOD can be used. The LOD datasets concern a big variety of application domains like government, life science, bibliographic, social and so on. Other datasets on the domain of biorefineries will be provided by our partner INRA (UMR IATE). It will be used in a context of determining new experimental tendencies in this domain and to improve the data quality.

This PhD project, we be under the supervision of Fatiha Saïs (Associate Professor at Paris Sud University), Patrice Buche (Research Engineer -HDR from INRA) and Rallou Thomoploulos (Researcher HDR at INRA).

- **Publications of the laboratory in the field (max 5)**

1. Saïs F. and R. Thomopoulos R. (2008) Reference Fusion and Flexible Querying. In Proceedings of OTM Conferences 2008, Mexico, November 2008, Lecture Notes in Computer Science #5332, Springer, pp. 1541-1549

2. Destercke S., Buche P., Charnomordic B. Evaluating data reliability: an evidential answer with application to a Web-enabled data warehouse. IEEE TKDE 25(1): 92-105 (2013).

3. Saïs F., Pernelle N., Rousset M.C. Combining a Logical and a Numerical Method for Data Reconciliation (2009). In Stefano Spaccapietra (Ed.): Journal on Data Semantics XII, Springer LNCS 5480, Volume: 12(66-94)

4. Buche P., Dibie-Barthélemy J., Khefifi R., Saïs F. (2011). An Ontology-based method for Duplicate Detection in Web Data Tables. Dexa 2011 proceedings, Lecture Notes in Computer Science #6860, pp 511-525.

5. Symeonidou D., Armant V., Pernelle P., Saïs F. (2014): SAKey: Scalable Almost Key Discovery in RDF Data. Semantic Web Conference (1) 2014: 33-49

Specific requirements to apply, if any

Skills in databases, Logics and Web technologies are welcomed. Interest in real-world applications and good level in English or French required. We target Phd candidates willing to have an experience in Semantic Web technologies, big data and data quality problems.