

Linked Data Annotation and Fusion driven by Data Quality Evaluation

Ioanna Giannopoulou⁽¹⁾, Fatiha Saïs⁽¹⁾ and Rallou Thomopoulos⁽²⁾

⁽¹⁾ LRI (Paris Sud University & CNRS)

Bât. 650 Ada Lovelace, F-91405 Orsay Cedex, France

⁽²⁾ INRA-Supagro, LIRMM (Montpellier 2 University, CNRS & Inria)

2 place P. Viala, F-34060 Montpellier cedex 1, France

Résumé. In this work, we are interested in exploring the problem of *data fusion*, starting from reconciled datasets whose objects are linked with semantic 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.

1 Introduction

Data linking, also known as *data reconciliation* Ferrara et al. (2013); Saïs et al. (2009); Pernelle et al. (2013), 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 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 most appropriate value must be chosen. 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.

Research on the data fusion problem has begun over two decades ago in the field of relational databases Bleiholder et Naumann (2008). However, as we examine the data fusion from the RDF point of view, we notice that the specificities of RDF mechanisms cannot be reflected in solutions offered by relational databases experts. Three main approaches have been proposed for data fusion in RDF Saïs et Thomopoulos (2008); Saïs et al. (2010); Flouris et al. (2012); Mendes et al. (2012). These different approaches attempt to evaluate the quality of each value, by taking into account various measures based on the value itself and/or its metadata.

In this work, we are interested in exploring the problem of data fusion. Our method combines different quality criteria based on the value and its data source, and exploits, whenever possible the ontology semantics, constraints and relations. What is more, we create a mechanism to provide explanations about the quality of each value, as estimated by our system. To achieve this, we generate annotations used for traceability and explanation purposes.

Our approach is described in detail in Section 2. A first evaluation is presented in Section 3. Finally, Section 4 concludes the paper and gives some future work.