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.

2 Proposed solution

Inspired by the approaches cited in Section 1, we build a system that attempts to offer a solid automated solution to the fusion problem. With respect to the description of conflict-handling methods described in Bleiholder et Naumann (2008), our approach is automatic and applies a conflict-resolution strategy. We will show that it is at the same time an instance based and a metadata based strategy.

As it will be shown in the next sections, this solution will (i) be adapted to each property, (ii) exclude values that are *implausible* according to specific rules, (iii) exploit, whenever possible, the ontology knowledge (constraints, relations, etc.) and (iv) natively treat multi-valued properties. Most importantly, it will store the information that led to each fusion decision, in order to offer explanations on how and why a value has been chosen.

2.1 Methodology

Detect implausible values. We use the notion of plausibility in order to exclude values that are *irrational*, according to some measures or predefined logic rules. This step is based on the following elements :
Frequency. The measures of homogeneity and occurrence frequency in the data sources as defined in Saïs et Thomopoulos (2008), are used here in order to detect the implausibility : if the occurrence frequency is lower than a preset threshold, then the value is considered as implausible.
Domain constraints. Some domain constraints and property typing restrictions can also be used to detect implausible values. For example, if a property *age* is typed as "*xsd :nonNegativeInteger*", then a negative value of *age* should be considered as implausible.

Calculate value quality score. For the values that are not considered as *implausible*, a quality score is computed. We believe that the aspects of the value itself (e.g. how often it appears in a specified set) and the aspects of the quality of its data source can be equally important, and that is why, we resulted in a list of the following for criteria :
The homogeneity of a value. The frequency of appearance of a value among all the values for a property within the group of reconciled references.
Its occurrence frequency. The frequency of appearance of a value among all the values for a property within the whole dataset. If a value appears several times inside a dataset for the same property, then it is more reliable as it is less likely to contain spelling errors.

The freshness of its source. The last update of the data source. Data sources that are recently updated tend to be considered as more reliable.

The reliability of its source. The users can explicitly give their preference towards one or another data source. These preferences are used to fix a reliability value in $[0..1]$. If no such preference is given, all data sources are considered equally reliable.

For more detailed descriptions on the homogeneity, the occurrence frequency and the freshness, see the definitions in Saïs et Thomopoulos (2008).

The global quality score of a given value is the average of these criteria is computed. Indeed, other aggregation functions can be used as a weighted average, a maximum, a minimum, etc. The score is then normalized and represented on a qualitative scale with the following values :

excellent, medium, poor. As we will see below, this scale will serve for querying the information on the metadata ontology.

Discover relations. For any value that is not considered as *implausible*, the system is trying to discover how the value is related to the other possible values for the property. The possible relations are :

More Precise : a string inclusion control is applied here, as well as querying a knowledge base to detect possible subsumption relations for hierarchical values to select the more precise value. Examples : *Paris* is more precise than *France*, "10/05/1999" is more precise than "1999".

Synonym : several APIs for detecting synonyms are used to determine whether two values are synonyms. Example : *England* is a synonym of *UK*. In case the property is explicitly defined as mono-valued, the value with the best quality score is chosen. Otherwise, both values are kept in the property.

Incompatible : the system refers to a list of expert rules to detect possible *logical* incompatibilities between the values of one property or several properties.

For example, if we declare in a knowledge base that a property $hasFunction$ is functionnal and : $hasFunction(X,'' PrimeMinister'') \wedge hasFunction(X,'' President'') \Rightarrow \bot$ then if the property $hasFunction$ has these two values, these last will be declared as *incompatible*.

According to the relations, the quality score is affected positively or negatively, by a bonus or a malus. The values are then sorted by their quality score. For mono-valued properties, the value with the highest quality score is chosen. For multi-valued properties, all the plausible values are kept. For each value, the system stores all the information that led to the calculation of the score, or the reason of *implausibility*.

2.2 Provenance of Fusion Decisions

Since the calculation of the quality score and the selection of the appropriate value is more complex, we realize that offering explanations on the provenance of the fusion decisions is a useful feature. To achieve this, we found the reification mechanism, also used in Saïs et Thomopoulos (2008), more flexible and suitable for our representation needs. The reification mechanism allows to enrich the RDF declarations by adding new elements. Substantially, it offers the possibility to create a metadata ontology describing existing RDF triples.

The data fusion metadata ontology. We introduce a data fusion metadata ontology in order to use the RDF reification mechanism to annotate fused data by the quality information that are exploited to achieve the fusion decisions (see Figure 1). The $rdf:Statement$ class is enriched by the object property $hasQuality$. The main class $Quality$ has three object properties which organize the different quality aspects. The $hasCriteria$ property groups up all the measures used to calculate the quality score, the $hasRelations$ brings together the relations of the value with other values, and the $hasIndicators$ contains the remaining information.

A metadata document conforming to this ontology is produced by the system as a result of the data fusion algorithm. It provides descriptions for the quality measures of the value and the reasons why it was chosen or excluded (e.g., which rule it violates, the values that are more precise than it, etc.). However, the initial purpose of the metadata document is to be automatically queried. Depending on the quality scale of a value ($excellent, medium, poor$), different queries are intended to provide a "story" explaining the aspects of the value's quality.

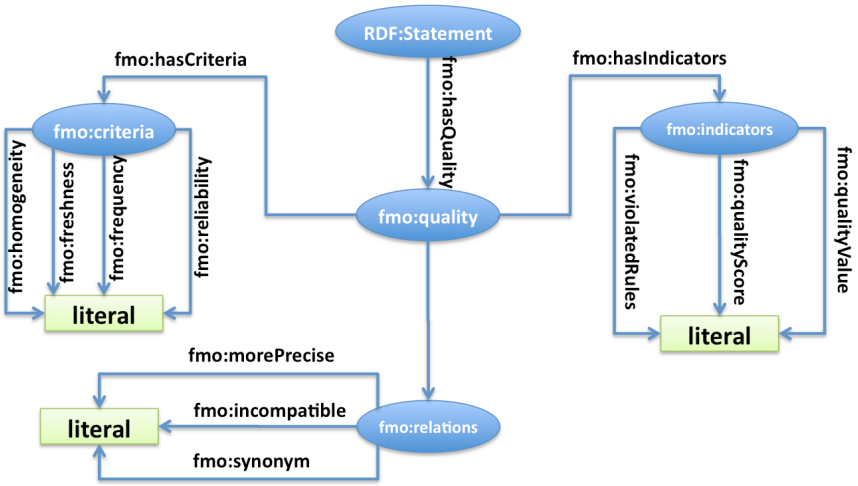

FIG. 1 – *The Metadata Ontology*

Example. An excerpt of the metadata document is displayed below. We are using the namespace *dfa*, standing for *data fusion annotation*. In this example, the two values for the property $dfa:first\_name$ are presented and annotated with all the useful information concerning their quality. Specifically, the implausible value *Jacues* contains only the information about its homogeneity, which is very low (0.015) and, thus, accounts for its exclusion from the list of plausible values. On the other side, for the plausible value *Jacques*, all the information concerning the calculation of its quality score are contained in the file. For this example, if the system queries the metadata document for the value *Jacues*, the answer would be *"The value is implausible due to very low appearance in the data sources. Possible reason : misspelling."*, while for the value *Jacques* it will be *"The value is the only plausible one. It has a combination of high appearance in the data source and high level of trust on its data source"*.

```
PERSON-17123430093 rdf:type PhysicalPerson
v1 rdf:type Value
q1 rdf:type Quality
c1 rdf:type Criteria
PERSON-17123430093 dfa:first_name v1
v1 dfa:hasValue Jacques
v1 dfa:isImplausible false
...
PERSON-17123430093 dfa:first_name v2
v2 rdf:type Value
q2 rdf:type Quality
c2 rdf:type Criteria
v2 dfa:hasValue Jacues
v2 dfa:isImplausible true
v2 fmo:hasQuality q2
q2 fmo:hasCriteria c2
c2 fmo:hasHomogeneity 0.015
```

3 Experiments

In this section, the main data sources that are used to test our approach are provided by the INA partner of the ANR project Qualinca.

We considered a set of groups of 10819 reconciled instances (pairwise linked using $owl:sameAs$ links) that represent french famous persons and content notices where they are involved. In Table 1.(a) we show the distribution of the reconciled instances groups. These instances are described using different properties, as $aPourNom$, $aPourTitreCollection$, $aPourDateDiffusion$, and so on.

Table (a)

Person	# instances
Jacques Martin	10288
Philippe Bouvard	264
Daniel Prevost	214
Frederic Martin	26
Emmanuel Petit	12
Luis Fernandez	7
Michel Leclerc	6
Virginie Lemoine	2

Table (b)

	#values	%
#distinct values	14588	–
#isImplausible = “true”	9370	64.23 %
#isImplausible = “false”	5218	34.76 %
#qualityValue = “excellent”	2	0.04 %
#qualityValue = “medium”	3233	61.95 %
#qualityValue = “poor”	1983	38 %

TAB. 1 – *Table (a) Groups of reconciled instances ; Table (b) First data fusion results*

In Table 1.(b) we show the first results obtained by our data fusion approach. From the annotation file we extracted the number of distinct values, the number of values that are detected as implausible and the ones that are plausible thanks to the frequency computation. The three-valued scale (*excellent, medium, poor*) was validated with the domain experts, since it captures the intuitive idea of neutral / positive / negative score. It could be more precise (with 5 values for instance) but this would still be a refinement of the three-valued scale. For the quality value that is computed we used three values of thresholds to determine them :

- if qualityScore $\geq$ *0.67 then qualityValue = “excellent”.*
- if 0.33 <qualityScore > 0.67 then qualityValue = “medium”.
- if qualityScore $\leq$ *0.33 then qualityValue = “poor”.*

What we can observe from these results is that more than 64% of values are detected as implausible and considered as inappropriate for the corresponding properties. Furthermore, from the plausible values almost 62% of the values have a quality value that is *medium* what is consolidate the results of the first step concerning the selection of the plausible values. More qualitative experiments are needed to better qualify the reasons why the 38% of values having a *poor* quality value appear as plausible.

4 Conclusion

In this work, we have presented an effort to study in depth the problem of data fusion in the context of Linked Data. We have shown that our approach provides strong additions

on the calculation of a value quality score and includes the original idea of keeping track of the provenance of the fusion decisions. The annotation ontology we have proposed in order to represent the aspects of the value quality, offers the possibility of automatical querying to obtain specified responses and explanations. We have achieved a first implementation of the system and studied several bibliographic datasets as well as ways to evaluate our experiment results.

Several directions will be worth exploring. In particular, it would be interesting to look into different scenarios to exploit the fused data, as well as use further quality criteria. We could also experiment with different combinations of the criteria. Additional experiments will allow defining generic thresholds for our variables and to discover the best combination of weights for the calculation of the quality score.

5 Acknowledgment

This work is supported by the French National Research Agency : "Quality and Interoperability of Large Catalogues of Documents" project (QUALINCA-ANR-2012-CORD-012-02).

Références

Bleiholder, J. et F. Naumann (2008). Data fusion. *ACM Comput. Surv. 41*(1).

Ferrara, A., A. Nikolov, et F. Scharffe (2013). Data linking. *J. Web Sem. 23*, 1.

Flouris, G., Y. R. andMaria Poveda-Villalon andPablo N. Mendes, et I. Fundulaki (2012). Using provenance for quality assessment and repair in linked open data. In *In Proceedings of the 2nd Joint Workshop on Knowledge Evolution and Ontology Dynamics (EvoDyn-12)*.

Mendes, P. N., H. Mühleisen, et C. Bizer (2012). Sieve : linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012*, pp. 116–123.

Pernelle, N., F. Saïs, et D. Symeonidou (2013). An automatic key discovery approach for data linking. *J. Web Sem. 23*, 16–30.

Saïs, F., N. Pernelle, et M. Rousset (2009). Combining a logical and a numerical method for data reconciliation. *J. Data Semantics 12*, 66–94.

Saïs, F. et R. Thomopoulos (2008). Reference fusion and flexible querying. In *Proceedings of On the Move to Meaningful Internet Systems : OTM 2008 Confederated International Conferences, CoopIS, DOA, GADA, IS, and ODBASE 2008, Part II*, pp. 1541–1549.

Saïs, F., R. Thomopoulos, et S. Destercke (2010). Ontology-driven possibilistic reference fusion. In *Proceedings of On the Move to Meaningful Internet Systems, OTM 2010 - Confederated International Conferences : CoopIS, IS, DOA and ODBASE, Part II*, pp. 1079–1096.

Summary

Dans cet article nous présentons une approche de fusion de données fondée sur l'utilisation d'informations sur la qualité des données pour résoudre les éventuels conflits entre valeurs.