DC Proposal: Towards Linked Data Assessment and Linking Temporal Facts

Anisa Rula

University of Milano-Bicocca,
Department of Computer Science, Systems and Communication (DISCo),
Innovative Techonologies for Interaction and Services (Lab),
Viale Sarca 336, Milan, Italy
anisa.rula@disco.unimib.it

Abstract. Since the Linked Data is continuously growing on the Web, the quality of overall data can rapidly degrade over time. The research proposed here deals with the quality assessment in the Linked Data and the temporal linking techniques. First, we conduct an in-depth study of appropriate dimensions and their respectively metrics by defining a data quality framework that evaluates, along these dimensions, linked published data on the Web. Second, since the assessment and improvement of the Linked Data quality such as accuracy or the resolution of heterogeneities is performed through record linkage techniques, we propose an extended technique that apply time in similarity computation which can improve over traditional linkage techniques. This paper describes the core problem, presents the proposed approach, reports on initial results, and lists planned future tasks.

Keywords: Linked Data Quality, Quality Assessment, Temporal Linking.

1 Problem Definition

Data quality is an important issue for data driven applications which should be deeply investigated and understood. As a consequence of non controlled quality of the data that flows across information systems, the overall data can rapidly degrade over time. The literatures provides a wide range of techniques used to assess and improve the quality of data, such as record linkage, business rules, and similarity measures [2]. However, the quality becomes more complex and controversial as a consequence of networked-based structure (such as the web), where the amount of data evolve and it becomes more complex to be controlled.

Our focus is based on the assessment of data sets represented by structured data published on the web, known as Linked Data [4]. The aspect of quality in Linked Data is considered as an important task to consumers for a number of obvious reasons: they need data to be correct, thus, they need to have the ability to select and compare data from different sources to detect and correct errors in the data sets. Missing values or duplication can lead to applications not realizing the full potential of exchanging data.

L. Aroyo et al. (Eds.): ISWC 2011, Part II, LNCS 7032, pp. 341–348, 2011.

Techniques as record linkage are mostly used to assess and improve the data quality of the information systems. Although these techniques have been adapted in the Linked Data context [11], they ignore as well as the traditional one that real-world entity can evolve over time and can fall short for temporal data. For example, a person can change her phone number and address and so facts that describe the same real-world entity at different times can contain different values. Identifying entities that refer to the same concept enables interesting longitudinal data analysis over such data. Thus, the representation of temporal entities within the Linked Data cloud is an essential step of the linking which provide linking in a temporal context and likewise evaluate the quality of the Linked Data.

As a possible scenario, we can consider the Digital Bibliography & Library Project (DBLP)[1], published as Linked Data. It is one of the largest collection of bibliographic metadata about computer science publications over many decades. Although in general the DBLP data is of a very high quality, we have noticed some quality problems by querying that data. We wish to identify individual authors such that we can list all publications by each author. This data set contains temporal entities over a long period of time; each entity is associated with a time stamp and describes some aspects of a real-world entity at that particular time. A necessary extension to the traditional linking should be approached by incorporating the time concept to the entities.

In particular, this PhD work will concentrate on assessing the quality of Linked Data which is divided into two parts. First, to solve the above mentioned problem we aim at providing data quality dimensions as well as related metrics and validation tools which are mandatory for the assessment of quality of the published data (Linked Data). Second, as a continuously work of quality assessment we aim at providing an in-depth study about the matching heuristics defined in the context of Linked Data and the application of the approach proposed in [16] for linking temporal facts that describe the same real-world entity over time and so be able to trace the history of that entity.

2 State of the Art

Quality Aspect in Linked Data

The assessment of data quality is considered as a continuous cycle involving four major steps: the definition of quality dimensions, measuring these dimensions through sound and measurable metrics, and analyzing the results [21]. Starting from some previous works which describe six most important classifications of quality dimensions, it is possible to define a basic set of data quality dimensions, including accuracy, completeness, consistency, and timeliness. Concerning the Web, a model that associates quality information with Web data is proposed in [18]. Several dimensions are considered, such as volatility, completability, and semantic and syntactic accuracy.

[1] http://dblp.l3s.de/d2r/

Linked Data, as all the data driven application need a thorough assessment. But the assessment of Linked Data poses a number of unique challenges: due to the structured nature of Linked data published in an open environment such as the web. A comprehensive study of various problems related to the quality in Linked Data have been conducted in [13]. In fact the use of incompatible levels of abstraction makes complex a true context-sensitive analysis in elaborating query answering and visualization scenarios. Broken links in Linked Data or the ambiguous use of owl:sameAs are some of the data quality errors that can reduce the usability of a Linked Data approach. Broken Linked Data appears when it is impossible to retrieve the content of structured data due to server errors or the general unavailability of a reference. The owl:sameAs property is used to connect different data element to support semantic data integration. That is, any two URI references connected by owl:sameAs should be the same thing. But in reality, the correctness cannot be ensured and some analysis in the literature underline the different semantic associated by different designers to the owl:sameAs properties. Therefore, the authors in [9] conduct an empirical study and proposed several components of a general strategy for integrating and fusing information from the URIs in an owl:sameAs network. An approach has been proposed for quality and trustworthiness assessment based on provenance information in [10]. A description of dimensions and related metrics for the assessment of Linked Data are partly a contribution of Web community [1]. With the goal to evaluate a quality-driven information filtering a framework is proposed in [3] which supports information consumers in their decision whether to accept or reject information. This framework requires the contribution of the consumer on writing the policies.

Temporal Linking Aspects in Linked Data

Record linkage considers a set of records as input and discovers which of them refers to the same real-world entity, even if the records are not identical. They typically relies on string comparison techniques which compare multiple properties of the entities that are to be interlinked. The first studies were introduced in the statistic community [7]. Record linkage, also called identity resolution or duplicate detection, is a well-known problem in database community [5] as well as in the ontology matching community [6].

Recently, this approach is finding application in a new community such as the Linked Data. The usual approach uses automatic or semi-automatic record linkage heuristics to generate links between data sources. Silk – a Link Discovery Framework is a toolkit used for discovering and maintaining data links between Web data sources [20].

Considering the time evolution of entities in the the record linkage, some approaches have been developed [16]. The value evolution over time has been addressed in [16] by introducing the concept of decay applied in a global fashion. This approach could be also employed in our solution. Within the Semantic Web community, the representation of temporal information encodes the semantics into a time ontology which describes the temporal content of Web pages and the temporal properties of Web services [12]. Some other existing approaches include *temporal RDF* for the representation of temporal information which introduce time in RDF by assigning a number t for the temporal validity of a triple [8]; *versioning* which suggests that the ontology has different versions, one per instance of time [15]; *named graph* to implement temporal graphs which were designed to handle statements temporal validity [19], etc. There is also another approach based on tracing knowledge evolution over time which extract temporal facts to build a large-scale temporal information system [22].

3 Proposed Approach and Methodology

The contribution of this PhD work is twofold: (i) enrich the actual quality assessment framework defined in [3] by adding a new component composed by a set of quality dimensions, new measures and validation tools for higher quality of the published data; (ii) define new algorithms for performing linking between two data sets considering the temporal aspect of entities in the Linked Data.

In the following, we give an overview of the methodology which we want to follow to come up with the aforementioned contributions.

New Component for Data Quality Assessment

While there are significant overlaps, our approach focuses on assessing the quality of structured data on the web by defining a ”filtered” set of quality dimensions which fit better to the user or application requirement, rather than a continuously creation of new dimensions and separated methodologies. In this context we introduce the concept of Data Profiling (DP) defined as the application of data analysis techniques to existing data sources for the purpose of determining their quality. Therefore, the aim of this work is to define a data quality framework as a set of guidelines and techniques that, starting from input information describing a given application context, defines a rational process to assess and improve the quality of published data. Furthermore, our intent is to drive through an automatic or semi-automatic data quality approach. The achievement of an automatic data quality framework raises complex research issues and challenges, which we intend to tackle in this PhD. More precisely, we will focus on the creation of a repository of quality dimensions that are interesting for Linked Data purpose (both consumer and producer view point), for each dimension at least one metric (subjective or objective) and a framework able to analyse the data sources by means of probes that implements the above defined metrics. Results of this analysis will be shown in a dashboard so that it could be easier to understand the quality of exposed data source.

As a first step work we introduce a framework based on green engineering aspects where we have extracted only 9 of the original 12 principles with a short description, the dimensions in which they expand and measures for the assessment of linked data on the Web. [14]. The evaluation of the data will be done through validators which will consist of open source or off-the-shelf algorithms offered in the Web of Data community, as well as new validators (e.g. to check comprehensibility) that we are implementing.

Temporal Linking in Linked Data

The approach followed in this paragraph is inspired by legacy related work for structured data from the database community since the record linkage techniques are used to assess and improve the quality. In fact, if data quality issues are related to the accuracy and completeness dimensions which represent a quality aspect of the data then the improvement method is targeted to the record linkage technique.

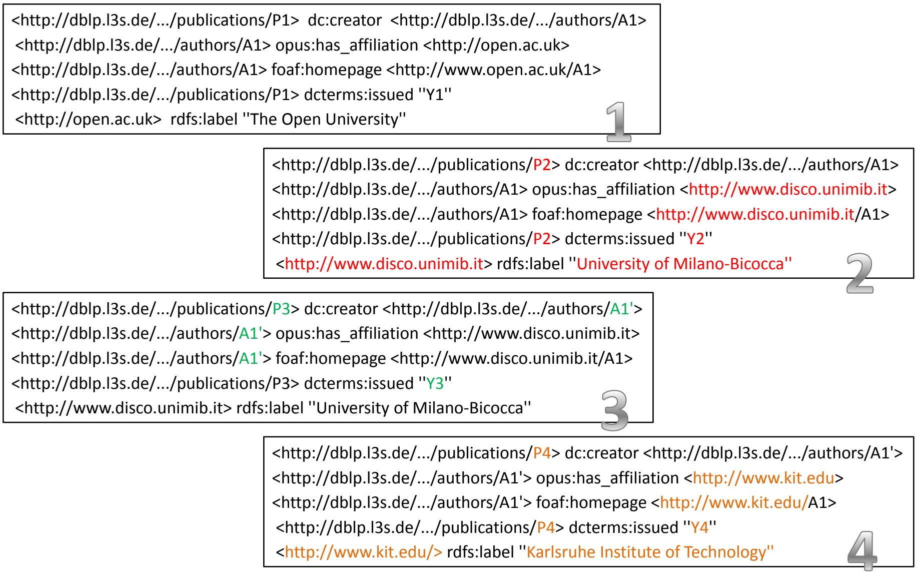
However, the temporal record linkage gain an important role since it goes further from the traditional record linkage for the quality assessment. Therefore, considering the temporal aspects within the Linked Data is an essential step in order to provide connection of the same entities expressed in different time stamp.

A first step of defining a temporal record linkage will be targeted at transferring techniques from relational databases to the Linked Data environment [17]. For instance, it has been observed that the record linkage is a well-known problem in database community [5] and many of the techniques from these fields are directly applicable in the Linked Data context [11]. Thus, we will appropriately adapt the use of "time decay", which aims to capture the effect of time elapse on entity value evolution. As an example, let us consider RDF triples that describe paper authors. Let consider two real world persons: A1 and A1'.

In Figure 1 we have author A1 who was at "The Open University" in year Y1; then A1 moved to "University of Milan Bicocca" in year Y2; A1' describe another entity, this author moved from "University of Milan Bicocca" in year Y3 to "Karlsruhe Institute of Technology" in year Y4.

Despite the challenges, temporal information does present additional evidence for linkage. We can notice that the if we consider A1 and A1' as the same person he is moving back and forth from one university to another. Exploring such evidence would require a global view of the facts with the time factor in mind. In particular we want to apply the concept of *false positive* and *false negative* related to the time decay. In particular, we consider *false negative* when there are changes on a value then it is not necessarily considered that these values are referring to different entities; we consider *false positive* when a value remain the same with a long time gap then it is not referring to the same entity. Afterwards we plan to learn decay from labelled data and apply it when computing links between entities.

An overview of what will be followed during the research is briefly described below. First, when creating explicit data links between entities, the traditional link discovery technique reward high value similarity and penalize low value similarity. However, as time slip away, values of a particular entity may evolve. Meanwhile, different entities are more likely to share the same value(s) with a long time gap. Thus, decay is introduced to reduce the penalty for value disagreement and reward for value agreement over a long period. We expect to have better results by applying decay in similarity computation.



```
<http://dblp.l3s.de/.../publications/P1>  dc:creator  <http://dblp.l3s.de/.../authors/A1>
<http://dblp.l3s.de/.../authors/A1> opus:has_affiliation <http://open.ac.uk>
<http://dblp.l3s.de/.../authors/A1> foaf:homepage <http://www.open.ac.uk/A1>
<http://dblp.l3s.de/.../publications/P1> dcterms:issued "Y1"
<http://open.ac.uk>  rdfs:label "The Open University"
```

```
<http://dblp.l3s.de/.../publications/P2> dc:creator <http://dblp.l3s.de/.../authors/A1>
<http://dblp.l3s.de/.../authors/A1> opus:has_affiliation <http://www.disco.unimib.it>
<http://dblp.l3s.de/.../authors/A1> foaf:homepage <http://www.disco.unimib.it/A1>
<http://dblp.l3s.de/.../publications/P2> dcterms:issued "Y2"
<http://www.disco.unimib.it> rdfs:label "University of Milano-Bicocca"
```

```
<http://dblp.l3s.de/.../publications/P3> dc:creator <http://dblp.l3s.de/.../authors/A1'>
<http://dblp.l3s.de/.../authors/A1'> opus:has_affiliation <http://www.disco.unimib.it>
<http://dblp.l3s.de/.../authors/A1'> foaf:homepage <http://www.disco.unimib.it/A1>
<http://dblp.l3s.de/.../publications/P3> dcterms:issued "Y3"
<http://www.disco.unimib.it> rdfs:label "University of Milano-Bicocca"
```

```
<http://dblp.l3s.de/.../publications/P4> dc:creator <http://dblp.l3s.de/.../authors/A1'>
<http://dblp.l3s.de/.../authors/A1'> opus:has_affiliation <http://www.kit.edu>
<http://dblp.l3s.de/.../authors/A1'> foaf:homepage <http://www.kit.edu/A1>
<http://dblp.l3s.de/.../publications/P4> dcterms:issued "Y4"
<http://www.kit.edu/> rdfs:label "Karlsruhe Institute of Technology"
```

Fig. 1. An example of RDF triples that describe the evolution of the entity author

4 Results and Conclusions

The PhD work is now in the first year. Current work involves analysing the data quality dimensions addressed by several research efforts. The results so far are: literature study on data quality methodologies by a comparative description, create a framework to be able to provide a set of dimensions and their respective measures for helping the consumer or the publisher on evaluating the quality of their Linked Data. Related to the second contribution we firstly considered a comparison of interlinking approaches. The idea was to verify if there exist a full cover of all steps presented in the traditional record linkage activity. The approach we considered was the following: (i) define the main steps present in the record linkage (ii) evaluate the current works based on those steps (iii) define the problems and future works.

In Figure 2 we can see the works[2] considered in this comparison which operate principally on the instance matching level. The basic idea is that only one or two of them cover all the steps used in the record linkage task. A key aspect has been underestimated so far in the research in the linking task, in particular the Search Space Reduction step. This step has been deeply investigated only in two works. Therefore, in the Search Space Reduction there are no many methods proposed to reduce efficiently the number of entity comparisons. Finally, we can conclude that a lot of works need to be done to improve the quality of linking techniques

[2] http://dl.dropbox.com/u/2500530/Linked%20Data%20vs%20Record%20Linkage.pdf

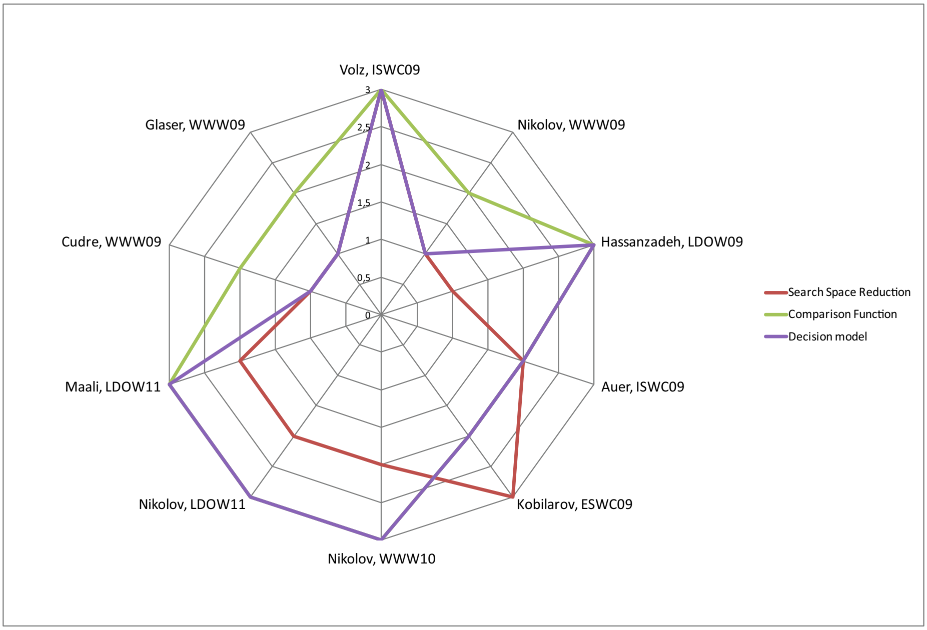


Fig. 2. A comparison of linking approaches in the Linked Data

in the Linked Data. With the perspective of improving the interlinking between data sets we concentrate on linking with temporal information such that to capture the effect of elapsed time on entity value evolution. We define a use case for applying the approach explained earlier. We consider that this work will be very beneficial for the above reasons.

In our future work, we will focus on the effectiveness of the assessment measures defined in the framework based on their nature (subjective or objective). We especially would be interested to investigate the role of the temporal linking technique in the overall framework, e.g., how some of the dimensions defined in the framework will be processed by the temporal linking technique to achieve better results.

Acknowledgements. I would like to thank my supervisor Andrea Maurino and my tutor Prof. Carlo Batini for their suggestions, and the anonymous reviewers for their helpful comments.

References

1. Quality criteria for linked data sources (2010), `http://sourceforge.net/apps/mediawiki/trdf/index.php?title=quality-criteria-for-linked-data-sources`

2. Batini, C., Cappiello, C., Francalanci, C., Maurino, A.: Methodologies for data quality assessment and improvement. Proceedings of the ACM Comput. Surv., 16:1–16:52 (2009)
3. Bizer, C., Cyganiak, R.: Quality-driven information filtering using the wiqa policy framework. Web Semantics (2009)
4. Bizer, C., Heath, T., Berners-Lee, T.: Linked Data - The Story So Far. IJSWIS 5(3), 1–22 (2009)
5. Elmagarmid, A., Ipeirotis, P., Verykios, V.: Duplicate record detection: A survey. IEEE Transactions on Knowledge and Data Engineering, 1–16 (2007)
6. Euzenat, J., Shvaiko, P.: Ontology Matching. Springer-Verlag New York, Inc. (2007)
7. Fellegi, I., Sunter, A.: A theory for record linkage. Journal of the American Statistical Association, 1183–1210 (1969)
8. Gutierrez, C., Hurtado, C.A., Vaisman, A.: Introducing time into rdf. IEEE Trans. on Knowl. and Data Eng., 207–218 (2007)
9. Halpin, H., Hayes, P.J., McCusker, J.P., McGuinness, D.L., Thompson, H.S.: When owl:sameAs isn't the same: An analysis of identity in linked data. In: Patel-Schneider, P.F., Pan, Y., Hitzler, P., Mika, P., Zhang, L., Pan, J.Z., Horrocks, I., Glimm, B. (eds.) ISWC 2010, Part I. LNCS, vol. 6496, pp. 305–320. Springer, Heidelberg (2010)
10. Hartig, O., Zhao, J.: Using web data provenance for quality assessment. In: Proceedings of the International Workshop on Semantic Web and Provenance Management, Washington DC, USA (2009)
11. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool (2011)
12. Hobbs, J.R., Pan, F.: An ontology of time for the semantic web, pp. 66–85 (2004)
13. Hogan, A., Harth, A., Passant, A., Decker, S., Polleres, A.: Weaving the pedantic web. In: International Workshop on LDOW at WWW (2010)
14. Hoxha, J., Rula, A., Ell, B.: Towards green linked data. Submission on the Semantic Web-ISWC (2011)
15. Klein, M., Fensel, D.: Ontology versioning on the semantic web, pp. 75–91. Stanford University (2001)
16. Li, P., Dong, X., Maurino, A., Srivastava, D.: Linking temporal records. In: Proceedings of the VLDB Endowment, vol. 4 (2011)
17. Ozsoyoglu, G., Snodgrass, R.T.: Temporal and real-time databases: A survey. IEEE Trans. on Knowl. and Data Eng., 513–532 (1995)
18. Pernici, B., Scannapieco, M.: Data Quality in Web Information Systems. In: Spaccapietra, S., March, S., Aberer, K. (eds.) Journal on Data Semantics I. LNCS, vol. 2800, pp. 48–68. Springer, Heidelberg (2003)
19. Tappolet, J., Bernstein, A.: Applied Temporal RDF: Efficient Temporal Querying of RDF Data with SPARQL. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 308–322. Springer, Heidelberg (2009)
20. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Discovering and Maintaining Links on the Web of Data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 650–665. Springer, Heidelberg (2009)
21. Wang, R.Y.: A product perspective on total data quality management. Commun. ACM, 58–65 (1998)
22. Wang, Y., Zhu, M., Qu, L., Spaniol, M., Weikum, G.: Timely yago: harvesting, querying, and visualizing temporal knowledge from wikipedia. In: Proceedings of the International Conference on Extending Database Technology. ACM (2010)