A Quality Model for Linked Data Exploration

Cinzia Cappiello[1], Tommaso Di Noia[2], Bogdan Alexandru Marcu[1], and Maristella Matera[1]

[1] Politecnico di Milano - DEIB
P.zza Leonardo da Vinci, 32 - 20133 Milano - Italy
name.surname@polimi.it
[2] Politecnico di Bari - DEI
via E. Orabona, 4 - 70125 Bari - Italy
name.surname@poliba.it

Abstract Linked (Open) Data (LD) offer the great opportunity to interconnect and share large amounts of data on a global scale, creating added value compared to data published via pure HTML. However, this enormous potential is not completely accessible. In fact, LD datasets are often affected by errors, inconsistencies, missing values and other quality issues that may lower their usage. Users are often not aware of the quality and characteristics of the LD datasets that they use for various and diverse tasks; thus they are not conscious of the effects that poor quality datasets may have on the results of their analyses. In this paper we present our initial results aimed to unleash LD usefulness, by providing a set of quality dimensions able to drive the selection and evaluation of LD sources. As a proof of concepts, we applied our model for assessing the quality of two LD datasets.

Keywords: Linked Data (LD), Data Quality, Quality Models for LD.

1 Introduction

In the last decade we have been facing with the transition of the Web towards a gigantic distributed knowledge base. Nowadays, we share knowledge in the Web by publishing documents, blog posts, reports as well as by deploying services and APIs. At the same time, all this information can be integrated to create new services able to aggregate data and to transform them in new knowledge.

Among the different obstacles to the development of these new breed of services we may identify two main elements: *i)* information/data are exposed in different formats and made available with reference to diverse data models; *ii)* it is hard to estimate the quality of a data source. As for the first point, the Linking Open Data initiative has given a big boost in the direction of a unified model to publish data on the Web. In fact, the hyperlinked structure of a Linked Data (LD) dataset looks the perfect nominee to distribute and connect data and resources by exploiting well known and established technologies such as HTTP. Nevertheless, since they are conceived to publish and distribute data in an open environment (as the Web is), LD datasets are affected by all the related issues:

incompleteness of the information, inconsistency and, more generally, unknown quality. In order to develop a Web application, the selection of the right data source may impact on the effectiveness of the application itself. The same is if one wants to perform data analysis. Thus, understanding the quality of a LD dataset may be a key factor in the selection of a data source among all the ones publicly available on the Web.

In this paper, we frame the problem of characterizing the quality of LD sources by providing a set of quality dimensions able to drive the evaluation and the selection of Linked Data sources. The literature extensively discusses in general the quality of data; however, there is a lack of contributions focusing on the peculiar nature of LD. This paper is an initial attempt to fill this gap and proposes a quality model that leverages traditional data quality dimensions but extends and specializes them in order to cover the peculiarities of Linked Data. After illustrating the main motivations that led us to define a new data model and analyzing the few contributions so far proposed in the literature (Section 2), we introduce our quality model for LD (Section 3). Then we show how it has been exploited for the evaluation of two small datasets gathered from two Linked Data sources, DBpedia and LinkedMDB (Section 4). In Section 5 we draw our conclusions and outline our future work.

2 Rationale and Background

The current Web has moved from a distributed base of interlinked documents to a distributed base of data. In the last years we have been confronting with the publication of data on the Web in diverse and heterogeneous formats, spanning from simple CSVs up to HTML tables. More recently, various data providers have also started to allow direct access to their datasets via Web APIs.

Thanks to the development of the Semantic Web technological stack and to the publication of the LD principles, we are also witnessing the ever increasing availability of a new breed of datasets on the Web. Linked (Open) Data relate to a set of best practices for providing an infrastructure that facilitates, encourages and promotes sharing data across the Web. The result is a distributed model that allows any data provider to publish data and also link them with other information sources on the Web.

The LD model was conceived in 2006 by Tim Berners-Lee as a set of rules and common practices on how to publish data on the Web in a semantics-aware, machine readable way. Since then, Linked Data sources have proliferated; their large volume and especially their interconnected nature have motivated researchers to define methods and tools to access and aggregate data from different LD sources [3,7,11,8].

2.1 Linked Data technology

LD provide a generic and flexible publishing paradigm, which makes it easier for data consumers to discover and integrate data from a large number of sources.

Indeed, by using URIs as global identifiers for resources, it is straightforward to set hyperlinks between entities in different datasets. As a result, in the current Web we find many Linked Data datasets connected with each other in a single global data space thus creating the so called Linked Open Data (LOD) cloud which somehow overcomes the incompleteness that often characterizes single data sources.

LD technology is characterized by the following elements:

- *A unifying data model.* Linked Data rely on RDF as a single, unifying data model. As it bases on the notion of triple, RDF is a perfect candidate to mimic the graph-based nature of the current Web thus allowing a smooth transition towards an interconnected network of entities.
- *Hyperlink-based data discovery.* By using URIs as global identifiers for entities, Linked Data allow to navigate and explore data spaces as well as to connect them with each other in a single global data space (the LOD cloud);
- *Self-descriptive data.* Linked Data ease the integration of data from different sources by relying on shared vocabularies and OWL ontologies.
- *A powerful query language.* Following their inspiring principles, LD are accessible through SPARQL endpoints. SPARQL is the standard query language to get data from a dataset. Its syntax is based on graph-pattern matching thus making it a perfect tool to retrieve data from RDF repositories.

In general, semantic applications based on LOD exploit the two major properties of the LOD architecture: the first one is the standardized data representation and access, as LD rely on a standardized data model (RDF); the second property is the openness of the Web of Data which enables the discovery of new data at runtime, in line with the LOD principles. The knowledge graph, represented through a set of RDF triples, is an ideal candidate for exploratory tasks whose aim can be manyfold. Just to cite the most frequent ones:

- Browsing a linked data space will ease the discovery of serendipitous entities;
- Via the exploration of a dataset starting from a seed node, a system can learn which are the top-k most similar or related entities thus allowing the user to perform approximate retrieval tasks;
- In case of an encyclopedic dataset, such as DBpedia, Freebase or Wikidata, one might partition the knowledge graph by considering reachability criteria.

Moreover, the interlinked nature of LOD dataset makes it possible to perform federated queries among different datasets. By leveraging the links between datasets, the user is allowed to integrate on the fly data that are retrieved from diverse knowledge sources.

All this exploration and integration tasks aim to generate new knowledge through the interlinking of already existing but distributed data. However, the usefulness of the resulting knowledge depends very much on the quality of each single LD dataset. This aspect can be pursued if specific quality models able to capture the novel aspects introduced by LD are considered.

2.2 LD quality

Despite the potential of LD and the interest they received by the scientific community, recent studies have shown that the majority of Linked Data datasets suffer from data quality problems [9,14] that still limit their adoption in real applications. In the last years some models and related methodologies have been proposed to assess the quality of data (a comprehensive survey is reported in [2]). Some papers concentrate on Web data sources [5,1,6,4]. However, very few contributions have specifically addressed the LD quality. Among them, in [11] the authors present a framework for assessing the Linked Data quality with focus on the fusion of LD datasets. They start from the definition of quality as "fitness for use" [10], and then pose emphasis on the importance of the task for which Linked Data are used. Thus, they present a framework supporting the specification and the automatic assessment of quality metrics for the specific task of data fusion. The paper, however, does not provide a quality model, while it concentrates on methods and technologies to support the automatic assessment of "any" quality metric within a framework for LD access. Similarly to the previous work, in [12] the authors propose a methodology for Linked Data quality assessment, outlining detailed phases and activities, but they do not clarify which quality dimensions are worth to be considering when LD have to be exploited.

A notable contribution towards the definition of a comprehensive quality model for LD is reported in [14]. This work exhaustively discusses the quality of Linked Data and motivates the need of methods for assessing it. It considers any single dimension of traditional data quality and tries to reformulate it for application to LD usage. The resulting model is comprehensive and valuable, but it can be cumbersome to apply, especially if one's goal is to assess Linked Data quality through the automatic computation of metrics. As explained in the following section, we adopt a minimalist approach, trying as much as possible to *i)* specialize only the most relevant traditional quality dimensions to the LD nature, *ii)* identify new dimensions (e.g., navigability) that are specific for exploratory tasks that characterize Linked Data and *iii)* define measures that can be objectively assessed through automatic methods.

3 Quality Model

As discussed in Section 2.1, LD rely on the RDF model, which provides mechanisms for describing groups of related *resources* and the relationships among them. More formally, the model is based on the notion of triple as "`subject-predicate-object`" expressions where the `subject` and the `object` are resources and the `predicate` defines the *properties* for the subject of the statement. *Nodes* are sets of subjects and objects. The subject is a URI or a blank node, the predicate is a URI and the object is a URI, a literal or a blank node. Literals and blank nodes are known as *RDF terms*.

In order to support the users in selecting the most suitable LD datasets for their tasks, we propose to annotate the different sources with a set of quality metadata that can make the users aware of their quality level. We identify a minimal set of dimensions, and related metrics, built by considering and redefining some dimensions gathered from the literature (i.e., [14]), and adding some new concepts. The resulting quality model is composed of five dimensions: *Amount of Data*, *Conciseness*, *Completeness*, *Navigability* and *Interlinking*. Amount of data, Conciseness and Completeness focus on the richness and redundancy of the LD dataset, while Navigability and Interlinking provide a measure for the possibility to explore respectively the considered dataset or different data sources starting from a specific resource. A thorough description of these dimensions is provided in the following.

Amount of data. It refers to "the extent to which the quantity or volume of data is appropriate for a particular task" [13]. When using LD to feed an application, it is important to have an idea of the richness of the dataset we are using. A simple metric that can help assess such a dimension is related to the average number of properties that characterize an entity in the considered data source.

Conciseness. In data quality literature, it is defined as "the extent to which data are compactly represented without being overwhelming" [13]. For the quality of LD datasets, we restrict the focus on the *intensional conciseness*, which refers to the degree to which a resource is characterized by a set of properties that is free of redundancy. A metric to measure such dimension for a LD entity is given by the ratio of the number of non-redundant properties and the total number of properties. Conciseness for a dataset that includes more resources is calculated by evaluating the average of the different conciseness measures.

Completeness. It refers to the degree to which all required information is present in a particular dataset. In general, completeness is "the extent to which data are of sufficient depth, breadth and scope for the task at hand" [13]. We measure LD completeness as the ratio of the number of retrieved properties of an entity and the number of required properties for a specific resource. The required properties are the ones included in an ideal list of properties that we consider as complete.

Navigability. It refers to the degree to which the different resources in the same dataset are linked. In particular, it highlights how much it is possible to navigate the dataset starting from a specific node. For assessing navigability, we start from a specific resource and all the triples in which it is involved as subject, and we then consider the percentage of objects that are URIs with respect to the total number of objects.

Interlinking. It refers to the degree to which resources in the dataset are linked with the same resource of an external dataset. For the assessment of this dimension, the statements with the predicate `owl:sameAs` are considered as it states that a resource of the graph is the same resource of another dataset. The metric to evaluate the Interlinking dimension of a whole dataset is the total number of `owl:sameAs` links.

Note that Conciseness, Completeness and Navigability metrics are defined as percentages; therefore their values belong to the interval [0,100]. Amount of data and Interlinking have instead values in $\mathbb{N}$.

4 Quality-aware source selection: an example

In order to start testing the usefulness and effectiveness of the proposed model, we performed some preliminary experiments to evaluate the quality of two datasets gathered from two important LD sources:

- *DBpedia*[3], a dataset containing data extracted from Wikipedia,
- *LinkedMDB*[4], a dataset containing movie-related content.

DBpedia is a project aimed to make Wikipedia data available on the Web in a structured way. It allows one to query Wikipedia data and to integrate them with other Web sources. The English version of DBpedia contains the description of about 4.6 million of resources and most of them (i.e., about 4.2 million) are classified on the basis of the DBpedia ontology. DBpedia is one of the largest LD datasets available; it has the big advantage to cover many knowledge domains and to be always updated, since it evolves as Wikipedia changes.

LinkedMDB publishes the first open semantic Web database for movies, including a large number of interlinks to several datasets on the open data cloud and references to related webpages.

For our experiments, we assume to be interested in developing an application for searching and exploring data about movies. We can rely on DBpedia or on a more specialized source as LinkedMDB. In this section, we show how the quality assessment, according to the model presented above, can be a good driver in the source selection and thus it can support users in understanding which is the source to prefer for their tasks.

Every resource in DBpedia is accessible through the URI pattern: `http://dbpedia.org/resource/<name_of_resource>`. Analogously, movies in LinkedMDB can be accessed via the URI `http://data.linkedmdb.org/directory/film/<movie_id>`. Before thoroughly analyzing the two sources, we just compared two descriptions of the same movie. We considered the movie “Minority Report” and compared information published at `http://dbpedia.org/page/Minority_Report_(film)` and `http://data.linkedmdb.org/page/film/333`. We noticed that LinkedMDB contains all the important data (e.g., *director*, *language*, *editor*, *music*, *producers*, *writers* , etc.) and also contains the actor list, which is instead missing in DBpedia. As regards conciseness, LinkedMDB does not contain any redundant value while in DBpedia there are four properties that

[3] http://wiki.dbpedia.org/
[4] http://www.linkedmdb.org/

Table 1. Quality metrics for DBpedia and LinkedMDB datasets

Quality Metric	DBpedia	LinkedMDB
Amount of data	51.21	39.22
Conciseness	96.07%	100%
Completeness	68.11%	72.85%
Navigability	83.66%	61.40%
Interlinking	20	0.74

are listed twice (i.e, budget, director, editing and gross). However, LinkedMDB is characterized by a lower interlinking value since it provides the external links only to DBpedia, Freebase, IMDB and RottenTomatoes pages of the movie. In DBpedia, Minority Report has instead 15 `owl:sameAs` links (out of which 11 refer to localized versions of DBpedia in different languages). This first analysis highlights that LinkedMDB has a richer content while DBpedia is less specialized but it has a good interlinking with the LOD cloud.

Considering the subgraph related to movies, we calculated the metrics described in Section 3. The results are shown in Table 1. Note that for assessing the completeness dimension we identified a set of mandatory properties, which are the most popular ones used in Wikipedia Web pages to describe a movie, that are: `starring`, `writer`, `director`, `producer`, `music composer`, `distributor`, `language`, `cinematography`, `editing`, `country`, `realase date`. All these properties are modeled in both data sets but cinematography.

Looking at Table 1, we can notice that DBpedia features a higher amount of data, navigability and interlinking, while LinkedMDB is preferable for completeness and conciseness. This might depend on the fact that LinkedMDB is a specialized data set and thus its content on movies is more accurate than in DBpedia. This in turn highlights that the selection of sources might depend on the target application domain. Therefore, LinkedMDB has to be preferred if high content quality on movies is required, while DBpedia has the capability to provide in general more information and to enable navigation along multiple and multi-language sources.

5 Conclusions

This paper has presented some preliminary results on the definition of a quality model for LD. Our model tries to capture the peculiar nature of Linked Data, considering those dimensions that are more significant for tasks typical of a LD-based application, and with a specific focus on metrics that can foster automatic quality assessment. Some preliminary experiments demonstrated that the model can be effectively used to evaluate LD datasets. Future work will be devoted to refine the model, through a systematic identification of possible tasks on Linked Data and the definition of corresponding dimensions and metrics. Model validation will be also conducted on larger datasets.

Acknowledgments

We are grateful to the students that helped us validate the model by developing tools to download and analyze the DBpedia and LinkedMDB datasets.

References

1. D. Barbagallo, C. Cappiello, C. Francalanci, and M. Matera. Reputation-based selection of information sources. In *Proc. of ICEIS 2010*, 2010.
2. C. Batini, C. Cappiello, C. Francalanci, and A. Maurino. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, 41(3), 2009.
3. D. Bianchini, V. D. Antonellis, and M. Melchiori. A linked data perspective for effective exploration of web apis repositories. In F. Daniel, P. Dolog, and Q. Li, editors, *Web Engineering - 13th International Conference, ICWE 2013, Aalborg, Denmark, July 8-12, 2013. Proceedings*, volume 7977 of *Lecture Notes in Computer Science*, pages 506–509. Springer, 2013.
4. C. Bizer and R. Cyganiak. Quality-driven information filtering using the {WIQA} policy framework. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(1):1 – 10, 2009. The Semantic Web and Policy.
5. C. Cappiello, F. Daniel, and M. Matera. A quality model for mashup components. In M. Gaedke, M. Grossniklaus, and O. Díaz, editors, *ICWE*, volume 5648 of *Lecture Notes in Computer Science*, pages 236–250. Springer, 2009.
6. C. Cappiello, F. Daniel, M. Matera, and C. Pautasso. Information quality in mashups. *IEEE Internet Computing*, 14(4):14–22, 2010.
7. G. Desolda. Enhancing workspace composition by exploiting linked open data as a polymorphic data source. In E. Damiani, R. J. Howlett, L. C. Jain, L. Gallo, and G. De Pietro, editors, *Intelligent Interactive Multimedia Systems and Services*, volume 40 of *Smart Innovation, Systems and Technologies*, pages 97–108. Springer International Publishing, 2015.
8. T. Di Noia, V. C. Ostuni, J. Rosati, P. Tomeo, E. Di Sciascio, R. Mirizzi, and C. Bartolini. Building a relatedness graph from linked open data: A case study in the IT domain. *Expert Syst. Appl.*, 44:354–366, 2016.
9. A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of linked data conformance. *J. Web Sem.*, 14:14–44, 2012.
10. J. M. Juran. *The Quality Control Handbook*. McGraw-Hill, 1974.
11. P. N. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. In D. Srivastava and I. Ari, editors, *Proceedings of the 2012 Joint EDBT/ICDT Workshops, Berlin, Germany, March 30, 2012*, pages 116–123. ACM, 2012.
12. A. Rula and A. Zaveri. Methodology for assessment of linked data quality. In M. Knuth, D. Kontokostas, and H. Sack, editors, *Proceedings of the 1st Workshop on Linked Data Quality co-located with 10th International Conference on Semantic Systems, LDQ@SEMANTiCS 2014, Leipzig, Germany, September 2nd, 2014.*, volume 1215 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2014.
13. R. Y. Wang and D. M. Strong. Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33, Mar. 1996.
14. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2016.