Methodology for Assessment of Linked Data Quality

Anisa Rula
University of Milano-Bicocca,
Department of Computer Science, Systems and Communication (DISCo)
Viale Sarca 336, Milan, Italy
anisa.rula@disco.unimib.it

Amrapali Zaveri
University of Leipzig,
Institute of Computer Science, AKSW Group
Augustusplatz 10, D-04009 Leipzig, Germany
zaveri@informatik.uni-leipzig.de

ABSTRACT

With the expansion in the amount of data being produced as Linked Data (LD), the opportunity to build use cases has also increased. However, a crippling problem to the reliability of these use cases is the underlying poor data quality. Moreover, the ability to assess the quality of the consumed LD, based on the satisfaction of the consumers' quality requirements, significantly influences usability of such data for a given use case. In this paper, we propose a data quality assessment methodology specifically designed for LD. This methodology consists of three phases and six steps with specific emphasis on considering a use case.

Keywords

data quality, linked data, assessment, improvement

1. INTRODUCTION

Recently, Linked Data (LD) has contributed a sea of information to the Web all represented in structured formats, linked with one another and made publicly available [4]. This information belongs to an enormous number of datasets covering various domains such as life sciences, geographic data, or governmental[1]. Publication of this information as Linked Data has enabled users in aggregating data from different sources to build mashups that assist in discovering new valuable information. However, recent studies have shown that majority of these datasets suffer from several data quality problems such as representational, inconsistency or interoperability issues [5]. These problems significantly hinder the uptake of these datasets in particular use cases and affect the results as the poor quality is propagated in the aggregated datasets. The ability to assess the quality of the consumed LD, based on the satisfaction of the consumers' quality requirements, significantly influences usability of such data for any given use case.

[1] `http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.html`

LDQ 2014, 1st Workshop on Linked Data Quality Sept. 2, 2014, Leipzig, Germany.

Data quality is usually defined as *fitness for use* [6] and is comprised of several data quality dimensions (e.g. completeness, accuracy, conciseness etc.) along with their respective metrics (means to measure the dimension). There have been several methodologies, which have been proposed to assess the quality of a dataset [2, 8, 10]. Even though these methodologies provide useful ways to assess the quality of a dataset, they often do not address a particular use case (usually involving several datasets) and demand a considerable amount of user involvement and expertise. Also, most of the output is not interpretable by humans and the methodologies are bound to one particular dataset and its characteristics.

Therefore, in this paper, we propose a data quality assessment methodology comprising of three phases and six steps (section 2). In contrast to the previously introduced methodologies, our methodology aims to bring an overview of the entire assessment methodology right from identifying the problems to fixing them. We discuss related work in section 3 and provide directions to future work in section 4.

2. DATA QUALITY ASSESSMENT METHODOLOGY

A data quality assessment methodology is defined as the process of evaluating if a piece of data meets the information consumers need in a specific use case [2]. In a comprehensive survey [12], it was observed that in the 30 identified approaches, there were no standardized set of steps that were followed to assess the quality of a dataset. Inspired from the methodology proposed in [1] and the lack of a standardized methodology in LD, we propose a methodology consisting of three phases and nine steps. In particular, from each of the 30 approaches, we extracted the common steps that were proposed to assess the quality of a dataset. We then adapted and revised these steps to propose a data quality assessment methodology particularly for LD as depicted in Figure 1.

Our methodology thus consists of the following phases and steps:

1. Phase I: Requirements Analysis
   (a) Step I: Use Case Analysis
2. Phase II: Quality Assessment
   (a) Step II: Identification of quality issues

(b) Step III: Statistical and Low-level Analysis

(c) Step IV: Advanced Analysis

3. Phase III: Quality Improvement

(a) Step V: Root Cause Analysis

(b) Step VI: Fixing Quality Problems

The following sections describe each of the steps in detail along with the list of data quality dimensions (from the 18 dimensions identified in [12] that are applicable for each step.

2.1 Phase I: Requirements analysis

The multi-dimensional nature of data quality makes it dependent on a number of factors that can be determined by analyzing the users requirements. Thus, the use case in question is highly important when assessing the quality of a dataset. This *requirement analysis* phase thus includes the gathering of requirements and subsequent analysis of the requirements based on the use case.

2.1.1 Step I: Use Case analysis

In this step, the user provides the details of a use case or an application that best describes the usage of the dataset in order to provide a tailored quality assessment process. For this step, we identify two types of users: (a) those who are already consumers of the dataset and thus provide their data quality experiences through use cases and (b) those who are potential consumers of the dataset and thus cannot provide such experiences. The first kinds of users already know what data quality problems they faced or are prone to face. In this case, the user guides the assessment process since they know the dataset problems before hand; in the second case the assessment process guides the user. However, both users are exploring the *fitness for use* of their dataset. This step facilitates the choice regarding not only which dataset should be assessed first, but also which aspects of individual dataset should be the initial target.

2.2 Phase II: Data Quality Assessment

In the previous phase, we identified the user requirements for her dataset with the particular use case she has in mind. This second phase involves the actual quality assessment based on the requirements. In particular, amongst the set of dimensions and metrics discussed in [12], the most relevant ones are selected. Thereafter, a quantitative evaluation of the quality of the dataset is performed using the metrics specific for each selected dimension. Thus, this phase consists of three steps: (II) Identification of quality issues (III) Statistical and Low-level analysis and (IV) Advanced analysis.

2.2.1 Step II: Identification of quality issues

The goal of this step is to identify a set of the most relevant data quality issues based on the use case. This identification is done with the help of a checklist, which can be filled by the user. The questions in the checklist implicitly refer to quality problems and their related quality dimensions. For example, questions such as whether the datasets provides a message board or a mailing list (pointing to the *understandability* dimension) or whether the data is provided in

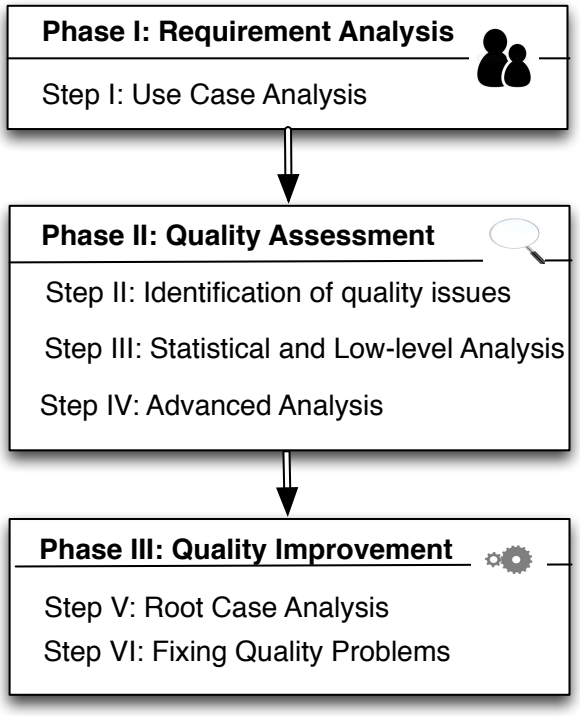

Figure 1: The quality assessment methodology

different serialization formats or languages (pointing to the *versatility* dimension), are presented to the user. In this step, the user involvement is entirely manual where the user must have knowledge about the details of the dataset to answer these questions. The output of this step is the result of the evaluation of the boolean dimensions, that is, a sum of 0's(no) or 1's(yes) which adds to the final data quality assessment score. Using this information, it is then possible to determine a set of relevant dimensions.

2.2.2 Step III: Statistical and Low-level Analysis

This step performs basic statistical and low-level analysis on the dataset. That is, generic statistics that can be calculated automatically are included in this step. For example, the number of blank nodes pointing towards the *completeness* of the dataset or number of interlinks between datasets showcasing the *interlinking* degree of the dataset are calculated. After the analysis, generic statistics on the dataset based on certain pre-defined heuristics are calculated and provided to the user. The end result is a score indicating the value for each of the metrics assessed.

2.2.3 Step IV: Advanced Analysis

This step, in combination with steps II and III, is used for assessing the overall quality of the dataset. The assessment can be performed in different ways for different quality dimensions. For example, in order to assess the *accuracy* of data values, a pattern-based approach can be applied, which generates data quality tests of RDF knowledge bases [7]. These patterns will capture incorrect values such as postal address, phone number, email address, personal identifica-

tion number, etc.

This step is performed by comparing values from the transformed dataset to the gold standard values (i.e. values from the original source) or to a dataset in the same domain. For example, in case of measuring the population completeness of a dataset, it needs to be compared with the original dataset. Thus, this step requires the target or derived dataset as well as the original or source dataset as input. The output of this step are (i) evaluation results performed between target and original datasets or those in the same domain and (ii) an aggregated value (score) of the results.

The data quality score metrics are based on simple ratio calculation. The ratio is measured by subtracting the ratio between the total number of instances that violate a data quality rule (V) and the total number of relevant instances (T) from one, as the following formula shows:

$$DQ_{score} = 1 - (V/T) \quad (1)$$

This score can be applied for each property of the dataset. In case we want to calculate the quality of the overall properties/attributes in a dataset, the above DQ_{score} is multiplied with a weight w_i representing the importance of the intended task for each property in the dataset and divide the sum of the weighted DQ_{score} by the sum of all weighting factors of the regarded properties (W).

$$DQ_{weightedscore} = \sum_{i=1}^{n} (DQ_{score} * w_i)/W \quad (2)$$

In case of equal importance of the properties for the task at hand or in case it is not possible to annotate importance values, all w_i are considered equal to 1 and the W value is gives the number of all properties that are tested in the dataset. While in the former case, the $DQ_{weightedscore}$ is a contextual metric in the latter case it is considered to be an intrinsic metric.

At the end of this phase, the total score from Steps II to IV are aggregated and provided as a result to the user indicating the quality of the dataset. A breakdown of the scores for each of the metrics assessed is provided so that the user is able to look at each metric separately. Additionally, explanations of how the assessment was performed i.e. details of the metrics are available to the user so that she is able to interpret the results in a meaningful way.

2.3 Phase III: Quality Improvement

This phase focuses towards improving the quality of the datasets based on the analysis performed in Phase II focusing on the use case identified in Phase I. This phase consists of two steps: (VI) Root Cause Analysis and (VII) Fixing Quality Problems.

2.3.1 Step V: Root Cause Analysis

In this step, the main aim is to find an explanation for the cause of the detected data quality issues i.e. performing *root cause analysis*. This step helps the user interpret and understand the results of the data quality assessment that is performed on her dataset. Moreover, this step is important as the decision of whether to trust the assessment results depends highly on the precise understanding of the evaluation of the data quality. Essentially, this step involves:

- detecting whether the problem occurs in the original dataset
- in case the original dataset is not available, analyze the dataset to detect the cause

For example, if the data quality assessment reports problem of inconsistency in the dataset, the data modeling should be checked or if the problem of completeness is reported, the values in the original dataset and target dataset should be compared to find the cause.

2.3.2 Step VI: Fixing Quality Problems

In this step, strategies to address the identified root cause of the problems are implemented. There are several strategies that can be implemented in this step such as:

- Semi-automatic or automated approaches
- Crowdsourcing mechanisms

Semi-automated or automated approaches can help detect quality issues and their causes on a large-scale. For example, inconsistencies in the ontology can be detected by running a reasoner on the entire ontology. Crowdsourcing, on the other hand, is highly appropriate for any assignment involving large to huge numbers of small tasks requiring human judgment. In terms of LD, crowdsourcing quality assessment may involve, for example, verifying the completeness or correctness of a fact wrt. the original dataset. Such a task does not require underlying knowledge about the structure of the data and can be done in a time and cost effective manner [11].

3. RELATED WORK

A number of data quality assessment methodologies and tools have been introduced, those particularly focusing on LD. These methodologies can be broadly classified into three categories: (i) automated, (ii) semi-automated and (iii) manual. There exist data quality assessment tools, which work completely automatically, such as LinkQA[2], which is designed to assess the quality of links in an automated way and LODStats[3], which gathers comprehensive statistics (no. of classes, properties, links etc.) about a dataset available as RDF. On the other hand, there are generic tools for validating the structure of the RDF document[4], which only provide a high-level analysis of the quality in terms of representational (or modeling) problems. Tools, which semi-automatically assess data quality, include Flemming's data quality assessment tool [3]; LODRefine[5]; DL-Learner[6] [8]

[2] https://github.com/cgueret/LinkedData-QA
[3] http://stats.lod2.eu/
[4] http://swse.deri.org/RDFAlerts/, http://www.w3.org/RDF/Validator/
[5] http://code.zemanta.com/sparkica/
[6] http://dl-learner.org

and ORE (Ontology Repair and Enrichment)[7] [9]. Tools, which entail manual assessment, are Sieve [10], which assesses the quality of data using an integration process and WIQA [2], which allows users to apply a wide range of quality-based policies to filter information.

However, the automatic tools are bound to certain datasets and do not allow the freedom to the user to choose a particular dataset nor focus on a specific use case. In case of semi-automated tools, the user needs to have adequate knowledge about the dataset in order to use this tool. However, these tools are not bound to a use case. In case of manual tools, they demand a huge amount of user involvement and expertise and are not sensitive towards the use case.

Our data quality assessment methodology is at the intersection of these tools as it not only focuses on a particular use case but also allows the user to obtain low-level as well aggregated and higher level analysis of the dataset. Moreover, the methodology supports the interpretation of the results and allows the user to retrace or, if required, even change the input metrics to obtain the desired quality for the particular use case. Furthermore, the methodology incorporates the one important component missing from the existing ones, the improvement of data quality problems once identified.

4. CONCLUSIONS AND FUTURE WORK

In this paper, we have introduced a data quality assessment methodology consisting of three phases and six steps. This methodology is generic enough to be applied to any use case. In order to validate its usability, we plan to apply it to specific use cases to assess the feasibility and effectiveness of the methodology. This validity will also help us measure its applicability in various domains. Moreover, we plan to build a tool based on this methodology so as to assist users to assess the quality of any linked dataset.

5. REFERENCES

[1] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[2] C. Bizer and R. Cyganiak. Quality-driven information filtering using the WIQA policy framework. *Web Semantics*, 7(1):1 – 10, Jan 2009.

[3] A. Flemming. Quality characteristics of linked data publishing datasources. Master's thesis, Humboldt-Universität zu Berlin, 2010.

[4] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*, chapter 2, pages 1 – 136. Number 1:1 in Synthesis Lectures on the Semantic Web: Theory and Technology. Morgan and Claypool, 1st edition, 2011.

[5] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of Linked Data conformance. *Journal of Web Semantics*, 2012.

[6] J. Juran. *The Quality Control Handbook*. McGraw-Hill, New York, 1974.

[7] `http://ore-tool.net`

[7] D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-driven evaluation of linked data quality. In *WWW*, pages 747–758, 2014.

[8] J. Lehmann. DL-Learner: Learning Concepts in Description Logics. *Journal of Machine Learning Research*, 10:2639–2642, 2009.

[9] J. Lehmann and L. Bühmann. ORE - A Tool for Repairing and Enriching Knowledge Bases. In *ISWC*, LNCS. Springer, 2010.

[10] P. Mendes, H. Mühleisen, and C. Bizer. Sieve: Linked Data Quality Assessment and Fusion. In *LWDM*, March 2012.

[11] A. Zaveri, D. Kontokostas, M. A. S. andLorenz Bühmann, M. Morsey, S. Auer, and J. Lehmann. User-driven Quality Evaluation of DBpedia. In *Proceedings of 9th International Conference on Semantic Systems, I-SEMANTICS '13, Graz, Austria, September 4-6, 2013*, pages 97–104. ACM, 2013.

[12] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality Assessment Methodologies for Linked Data: A Survey. Under review, available at http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-open-data.