International Journal on Semantic Web and Information Systems
Volume 12 • Issue 3 • July-September 2016

Towards An Objective Assessment Framework for Linked Data Quality:
Enriching Dataset Profiles with Quality Indicators

Ahmad Assaf, EURECOM, London, UK
Aline Senart, SAP Labs France, Mougins, France
Raphaël Troncy, EURECOM, London, UK

ABSTRACT

Ensuring data quality in Linked Open Data is a complex process as it consists of structured information supported by models, ontologies and vocabularies and contains queryable endpoints and links. In this paper, the authors first propose an objective assessment framework for Linked Data quality. The authors build upon previous efforts that have identified potential quality issues but focus only on objective quality indicators that can measured regardless on the underlying use case. Secondly, the authors present an extensible quality measurement tool that helps on one hand data owners to rate the quality of their datasets, and on the other hand data consumers to choose their data sources from a ranked set. The authors evaluate this tool by measuring the quality of the LOD cloud. The results demonstrate that the general state of the datasets needs attention as they mostly have low completeness, provenance, licensing and comprehensibility quality scores.

KEYWORDS

Data Quality, Dataset Profile, Linked Data, Profile Generation, Quality Framework, Semantic Web

1. INTRODUCTION

In the last few years the Semantic Web gained a momentum supported by the introduction of many related initiatives like the Linked Open Data (LOD)[1]. From 12 datasets cataloged in 2007, the Linked Open Data cloud has grown to nearly 1000 datasets containing more than 82 billion triples. Data is being published by both public and private sectors and covers a diverse set of domains from life sciences to military. This success lies in the cooperation between data publishers and consumers where users are empowered to find, share and combine information in their applications easily.

We are entering an era where open is the new default. Governments, universities, organizations and even individuals are publicly publishing huge amounts of open data. This openness should be accompanied with a certain level of trust or guarantees about the quality of data. The Linked Open Data is a gold mine for those trying to leverage external data sources in order to produce more informed business decisions (Crawford, 2011).

However, the heterogeneous nature of sources reflects directly on the data quality as these sources often contain inconsistent as well as misinterpreted and incomplete information.

Traditional data quality is a thoroughly researched field with several benchmarks and frameworks to grasp its dimensions (Kahn, 2002; Stvilia, 2007; Wang, 1996). Data quality principles typically rely on many subjective indicators that are complex to measure automatically. The quality of data

DOI: 10.4018/IJSWIS.2016070104

in indeed realized when it is used (Godfrey, 1999), thus directly relating to the ability of satisfying users' continuous needs.

Web documents that are by nature unstructured and interlinked require different quality metrics and assessment techniques than traditional datasets. For example, the importance and quality of Web documents can be subjectively calculated via algorithms like Page Rank (Page, 1999). Ensuring data quality in Linked Open Data is a complex process as it consists of structured information supported by models, ontologies and vocabularies and contains queryable endpoints and links. This makes data quality assurance a challenge. Despite the fact that Linked Open Data quality is a trending and highly demanded topic, very few efforts are currently trying to standardize, track and formalize frameworks to issue scores or certificates that will help data consumers in their integration tasks.

Data quality assessment is the process of evaluating if a piece of data meets the consumers need in a specific use case (Bizer, 2009a). The dimensionality of data quality makes it dependent on the task and users requirements. For example, DBpedia (Bizer, 2009b) and YAGO (Suchanek, 2007) are knowledge bases containing data extracted from structured and semi-structured sources. They are used in a variety of applications e.g., annotation systems (Mendes, 2011), exploratory search (Marie, 2013) and recommendation engines (Di Noia, 2012). However, their data is not integrated into critical systems e.g., life critical (medical applications) or safety critical (aviation applications) as its data quality is found to be insufficient. In this paper, we first propose a comprehensive objective framework to evaluate the quality of Linked Data sources. Secondly, we present an extensible quality measurement tool that helps on one hand data owners to rate the quality of their dataset and get some hints on possible improvements, and on the other hand data consumers to choose their data sources from a ranked set. The aim of this paper is to provide researchers and practitioners with a comprehensive understanding of the objective issues surrounding Linked Data quality.

The framework we propose is based on a refinement of the data quality principles described in Assaf (2012) and surveyed in the work of Zaveri (2013). Some attributes have been grouped for more detailed quality assessments while we have also extended them by adding for each attribute a set of objective indicators. These indicators are measures that provide users with quality metrics measurable by tools regardless of the use case. For example, when measuring the quality of DBpedia dataset, an objective metric would be the availability of human or machine readable license information rather than the trustworthiness of the publishers.

Furthermore, we surveyed the landscape of Linked Data quality tools to discover that they only cover a subset of the proposed objective quality indicators. As a result, we extend Roomba which is a framework to assess and build dataset profiles with an extensible quality measurement tool and evaluate it by measuring the quality of the LOD cloud group. The results demonstrate that the general quality of LOD cloud needs more attention as most of the datasets suffer from various quality issues.

This paper is structured as follows: Section 2 presents the related work in data quality assessment methodologies. Section 3 presents our framework with its objective quality measures and indicators. Section 4 reviews the existing tools and frameworks in the Linked Open Data quality landscape. Section 5 presents our tool for evaluating those indicators. Section 6 presents concluding remarks and identifies future work.

2. RELATED WORK

According to Zaveri (2013) a comprehensive systematic review of data quality assessment methodologies applied to LOD. They have extracted 26 quality dimensions and a total of 110 objective and subjective quality indicators. However, some of those objective indicators are dependent on the use case thus there is no clear separation on what can be automatically measured. For example, data completeness is generally a subjective dimension. However, the authors specified that the detection of the degree on which all the real-world objects are represented, detection of number of missing values for specific property and detection of the degree to which instances in the dataset are interlinked are

Pattern Based Feature Construction in Semantic Data Mining

Agnieszka awrynowicz and Jdrzej Potoniec (2014). *International Journal on Semantic Web and Information Systems (pp. 27-65).*

www.igi-global.com/article/pattern-based-feature-construction-in-semantic-data-mining/113713?camid=4v1a

A Tool Suite to Enable Web Designers, Web Application Developers and End-users to Handle Semantic Data

Mariano Rico, Óscar Corcho, José Antonio Macías and David Camacho (2010). *International Journal on Semantic Web and Information Systems (pp. 38-60).*

www.igi-global.com/article/tool-suite-enable-web-designers/47108?camid=4v1a