Linked Data Quality Assessment and its Application to Societal Progress Measurement

Abstract

In recent years, the Linked Data (LD) paradigm has emerged as a simple mechanism for employing the Web as a medium for data and knowledge integration where both documents and data are linked. Moreover, the semantics and structure of the underlying data are kept intact, making this the Semantic Web. LD essentially entails a set of best practices for publishing and connecting structure data on the Web, which allows publishing and exchanging information in an interoperable and reusable fashion. Many different communities on the Internet such as geographic, media, life sciences and government have already adopted these LD principles. This is confirmed by the dramatically growing Linked Data Web, where currently more than 50 billion facts are represented.

With the emergence of Web of Linked Data, there are several use cases, which are possible due to the rich and disparate data integrated into one global information space. Linked Data, in these cases, not only assists in building mashups by interlinking heterogeneous and dispersed data from multiple sources but also empowers the uncovering of meaningful and impactful relationships. These discoveries have paved the way for scientists to explore the existing data and uncover meaningful outcomes that they might not have been aware of previously.

In all these use cases utilizing LD, one crippling problem is the underlying *data quality*. Incomplete, inconsistent or inaccurate data affects the end results gravely, thus making them unreliable. Data quality is commonly conceived as *fitness for use*, be it for a certain application or use case. There are cases when datasets that contain quality problems, are useful for certain applications, thus depending on the use case at hand. Thus, LD consumption has to deal with the problem of getting the data into a state in which it can be exploited for real use cases. The insufficient data quality can be caused either by the LD publication process or is intrinsic to the data source itself.

A key challenge is to assess the quality of datasets published on the Web and make this quality information explicit. Assessing data quality is particularly a challenge in LD as the underlying data stems from a set of multiple, autonomous and evolving data sources. Moreover, the dynamic nature of LD makes assessing the quality crucial to measure the accuracy of representing the real-world data. On the document Web, data

quality can only be indirectly or vaguely defined, but there is a requirement for more concrete and measurable data quality metrics for LD. Such data quality metrics include correctness of facts wrt. the real-world, adequacy of semantic representation, quality of interlinks, interoperability, timeliness or consistency with regard to implicit information. Even though data quality is an important concept in LD, there are few methodologies proposed to assess the quality of these datasets.

Thus, in this thesis, we first unify 18 data quality dimensions and provide a total of 69 metrics for assessment of LD. The first methodology includes the employment of LD experts for the assessment. This assessment is performed with the help of the TripleCheckMate tool, which was developed specifically to assist LD experts for assessing the quality of a dataset, in this case DBpedia. The second methodology is a *semi-automatic* process, in which the first phase involves the detection of common quality problems by the automatic creation of an extended schema for DBpedia. The second phase involves the manual verification of the generated schema axioms. Thereafter, we employ the wisdom of the crowds i.e. workers for online crowdsourcing platforms such as Amazon Mechanical Turk (MTurk) to assess the quality of DBpedia. We then compare the two approaches (previous assessment by LD experts and assessment by MTurk workers in this study) in order to measure the feasibility of each type of the user-driven data quality assessment methodology.

Additionally, we evaluate another semi-automated methodology for LD quality assessment, which also involves human judgement. In this semi-automated methodology, selected metrics are formally defined and implemented as part of a tool, namely R2RLint. The user is not only provided the results of the assessment but also specific entities that cause the errors, which help users understand the quality issues and thus can fix them. Finally, we take into account a domain-specific use case that consumes LD and leverages on data quality. In particular, we identify four LD sources, assess their quality using the R2RLint tool and then utilize them in building the Health Economic Research (HER) Observatory. We show the advantages of this semi-automated assessment over the other types of quality assessment methodologies discussed earlier. The Observatory aims at evaluating the impact of research development on the economic and healthcare performance of each country per year. We illustrate the usefulness of LD in this use case and the importance of quality assessment for any data analysis.