Abstract

The main objective of the Linked Open Data paradigm is to crystallize knowledge through the interlinking of already existing but dispersed data.The usefulness of the developed knowledge depends strongly on the quality of the aggregated and published data. Therefore, the goal of this research is proposing a metrics-driven framework for predicting the quality of linked open datasets from an inherent point of view. To achieve this goal, we have followed an approach which is started by analysis of the well-known data quality frameworks, and comparing existing dimensions of data quality presented in these models. we tried to identify the most appropriate quality dimensions that could be applied to inherent quality characteristics of LOD datasets. These inherent quality characteristics are completeness, semantic accuracy, syntactic accuracy, uniqueness, consistency and interlinking. In order to make the characteristics quantifiable, we define a set of metrics to measure the above six inherent quality characteristics using Goal-Question-Metric(GQM) approach.

To evaluate our work, we have theoretically supported our claim by validation of the metrics and evaluation of the quality model. To put the proposed metrics into practice, we have implemented an automated tool and computed the metric values for various datasets from different domains of LOD. Furthermore, we have subjectively evaluated our proposed model using expert opinion. The proposed metrics are shown to have meaningful correlation with the quality dimensions, thus we are able to predict the inherent quality dimensions of any dataset, once it is integrated into the LOD, by only observing the values of proposed metrics. The results help publishers to filter out low-quality data, which in turn enables data consumers to make better and more informed decisions when using the shared datasets.