Should you Link(ed) Data?

Quality Assessment for Linking to Third-Party Data

Jesse Bakker[1], Wouter Beek[1], and Erwin Folmer[2]

[1] VU University Amsterdam, Netherlands
[2] University of Twente, Netherlands

1 Introduction

The key feature of Linked Data is, linking data. By assigning URIs to things, we can uniquely identify them out of the context of a single dataset. With this, we can in a way, ascent from the data silos, to a single knowledge graph. In this knowledge graph we have the ability to combine data from numerous databases. But this comes with a risk. Before associating your own data with veiled, third-party data, one should know whether it suits your quality needs. For this, the following research question was posed; *How can a quality assessment help justify the interlinking of Linked Data and foster trust?* A methodology is presented as the answer to the research question, see Figure 1. This Methodology constitutes three phases, namely a Preparation, an Assessment and an Interpretation phase. After which, an established opinion (as quality is subjective/context dependent) is formed based on an extensive quality measurement dataset. This dataset contains 1. measurements 2. metadata 3. URIs of erroneous "things".

The methodology has tested with real Linked Open Data. This uncovered both relevant statistics and otherwise secluded erroneous instances. By using generic SPARQL queries and python code, the metrics were easily measurable and could quickly be deployed. However, the standard metrics, in most use cases, had an unsatisfactory coverage. This can be solved with the custom metrics, of which the creation is costly in both time and (cross-domain) expertise.

During the preparation phase, two datasets are selected, one to be assessed and one as the context. Standard and Custom metrics are employed in order to cover as many quality facets as possible. The Assessor defines quality requirements, based on the context dataset, in SHACL, in order to automate parts of the interpretation.

A knowledge graph is iteratively constructed. When following the methodology, the assessor iteratively interprets partitions of the measurements. Prior to each interpretation step, the knowledge graph is automatically updated, such that all to-be-interpreted measurements are described in the knowledge graph. Moreover, the knowledge graph includes documentation on the metrics itself and concepts related to the notion of quality. For each interpretation step, the assessor utilises predefined SPARQL patterns to retrieve a set of measurements, related metrics and other concepts. In addition the assessor retrieves a validation report, generated with SHACL. These are foundation of the opinion on whether the datasets should be interlinked.

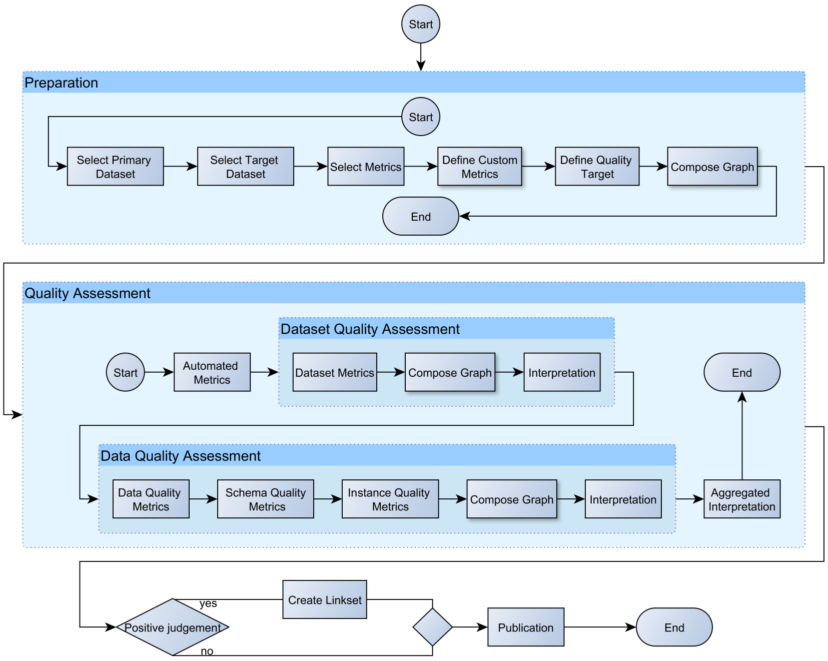

Fig. 1. Quality Assessment Methodology

SPARQL and Python are leveraged to compute the measurements. This is done semi-automatically, as metrics can sporadically require manual input. The measurement procedure is generic and can easily be reused, and extended for other use cases. Since resources have unique URIs, we can easily identify resources which negatively affect the resulting value of the metric, during the measurements and say something about them. This information is separately stored in the knowledge graph.

The knowledge graph is leveraged into an established judgement on whether one dataset should be interlinked with another, based on the quality of both the data and the dataset. Several Semantic Web standards are utilised in the process, such as SHACL, SPARQL, RDF(s), OWL and PROV-O. Regardless of whether the judgement favours the creation of a linkset (the set of triples relating the two datasets to each other), the knowledge graph should be published. This offers several benefits. First, the owner of the assessed dataset can use the knowledge graph to identify faulty resources and quality issues. Allowing the owner to amend them. Second, it provides a wealth of information to data consumers. Allowing them to inform themselves about possible uses, strengths and weaknesses of the data. Furthermore, even when no linkset resulted from the methodology, the knowledge graph can function as an admonition to inform data consumers on the risks of interlinking the two datasets.