Luzzu – A Framework for Linked Data Quality Assessment

Jeremy Debattista, Christoph Lange, Sören Auer

University of Bonn
{debattis,langec,auer}@iai.uni-bonn.de

Abstract. With the increasing adoption and growth of the Linked Open Data cloud, the variety of the web of data makes it challenging to determine the quality of the data published on the Web and to subsequently make this information explicit to data consumers. In this demo paper we describe Luzzu, a scalable Quality Assessment framework for Linked Data Apart from providing quality metadata and quality problem reports that can be used for data cleaning, Luzzu is extensible: third party metrics can be easily plugged-in the framework. Hence, the extensibility of Luzzu enables the quality assessment in light of "fitness for use".

Keywords: Data Quality, Assessment Framework, Quality Metadata, Quality Metrics

1 Introduction

With the increasing adoption of *Linked Open Data*, covering formats such as *RDFa*, *Microformats*, and *JSON-LD*, as well as initiatives such as *schema.org*, the Web is currently being complemented with a Web of Data. This enriched data shares many characteristics with the original documents, including the varying quality. Quality on the documents is usually measured indirectly using techniques such as page rank. This is because the quality of a document is only subjectively assessable, and thus an indirect measure, such as the number of links created by others to a certain web page, gives a good approximation of its quality.

In this paper we introduce Luzzu[1], a Linked Data Quality Assessment Framework. The rationale behind Luzzu is to provide an integrated platform that: (1) assesses Linked Data quality using a library of generic and user-provided domain specific quality metrics in a scalable manner; (2) provides queryable quality metadata on the assessed datasets; (3) assembles detailed quality reports on assessed datasets. For the demo we present Luzzu Web (`http://purl.org/net/luzzu`), a web user interface that enables the (a) exploration and ranking of quality assessed datasets; (b) visualisation of quality metadata; and (c) assessment of datasets, which invokes the Luzzu framework described in this paper. Within the demo video (`https://vimeo.com/132264226`) we assess the Democratic City

[1] GitHub: `https://github.com/EIS-Bonn/Luzzu`; Website: `http://eis-bonn.github.io/Luzzu/`; Luzzu Web: `http://purl.org/net/luzzu`

RDF dataset (`http://datahub.io/dataset/democratic-city`). These steps can be replicated on Luzzu Web.

2 Approach

The framework follows a three step workflow, starting with the metric initialisation process (**Step 1**). In this step, user defined metrics are compiled and initialised together with metrics implemented in Java. The quality assessment process is then commenced (**Step 2**) by sequentially streaming statements of the candidate dataset into the initialised quality metrics. Once this process is completed, the annotation process (**Step 3**) generates quality metadata and compiles a comprehensive quality report. The quality report produced in this framework enables data curators to improve the dataset's quality by using the report to identify quality issues within the dataset.

The framework comprises three layers: *Communication Layer*, *Assessment Layer* and *Knowledge Layer*. The *Communication Layer* exploits the framework's interfaces as a REST service, whilst the latter two layers are described in the remainder of this section.

2.1 Knowledge Layer

The *Knowledge Layer* is composed of three units, namely the *Semantic Schema Layer*, the *Annotation Unit*, and the *Operations Unit*. These units assist to the provision of quality metadata and assessment reports, and other operations that can be performed upon the same metadata. This layer, and subsequently Luzzu, is driven by a number of schemas that enables the representation of quality metadata (daQ), quality problem reports (QPRO) and other operational schemas to operate the framework[2]. The *Dataset Quality Ontology (daQ)* [2] is the core vocabulary, based on the RDF Data Cube vocabulary[3] and PROV-O[4], that defines how quality metadata should be represented at an abstract level. It is used to attach the results of quality benchmarking of a Linked Open Dataset to the dataset itself. These results can be used to rank (cf. Section 3) or visualise (cf. Section 4) datasets according to their quality.

2.2 Assessment Layer

The *Assessment Layer* is composed of three units, namely the *Processing Unit*, the *LQML Compilation Unit*, and the *Quality Assessment Unit*. These units handle the operations related to the quality assessment of a dataset.

The Quality Assessment Unit is the most important unit of the framework. Third parties can extend the framework by creating custom metrics by either

[2] All defined ontologies in Luzzu have the namespace `http://purl.org/eis/vocab/{prefix}`

[3] `http://www.w3.org/TR/vocab-data-cube/`

[4] `http://www.w3.org/TR/prov-o/`

implementing simple Java interfaces[5], or LQML [3], a novel quality metric language. The main advantage of LQML is that creators of quality metrics do not need to go through all the process to create a Java package, but can declaratively define a metric in a few lines of code. We are currently in the process of implementing functionality that allows more complex metrics to be implemented in LQML and not just simple pattern matching rules.

The Processing Unit controls the whole execution of the quality assessment of a chosen dataset. Luzzu implements two stream processing units; one based on the Jena RDF API and the other on the Spark processing framework. Streaming ensures scalability (since we are not limited by main memory) and parallelisability (since the parsing of a dataset can be split into several streams to be processed on different threads, cores or machines).

Jena provides the possibility of streaming triples sequentially in a separate thread implementing a producer-consumer queue. A dataset, which can be serialised in many typical RDF formats (RDF/XML, N-Triples, N-Quads etc.), is read directly from the disk storage. Luzzu can also consume data from a SPARQL endpoint[6] instead of a data dump. This way of consuming Linked Data for quality assessment is discouraged due to the fact that, if SPARQL *write* transactions (insert or update) are performed during the assessment, the whole quality assessment might be invalid. We are also in the process of enabling the assessment of datasets compressed with the HDT format. HDT [1] is a compact binary data structure for RDF, maintaining the search and browsing operations without having to decompress.

Another approach for sequential stream processing is to use the *Hadoop* MapReduce technology[7] or its in-memory equivalent *Spark*. The idea is to *map* the processing of large datasets on multiple clusters, streaming triples in the process. A simple *reduce* function then takes the results of the *map* to populate a queue that feeds the metrics processors. Metric computations are not "associative" in general, which is one of the main requirements to implement a MapReduce job; therefore, instantiated metrics are split into different threads in the master node.

3 Ranking Datasets using the Quality Metadata

Our framework enables flexible filtering and ranking in that the daQ vocabulary facilitates access to dataset quality metrics in these different dimensions and thus facilitates the (re)computation of custom aggregated metrics derived from base metrics. To keep quality metric information easily accessible, each assessed dataset should store the relevant daQ quality metadata graph in the dataset

[5] Information on how to implement metrics can be found at `http://eis-bonn.github.io/Luzzu/howto.html`.

[6] Note that the metrics are still implemented in Java and no SPARQL queries are executed on the endpoint during the assessment. This enables the assessment of metrics such as Dereferenceability that cannot be implemented using SPARQL queries.

[7] `http://hadoop.apache.org/`

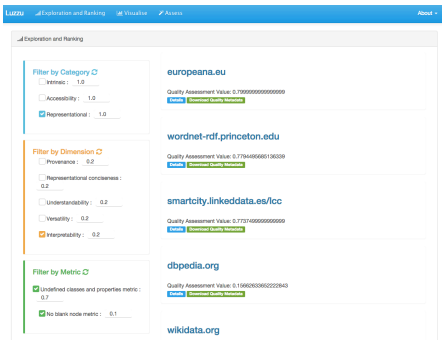

Fig. 1: The Web Interface showing the Assessment Process.

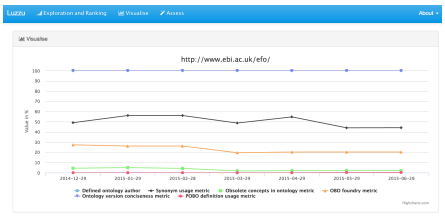

Fig. 2: Visualising a dataset over time.

itself, once it has been computed. In the spirit of "fitness for use", the Luzzu ranking algorithm enables users to define weights on their preferred categories, dimensions or metrics, that are deemed suitable for her task at hand. Figure 1 shows the ranking view of the Luzzu web application.

4 Visualising Quality Metadata

Apart from displaying ranked lists, the Luzzu web application visualises quality metadata as charts. A visualisation wizard helps the user to choose the right visualisation type and charts. Currently, the following three types can be visualised are (a) multiple datasets vs metric; (b) dataset vs Metric over time; (c) quality of dataset. Figure 2 depicts a dataset's quality evolution over time.

5 Conclusion

Data quality assessment is crucial for the wider deployment and use of Linked Data. With Luzzu we presented in this paper an approach for a scalable and easy-to-use Linked Data quality assessment framework. We see Luzzu as the first step on a long-term research agenda aiming at shedding light on the quality of data published on the Web.

References

1. Arias, M., Fernández, J.D., Martínez-Prieto, M.A., Gutiérrez, C.: Hdt-it: Storing, sharing and visualizing huge rdf datasets. In: 10th International Semantic Web Conference (ISWC 2011) (2011), `http://dataweb.infor.uva.es/wp-content/uploads/2011/10/iswc2011.pdf`
2. Debattista, J., Lange, C., Auer, S.: Representing dataset quality metadata using multi-dimensional views. pp. 92–99 (2014)
3. Debattista, J., Lange, C., Auer, S.: Luzzu quality metric language – a dsl for linked data quality assessment. CoRR (2015), `http://arxiv.org/abs/1503.05157`