DaQAR - An Ontology for the Uniform Exchange of Comparable Linked Data Quality Assessment Requirements

André Langer(✉) and Martin Gaedke

Technische Universität Chemnitz, Chemnitz, Germany
{andre.langer,martin.gaedke}@informatik.tu-chemnitz.de

Abstract. The World Wide Web represents a tremendous source of information with resources of varying data quality from almost arbitrary knowledge domains. The decision process to select the best data source for current business requirements is not trivial. In the past, research has already focused on vocabularies to represent data quality metrics and measurements (W3C's DQV) or notations to represent and validate structural requirements (W3C's SHACL). But a consistent universal semantic approach to define specific quality requirements for assessment purposes from the data consumer perspective is still missing. Therefore, we address this challenge and present DaQAR - an ontology that is capable of defining arbitrary quality requirements on both data instance, schema and service level in a uniform fashion. It can be used for data quality assessment purposes to compare multiple eligible data resources on particular metrics and attributes of current interest.

Keywords: Linked Data · Data quality · Quality requirements
Quality assessment

1 Introduction

Data Quality (DQ) is a term that describes how good data at hand fits to current business needs ("fitness for use" [12]). Data is classified as of good quality, if it conforms to all specified requirements and if it is free of defects [9]. As a consequence, the same data can be of poor quality for another use case with different requirements.

In the past, research primarily focused on reasonable standard metrics that could be used and measured as an indicator for data quality. A survey in 2014 identified 18 appropriate quality dimensions with 69 different metrics from 118 related articles [13]. Depending on the type of data source, they allow conformance measurements on data instance level, ontological schema level as well as on service level [7]. In addition to conventional metrics, other approaches made use of RDF graph structures and suggested a test-driven approach to

T. Mikkonen et al. (Eds.): ICWE 2018, LNCS 10845, pp. 234–242, 2018.
https://doi.org/10.1007/978-3-319-91662-0_18

identify data quality issues (SHACL[1], DQTP [6], SPIN [4]) for an identification of requirement violations on instance level.

In order to express quality measurement results, data quality vocabularies can be used such as W3Cs DQV[2]. The measurement results based on such a vocabulary provide information that a potential consumer of a data service can use to manually evaluate the suitability of the provided data for the target usage scenario. The information can also be used to automatically process concrete measurement results by a consecutive tool chain.

However, the unmanaged measurement of a set of available standard quality metrics is often not sufficient in practical scenarios. Instead, a predefined selection of metrics of interest depending on the current usage scenario of the data can be more valuable. Additionally, a measurement result value has also to be rated somehow, how good it fits to current expectations. The quality assessment step is neither trivial nor the calculation basis for an overall assessment score reproducible for a user in many cases. Ideally, a domain-independent vocabulary exists to define a profile with an exchangeable requirement description.

With this paper we aim to fill this gap and provide the following contribution:

- A basic vocabulary to define quality requirements of arbitrary quality concepts on one or multiple data resources from a consumer perspective
- A method to specify tolerated measurement thresholds and their impact on the quality assessment
- An approach to uniformly calculate and return an overall assessment score

The rest of the paper is structured in the following way: Sect. 2 describes in detail the challenge in defining data quality requirements among multiple data resources in a simple use case scenario. Section 3 presents our proposed solution how a semantic approach can be used to describe a specification of desired requirements, which is then verified and discussed in Sect. 4. Finally, we mention in Sect. 5 related concepts and briefly summarize our entire DaQAR approach in Sect. 6.

2 Conceptual Problem Analysis

Data is usually requested, collected or processed for a specific purpose. As a consequence, expressing the quality of an investigated dataset claims no absolute truth, but is dependent from the concrete usage scenario.

In order to assess the quality of available data sources for a certain use case, a set of requirements *REQ* of varying complexity should be definable by a data consumer as depicted in Fig. 1. These requirements could rely on well-known general metrics such as the ISO/IEC 25012 standard or other classifications [13], [2]. They could also be expressed by using generic constraint language template definitions, or locally by defining own individual concepts.

[1] https://www.w3.org/TR/shacl/.
[2] https://www.w3.org/TR/vocab-dqv/.

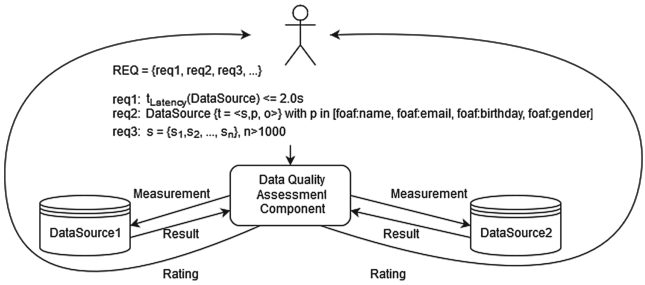

Fig. 1. Example scenario for assessing CRM customer data, specified by using the FOAF ontology, from two different data sources

We are convinced, that available technologies from the Semantic Web can assist a user in the selection process of a good data source and in associated quality monitoring tasks. It can even free the user from the tedious work of reviewing data quality measurement values and assessing its appropriateness in a (semi-)automated fashion as long as an explicit, machine-readable requirement profile exists.

Overall Objective. To make the quality assessment task of eligible data sources more comprehensible, automatable and adjustable for the current usage scenario, we aim at explicitly defining particular quality requirements and tolerated measurement boundaries with semantic means in a quality requirement description

Requirements. To contribute to the overall objective, the following challenges had to be considered for defining a quality assessment requirement vocabulary:

1. To enable human data consumers to define multiple usage requirements on a data source in a semantic fashion
2. To reuse existing standards and established references for quality concepts
3. To allow the interpretation of quality measurement values
4. To determine an appropriate assessment rating score
5. To increase comparability among different quality checkers and data sources

3 The DaQAR Approach

We propose a vocabulary for describing "Data Quality Assessment Requirements" (DaQAR)[3]. Its main application domain is the description of quality requirements on Semantic Data in an RDF representation, but it can obviously also be applied to traditional data sources as long as they are addresseable by a URI. It includes the concepts depicted in Fig. 2 and seeks to reuse existing W3C standards and already established data-related vocabularies, primarily DCAT and DQV.

A DaQAR description starts with a requirement specification as shown in Listing 1.1. It contains a list of all conceptual quality requirements of interest.

[3] http://purl.org/net/vsr/daqar#.

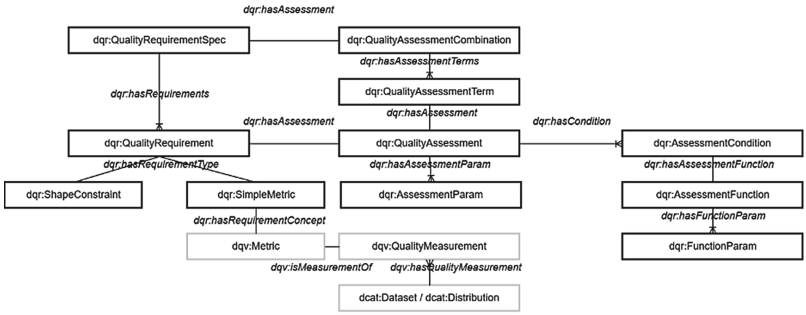

Fig. 2. DaQAR data model with main relations

Additionally, further expectations could be mentioned, such as the desired type of examined resources (e.g., *foaf:Person*) that is wanted for this requirement specification.

```
:OverallRequirement
    a dqr:QualityRequirementSpec ;
    dqr:hasRequirements :requirement1, :requirement2 ;
    dqr:hasAssessment :OverallAssessment .
```

Listing 1.1. General Requirement Specification

The description of a particular requirement contains the expected concept to measure and additional desired meta properties as shown in Listing 1.2. DaQAR is designed in such a way that arbitrary approaches to measure requirements can be used as long as their concepts can be referenced by a URL. This includes well-known quality daQ, DQV and ISO quality metrics (we refer to them as *sqr:SimpleMetric* in our vocabulary), constraint descriptions via SHACL shapes as well as other addressable SE metrics such as from QoS ontologies.

```
:requirement1
    a dqr:QualityRequirement ;
    dqr:hasRequirementType dqr:SimpleMetric ;
    dqr:hasRequirementConcept dqv:LowLatencyMetric ;
    dqr:expectedDataType xsd:double ;
    dqr:hasUnit muo:Seconds
    dqr:hasAssessment :assessment1 .
```

Listing 1.2. Description of a particular requirement

The *dqr:hasAssessment* property can either refer to an individual description on how to assess measured values within expected boundaries (see Listing 1.3); or it can refer to a globally known assessment description that utilizes additionally provided *dqv:assessmentParam* parameters.

```
:assessment1
    a dqr:QualityAssessment ;
    dqr:hasReturnType dqr:QualityRatio ;
    dqr:hasReturnValue dqr:FunctionalValue ;
    dqr:hasCondition [
        a dqr:AssessmentCondition ;
        dqr:hasValueGreaterEqualThan "0.0"^^xsd:double ;
        dqr:hasValueLowerEqualThan "2.0"^^xsd:double ;
        dqr:hasAssessmentFunction [
            a dqr:AssessmentFunction ;
            dqr:hasFunctionType dqr:LinearFunction ;
            dqr:hasFunctionParam [
                a dqr:FunctionParam ;
                dqr:hasStartValue "1.0"^^xsd:double
            ], [
                a dqr:FunctionParam ;
                dqr:hasEndValue "0.0"^^xsd:double
    ]]] .
```

Listing 1.3. Partial description of an assessment function

Obviously, such a description of a mathematical relationship in a declarative, semantic fashion becomes complex very soon and will probably not be written by an unexperienced user. Luckily, assessment definitions for quality requirements on metrics as e.g. specified in Listing 1.3 are often similar and only differ in concrete parameter values. From a semantic point of view, it makes sense to not solely define custom assessment functions locally, but to represent same assessment calculation descriptions under a global URI and only specify concrete parameter values as exemplary illustrated in Listing 1.4 as a modification of Listing 1.2.

```
:requirement1
    a dqr:QualityRequirement ;
    dqr:hasRequirementType dqr:SimpleMetric ;
    dqr:hasRequirementConcept dqv:LowLatencyMetric ;
    dqr:expectedDataType xsd:double ;
    dqr:hasUnit muo:Seconds ;
    dqr:hasAssessment <http://purl.org/net/vsr/daqar/
            assessmentdefinitions/LowerThanMaxBoundary> ;
    dqr:hasAssessmentParam :maxResponseTime .
```

Listing 1.4. Description of a particular requirement by referencing a globally known assessment definition

The assessments of multiple requirements can then be combined to an assessment of the entire data source to calculate an overall rating score as shown in Listing 1.5. A customer can rate the personal importance of each requirement for the overall assessment. We therefore allow a linear combination with weights (aka weighted scoring model) of each measurement result. Advanced concepts (such as regularization terms) could be considered as well.

```
:OverallAssessment
    a dqr:QualityAssessmentCombination ;
    dqr:ReturnType dqr:QualityRatio ;
    dqr:hasAssessmentTerms [
    a dqr:QualityAssessmentTerm ;
    dqr:hasWeight "0.6"xsd:double ;
    dqr:hasAssessment :assessment1
    ], [
    a dqr:QualityAssessmentTerm ;
    dqr:hasWeight "0.4"^^xsd:double ;
    dqr:hasAssessment :assessment2
    ] .
```

Listing 1.5. Definition of an overall assessment function

The result is commonly a *dqr:QualityRatio* numeric value indicating how good all posed requirements are fulfilled within their expected value limits. Instead of a ratio, the vocabulary is open for other *dqr:ReturnTypes* such as a star classification or a grade-based scale which could be derived from portioning the decimal assessment results in multiple groups. The requirement specification of Listing 1.1 can be instantiated for any data source state at a given point of time and result in a comparable assessment report value based on the explicit calculation instruction of Listing 1.5. This is independent from the assessment tool implementation as long as it sticks to the provided DQ requirement descriptions.

4 Evaluation

The ontology description is publicly available via GitHub and Zenodo[4], and a prototypical implementation of a DaQAR enabled quality assessment tool (SemQuire) has already been done in an industrial Linked Enterprise Data Services (LEDS) growth-core project context in Germany.

4.1 Methodology

We evaluated the DaQAR approach by using the objectives-based evaluation method [11] as the most prevalent approach in program evaluation.

As the practical adoption of our approach has not yet reached a certain level so that a credible field experiment or case study could be conducted and profoundly reviewed, an objectives-based study is a questions-oriented evaluation approach that can be performed in the meantime internally by program developers [11].

4.2 Discussion of Findings

To enable human data consumers to define multiple usage requirements on a data source in a semantic fashion, we followed the idea of a declarative, semantically-enabled description file and proposed in the DaQAR vocabulary the concept of

[4] https://doi.org/10.5281/zenodo.1039659.

a-priori definitions of the type *dqr:QualityRequirementSpec*. Such a specification consists of a set of requirements posed on a data source. The descriptions itself are human-readable by nature, but they obviously do not have to be written by hand but are also possible to create via some wizard GUI in future applications.

To reuse existing standards and established references for quality concepts or metrics, we fill a practical gap as an extension for W3C's DQV vocabulary and rely on RDF and OWL as proven techniques from the Semantic Web community.

To allow the interpretation of quality measurement values is done by allowing the definition of individual, expected or tolerated thresholds on the measurement results and custom rating functions. Additionally, we briefly discussed in Sect. 3 the idea to globally define referenceable assessment functions via a URI that can be instantiated by solely specifying parameter values. To the best of our knowledge, no other similar approach exists so far in the Semantic Web community. We will further investigate this concept in the future.

To determine an appropriate assessment rating score is ensured in DaQAR by providing a comprehensive assessment rating description. This is ensured both for each requirement itself as well as for the overall assessment score.

To increase comparability among different quality checkers and data sources is ensured, when each evaluation relies on the same DaQAR quality assessment requirement description for a particular, intended use case. We made both the requested measurement concepts as well as the definition on how to compute the overall assessment score explicit, so a Quality Checker or Quality Assessment Rating Tool that is compatible with DaQAR description can stick to this specification and deliver comparable results.

5 Related Work

A wide range of quality measurement and assessment tools already exists. A survey compared the most relevant tools [13]. Among these tools, the quality calculation and assessment is done in different ways. For instance, the Open-Data Checker [1] calculated metrics from data quality indicators specifically for CKAN data stores and simply outputted them in percent. KBMetrics [10] used a scoring system to make different data sources comparable. SWIQA [5] calculated a quality score based on the percentage how many instances violate given data quality rules.

Although multiple quality assessment tools already existed in the past and made use of RDF concepts, their assessment results of the quality of a data source were difficult to compare so far. Luzzu [3] for instance implemented several Linked Data quality metrics and introduced LQML as a domain-specific language for Linked Data quality assessment. RDFUnit [6] provides an own validation ontology and focuses a test-driven approach that can be run against an endpoint to validate RDF data. The support of DaQAR descriptions could enrich these tools; they do not mutually exclude each other. The DQV working group recently focused on harmonizing the output and comparison of quality metric values. For stating overall quality results, they considered to state the conformance degree

of a data source with predefined data policies by using ontologies such as ODRL. However, ODRL was primarily designed for representing permissions for data access policies and lacked further concepts to express desired value constraints. A first proposition to overcome this challenge was made by [8] where an extension of the DQV was described in the form of a quality model for Linked Data to compare and benchmark evaluation results and to include information about measurements and metric implementations. Our DaQAR approach also tries to fill this mismatch and extends the previous DQV work with different means for assessment requirement descriptions.

6 Conclusion

In this paper, we presented the DaQAR ontology, a vocabulary to describe Data Quality assessment requirements built on established W3C standards. This description includes both a list of required metrics and structural constraints, a method to predefine desired and tolerated measurement values, and an exchangeable assessment function description. All descriptions are independent from a particular data source and a concrete tool implementation. They are formulated as RDF triples and reuse existing data quality standards as good as possible. The DaQAR descriptions can be run on a particular dataset and commonly output a decimal assessment score representing a percentage to which the assessed data fulfills all specified requirements and limits. A conversation to other output formats is possible as well.

Acknowledgment. This work was supported by the grant from the German Federal Ministry of Education and Research (BMBF) for the LEDS Project under grant agreement No 03WKCG11D.

References

1. Assaf, A., Troncy, R., Senart, A.: Roomba: an extensible framework to validate and build dataset profiles. In: Gandon, F., Guéret, C., Villata, S., Breslin, J., Faron-Zucker, C., Zimmermann, A. (eds.) ESWC 2015. LNCS, vol. 9341, pp. 325–339. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-25639-9_46
2. Debattista, J., Lange, C., Auer, S.: DaQ, an ontology for dataset quality information. In: CEUR Workshop Proceedings. vol. 1184 (2014)
3. Debattista, J., Lange, C., Auer, S.: Luzzu - a framework for linked data quality assessment. In: ISWC 2015, 601043, pp. 1–16 (2015)
4. Fürber, C., Hepp, M.: Using semantic web resources for data quality management. In: Cimiano, P., Pinto, H.S. (eds.) EKAW 2010. LNCS (LNAI), vol. 6317, pp. 211–225. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-16438-5_15
5. Fürber, C., Hepp, M.: SWIQA a semantic web information quality assessment framework. In: Proceedings of the 19th European Conference on Information Systems (ECIS 2011), p. 76 (2011)
6. Kontokostas, D., Westphal, P., Auer, S., et al.: Test-driven evaluation of linked data quality. In: Proceedings of the 23rd International Conference on World Wide Web - WWW 2014, pp. 747–758 (2014). http://doi.acm.org/10.1145/2566486.2568002

7. Langer, A., Gaedke, M.: FAME.Q - a formal approach to master quality in enterprise linked data. In: Proceedings of the 15th International Conference on WWW/Internet (ICWI 2016), pp. 51–58, October 2016
8. Radulovic, F., Mihindukulasooriya, N., García-Castro, R., Gómez Pérez, A.: A comprehensive quality model for Linked Data. Semant. Web J. **1**, 1–5 (2015). http://www.semantic-web-journal.net/system/files/swj1488.pdf
9. Redman, T.C.: Data Quality: The Field Guide. Digital Press, Newton (2001)
10. Ruan, T., Dong, X., Li, Y., Wang, H.: KBMetrics - A Multi-purpose Tool for Measuring the Quality of Linked Open Data Sets (2015)
11. Stufflebeam, D.: Evaluation Models. New Directions for Evaluation, vol. 89, pp. 7–98. Jossey-Bass, San Francisco (2001)
12. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manag. Inf. Syst. **12**(4), 5–33 (1996)
13. Zaveri, A., Rula, A., Maurino, A., et al.: Quality assessment for linked open data: a survey. Semant. Web J. **1**, 1–31 (2014). http://www.semantic-web-journal.net/content/quality-assessment-methodologies-linked-open-data