Giving Dataset Quality Metadata Multi-Dimensional Views

Jeremy Debattista
University of Bonn /
Fraunhofer IAIS, Germany
name.surname@iais-extern.fraunhofer.de

Christoph Lange
University of Bonn /
Fraunhofer IAIS, Germany
math.semantic.web@gmail.com

Sören Auer
University of Bonn /
Fraunhofer IAIS, Germany
auer@cs.uni-bonn.de

ABSTRACT

Data quality is commonly defined as *fitness for use*. The problem of identifying quality of data is faced by many data consumers. On the other hand, data publishers do not have the means to identify quality problems in their data. To make the task for both stakeholders easier, we extend the Dataset Quality Ontology (daQ) with multi-dimensional and statistical properties from the Data Cube. The daQ is a light-weight, extensible vocabulary for attaching the results of quality benchmarking of a linked open dataset to the dataset. We discuss the design considerations, give examples for extending daQ by custom quality metrics, and present use cases such as analysing data versions, browsing datasets by quality and link identification. We also discuss how visualisation tools enable data publishers to analyse better the quality of their data.

Categories and Subject Descriptors

The Web of Data [**Vocabularies, taxonomies and schemas for the web of data**]

General Terms

Documentation, Measurement, Quality, Ontology

1. INTRODUCTION

There are various definitions for the term 'Data Quality'. Robert Pirsig defines quality as *the result of care* [20], whilst Juran defines quality as being *fitness for use* [17]. Juran's views on data quality were shared by Phillip Crosby, where he defined quality as *conformance to requirements* [8]. A substantial amount of Linked (Open) Datasets have already been published[1]. Most of these facts are extracted from heterogeneous sources, including semi-structured data and unstructured data, which do not guarantee high quality. Therefore such sources could lead to various problems such as inconsistencies and incompleteness, which could render a dataset to not be fit for a certain tasks. Moreover datasets might also evolve during their life span, leading to increase or decrease in quality.

Identifying the right quality factors for a dataset is a challenge which is always faced by many data consumers. The main problem is contributed by the fact that different domains require different quality metrics. Various research work [5, 11, 15] defines a number of factors that are pertinent to linked open datasets. In addition, Zaveri et al. [22] provides a systematic literature review categorising the different metrics.

To make the task of allowing different metrics defined in a standardised manner, we introduced the generic Dataset Quality Ontology (daQ) framework in [10]. The daQ is a light-weight ontology that allows datasets to be "stamped" with a number of respective quality measures, allowing for the expression of concrete, tangible values that represent the quality of the data. In this paper, we present the Data Cube extension of daQ, which allows for representing statistical observations in multidimensional spaces.

To put the reader into the context of this work, we introduce a use case:

> Quimp is a startup company providing various life science linked datasets. Their business model is that their customers (the real publishers) provide their data in various formats to the Quimp Portal. Meanwhile Quimp semantically lift the data to a more standardised Linked RDF representation using a number of pre-defined ontologies. This freshly-lifted data is periodically updated, having new resources added and others becoming obsolete. Quimp will then offer access for these linked datasets to data consumers. Data consumers often complain about the fact that datasets suffer a lot from incorrect and inconsistent facts. Concerned with these complaints, Quimp decided to start computing quality metrics on the semantically-enriched data. The quality metadata is also stored in order to visualise the quality change between different versions. The quality metrics through the latter visualisations help Quimp identify what aspects of their data is not up to standard, and therefore ensuring that quality over the different versions does not diminish.

The remainder of this paper is structured as follows: in Section 2 we discuss use cases for the extended daQ vocabulary. Then, in Section 3 and 4 we discuss the vocabulary design and show how data can be explored and used. Finally, in Section 5 we give an overview of similar ontology approaches before giving our final remarks in Section 6.

2. USE CASES

[1]Some statistics can be found in `http://lod-cloud.net` and `http://stats.lod2.eu`

Linked Open Data quality has different stakeholders in a myriad of domains, however, the stakeholders can be cast under either *publishers* or *consumers*. *Publishers* are mainly interested in publishing data that others can reuse. *Consumers*, both human end users and machine agents assisting them, require to use this published data in their applications.

Data consumers, both human and machine, may find it challenging to assess the quality of a dataset, i.e. its fitness for use. Currently, there is no standard way of how data publishers can assess the quality of their linked datasets. Most of these publishers rely on their mutual trust they have with data providers, believing that the data they provide is good for data consumers, leaving the data publishers in the dark of the value and quality of their data. Undoubtedly, this quality metadata is also significant to the data consumer who ultimately has to decide what data is fit to their use case. Activities of the recently formed *W3C Data on the Web Best Practices Working Group* (DWBP)[2][3] include developing a standard vocabulary for assessing and representing metadata about quality. Our daQ [10], or an adaptation or extension of it, is a candidate for this vocabulary.

The following use cases (UC) show how both data publishers and consumers can benefit from having multi-dimensional and statistical views of quality metadata on their published datasets. The UC thus motivate the need for extending daQ with the Data Cube vocabulary, the W3C standard for multi-dimensional data [9].

2.1 UC1: Analysis of Data Versions

Ideally, data publishers update their published datasets regularly to (a) keep the data fresh and up-to-date; (b) clean data to improve quality; (c) keep up with the data curation lifecycle. However, it is sometimes difficult to identify which aspects of the data are lacking quality standards. Furthermore, it is even more difficult to analyse how data quality changed over time. Given the necessary tools to analyse data quality[4], our proposed extension of daQ with Data Cube provides means for data publishers to have quality metadata represented in a multi-dimensional manner. Therefore, this ontology can represent metadata such as quality metric values against a set of different versions of a dataset. To keep quality metrics information easily accessible, we recommend that each dataset contains the relevant daQ metadata graph within the dataset itself.

2.2 UC2: "Fit" Dataset for Retrieval

Alexander et al. [1] provide the readers with a motivational use case with regard to how the voID ontology (cf. Section 5) can help with effective data selection. The authors describe that a consumer can find the appropriate dataset by:

- defining a criterion for content (what is the dataset mainly about);
- interlinking (to which other dataset is the one in question interlinked);
- vocabularies (what vocabularies are used in the dataset).

The daQ vocabulary gives an extra edge to "appropriateness" by providing the consumer with added quality criteria on the candidate datasets.

[2] `https://www.w3.org/2013/dwbp`

[3] The authors are affiliated with this WG, contributing to the standardisation of quality assessment of LOD.

[4] We are currently implementing a quality framework with a number of metrics which can be calculated over linked open datasets

An objective assessment of data quality enables data consumers to determine if a dataset is fit for a certain use case. Currently, tools catered for human data consumers such as semantic web search engines [16] or Data Web browsers [4, 13, 14], do not focus on dataset quality when presenting search results. With the introduction of the daQ framework, tools that provide faceted browsing facilities, such as the CKAN data portal engine[5], are enabled to provide more information about a dataset's quality attributes. Such functionality is attributable to the flexibility of the vocabulary, providing various filtering and ranking possibilities of the dataset quality metrics. This would permit human data consumers to have a better idea about the quality attributes of a dataset, and thus choose which is the most fitting to their use case. The daQ extension of multi-dimensional quality metadata not only enables this filtering and ranking functionality for open data management portals, but also enables data consumers to track and follow quality improvements of data publishers on their datasets. This also opens a sundry of opportunities leading to the assessment of data publishers regarding their willingness to enhance the value of the data in terms of quality.

2.3 UC3: Link Identification

Identifying links between existing datasets is one of the main drivers that makes the Linked Open Data cloud more coherent. Tools such as LIMES [19] and Silk [21] support the automatic identification of links according to built-in as well as user-defined criteria. The introduction of quality metadata to datasets will add another criterion for link identification, in that linking algorithms can also take the quality of the data into consideration before linking two resources. Linking tools could also consider the needs of a data consumer who might not only require to link to any high quality entity, but possibly even to those datasets which the consumer deems "fit" to her cause. This can be done by ranking and filtering candidate datasets according to criteria such as weights on specific quality metrics defined by the consumer (as described in *UC2*). Linking resources of proven quality helps to improve the quality of both datasets participating in the link. The generic framework proposed for the daQ vocabulary ensures that any custom metric defined by third parties can be easily integrated into any tool supporting such quality metadata for linking.

2.4 UC4: Extension of the Five Star Scheme

The popular five star scheme for deploying open data[6], which we propose to extend by a sixth star for quality, defines a set of widely accepted criteria that serve as a baseline for assessing data reusability. The reusability criteria defined by the five star scheme and by the quality metrics are largely measurable in an objective way. Thanks to such objective criteria, one can assess the reusability of any given dataset without the major effort of, for example, running a custom survey to determine whether its intended target audience finds it reusable. Such a survey may, of course, still help to get an *even better* understanding of quality issues. As a consumer, the benefits of a sixth star is that good quality datasets can be discovered. On the other hand, as a data publisher, the benefits of having the sixth star are that (i) the published data conforms to the established domain quality metrics; and (ii) catalogued and archived datasets (refer to [10]) can be easily discovered when consumers filter by quality aspects.

[5] `http://ckan.org`

[6] `http://5stardata.info`

3. THE DATASET QUALITY ONTOLOGY (DAQ)

The idea behind the Dataset Quality Ontology[7] (daQ) is to provide a comprehensive generic vocabulary framework, allowing a uniform definition of specific data quality metrics. This metric definition would then allow publishers to attach data quality metadata with quality benchmarking results to their linked dataset. In [10], the basic and most fundamental concepts were introduced.

3.1 The basic daQ Concepts

In this section we will briefly describe the daQ concepts introduced and formalised in [10]. Using daQ, the quality metadata is intended to be stored in what we defined to be the *Quality Graph*. The latter concept is a subclass of `rdfg:Graph` [6]. This means that the quality metadata is stored and managed in a separate named graph from the calculated dataset. Named graphs also allow the digital signing of graphs [7] , ensuring trust in the computed metrics.

The daQ ontology distinguishes between three layers of abstraction, based on the survey work by Zaveri et al. [22]. As shown in Figure 1 Box B, a quality graph comprises of a number of different *Categories*, which in turn possess a number of quality *Dimensions*[8]. A quality dimension groups one or more computed quality *Metrics*.

3.2 Extending daQ for Multi-Dimension Representation and Statistical Evaluation

The Data Cube Vocabulary [9] allows the representation of statistics about observations in a multidimensional attribute spaces. Previously, computing quality metrics of different resources (usually: datasets), or even different revisions of the same resource, resulted in multiple *quality graph*s, consisting of multiple instances of *Metric* classes representing the individual observations. Multidimensional analysis of these observations, e.g. across the revision history of a dataset, would thus have required complex querying. Extending daQ with the standardised Data Cube Vocabulary allows us to represent quality metadata of a dataset as a collection of *Observations*, dimensions being the different quality metrics computed, the resources whose quality is assessed, revisions of these resources, and arbitrary further dimensions, such as the intended application scenario. It also permits applying the wide range of tools that support data cubes to quality graphs, including the CubeViz visualisation tool[9].

Figure 1 shows the current state of daQ including its data cube extension, which entails some structural changes over the initial version of the vocabulary as it was presented in [10]. A *Quality Graph* is a special case of `qb:DataSet`, which allows us to represent a collection of quality observations complying to a defined dimensional structure. Each observation represents a quality metric measured out against a particular resource (e.g. a specific revision of a dataset). daQ defines the structure of such observations by the `qb:DataStructureDefinition` shown in Listing 1.

```
daq:dsd a qb:DataStructureDefinition ;
  # Dimensions: metrics and what they were computed on
  qb:component [
    qb:dimension daq:metric ;
    qb:order 1 ; ] ;
  qb:component [
    qb:dimension daq:computedOn ;
    qb:order 2 ; ] ;
  # Measures (here: metric values)
  qb:component [ qb:measure daq:value ; ] ;
  # Attribute (here: the unit of measurement)
  qb:component [
    qb:attribute sdmx-attribute:unitMeasure ;
    qb:componentRequired false ;
    qb:componentAttachment qb:DataSet ; ] .
```

Listing 1: The Data Structure Definition (Turtle Syntax)

The `daq:QualityGraph` also defines one restriction that controls the property `qb:structure` and its value to the mentioned definition, thus ensuring that all *Quality Graph* instances make use of the standard definition. Having a standard definition ensures that all *Quality Graph*s conform to a common data structure definition, thus datasets with attached quality metadata can be compared. Listing 2 describes the definition of `daq:QualityGraph`.

```
daq:QualityGraph
  a rdfs:Class, owl:Class  ;
  rdfs:subClassOf rdfg:Graph , qb:DataSet ,
    [ rdf:type owl:Restriction ;
      owl:onProperty qb:structure ;
      owl:hasValue daq:dsd ];
  rdfs:comment "Defines a quality graph which will
       contain all metadata about quality metrics on the
       dataset." ;
  rdfs:label "Quality Graph Statistics" .
```

Listing 2: The Quality Graph Definition (Turtle Syntax)

All abstract classes in Figure 1 Box B, except for `daq:Metric` have the same definition as in [10]. The `daq:Metric` class is now linked to a `qb:Observation` using the newly defined property `daq:hasObservation`. The properties `daq:computedOn` and `daq:value` are now defined as `qb:DimensionProperty` and `qb:MeasureProperty` respectively. The former is defined in each observation instance rather than once as a *Quality Graph* property. We also introduce the `daq:metric qb:DimensionProperty` to represent the metric being observed. Each observation will also include the `daq:dateComputed` property, which holds the timestamp of when it was computed. Each custom metric definition should also include the `daq:expectedDataType` property. This will indicate the observation's value datatype. The optional property `sdmx-attribute:unitMeasure` can be defined on an observation instance, enabling a system (application) to "understand" the measure of the value.

4. USING THE ONTOLOGY

4.1 Extending daQ

The classes of the core daQ vocabulary can be extended by more specific and custom quality metrics. In order to use the daQ, one should define the quality metrics that characterise the "fitness for use" [17] in a particular domain. We are currently in the process of defining the quality dimensions and metrics described in [22], some of which are being considered to be standard metrics to calculate quality on Linked Open Data sets whilst others are specific to the DIACHRON project[10] (refer to Section 4.5). **Extending** the daQ vocabulary means adding new quality protocols that inherit the abstract concepts (Category-Dimension-Metric). Custom quality metrics do not need to be included in the daQ namespace itself; in fact, in accordance with LOD best practices, we recommend extenders to make them in their own namespaces. In Figure 2 we show an illustrative example of extending the daQ ontology (TBox)

[7] `http://purl.org/eis/vocab/daq`

[8] In this paper we will refer to these as quality dimensions, in order to distinguish between the data cube dimensions

[9] `http://cubeviz.aksw.org`

[10] `http://diachron-fp7.eu`

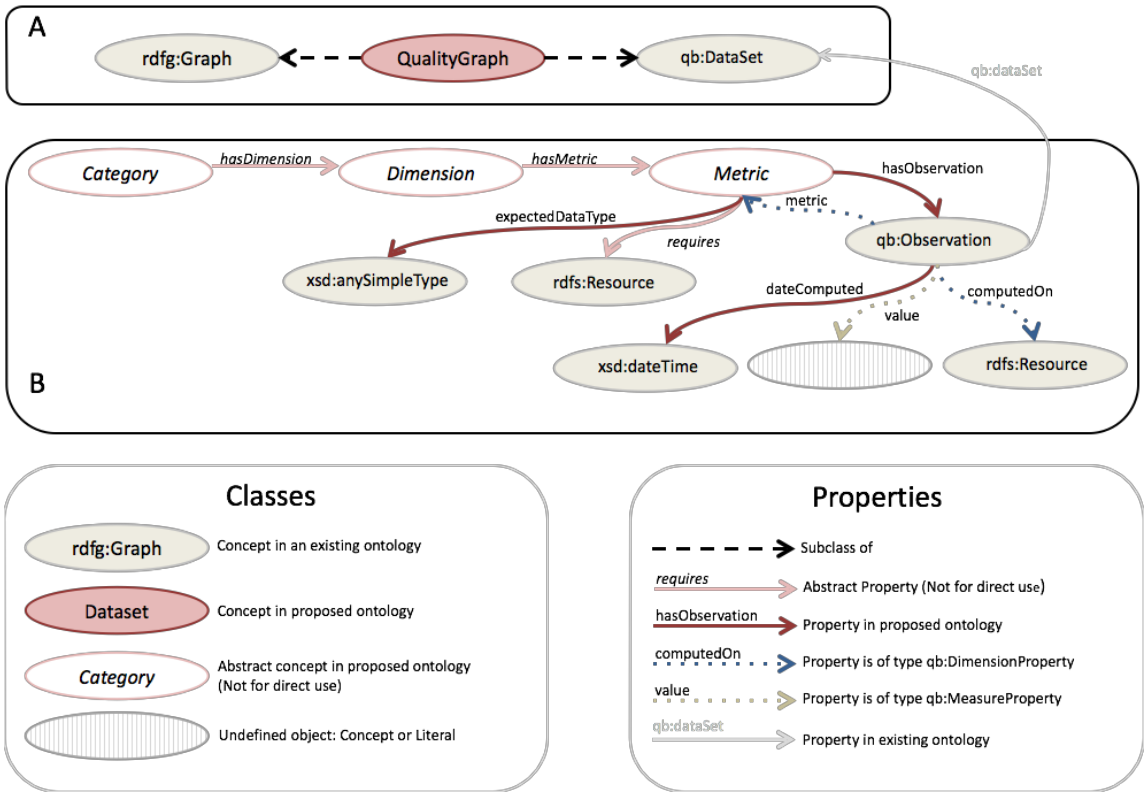


Figure 1: The extended Dataset Quality Ontology (daQ)

with a more specific quality attribute, i.e. the RDF Availability Metric as defined in [22], and an illustrative instance (ABox) of how it would be represented in a dataset.

The `Accessibility` concept is defined as an `rdfs:subClassOf` the abstract `daq:Category`. This category has five quality dimensions, one of which is the *Availability* dimension. This is defined as an `rdfs:subClassOf daq:Dimension`. Similarly, *RDFAvailabilityMetric* is defined as an `rdfs:subClassOf daq:Metric`. The specific properties *hasAvailabilityDimension* and *hasRDFAccessibilityMetric* (sub-properties of `daq:hasDimension` and `daq:hasMetric` respectively) are also defined (Figure 2).

4.2 Publishing daQ Metadata Records

Dataset publishers should offer a daQ description as an RDF Named Graph in their published dataset. Since such a daQ metadata record requires a lot of metrics to be computed, it is not normally intended to be authored manually. Publishing platforms such as CKAN should offer such an on-demand computation to dataset publishers (refer to [10]). Listing 3 shows an instance of the `daq:QualityGraph` in a dataset. *ex:qualityGraph1* is a named `daq:QualityGraph`. The defined graph is automatically a `qb:DataSet`, and due to the restriction placed on the `daq:QualityGraph` (see Listing 2), the value for the `qb:structure` property is defined as `daq:dsd` (see Listing 1). In the named graph, instances for the `daq:Accessibility`, `daq:Availability`, `daq:EndPointAvailabilityMetric` and `daq:RDFAvailabilityMetric` are shown. A metric instance has a number of observations. Each of these observations specifies the metric value (`daq:value`), the resource the metric was computed on (`daq:computedOn` – here: different datasets, which are actually different revisions of one dataset), when it was computed (`daq:dateComputed`), the metric instance (`daq:metric`) and finally to what dataset the observation is defined in (`qb:dataSet`).

```
# ... prefixes

# ... dataset triples

ex:qualityGraph1 a daq:QualityGraph ;
  qb:structure daq:dsd .

ex:qualityGraph1 {

  # ... quality triples
  ex:accessibilityCategory a dqm:Accessibility ;
    dqm:hasAvailabilityDimension ex:availabilityDimension
          .

  ex:availabilityDimension a dqm:Availability ;
    dqm:hasEndPointAvailabilityMetric ex:endPointMetric ;
    dqm:hasRDFAvailabilityMetric ex:rdfAvailMetric .

  ex:endPointMetric a dqm:EndPointAvailabilityMetric ;
    daq:hasObservation ex:obs1, ex:obs2 .

  ex:obs1 a qb:Observation ;
    daq:computedOn <efo-2.43> ;
    daq:dateComputed "2014-01-23T14:53:00"^^xsd:dateTime
          ;
    daq:value "1.0"^^xsd:double ;
    daq:metric ex:endPointMetric ;
    qb:dataSet ex:qualityGraph1 .

  ex:obs2 a qb:Observation ;
    daq:computedOn <efo-2.44> ;
    daq:dateComputed "2014-01-25T14:53:00"^^xsd:dateTime
          ;
```

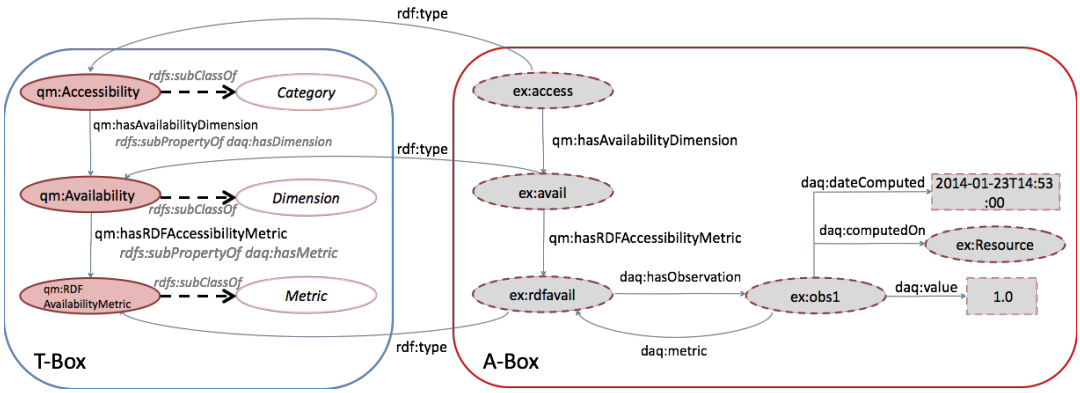


Figure 2: Extending the daQ Ontology – TBox and ABox

```
    daq:value "1.0"^^xsd:double ;
    daq:metric ex:endPointMetric ;
    qb:dataSet ex:qualityGraph1 .

  ex:rdfAvailMetric a dqm:RDFAvailabilityMetric ;
      daq:hasObservation ex:obs3, ex:obs4 .

  ex:obs3 a qb:Observation ;
    daq:computedOn <efo-2.43> ;
    daq:dateComputed "2014-01-23T14:53:01"^^xsd:dateTime
        ;
    daq:value "1.0"^^xsd:double ;
    daq:metric ex:rdfAvailMetric ;
    qb:dataSet ex:qualityGraph1 .

  ex:obs4 a qb:Observation ;
    daq:computedOn <efo-2.44> ;
    daq:dateComputed "2014-01-25T14:53:01"^^xsd:dateTime
        ;
    daq:value "0.0"^^xsd:double ;
    daq:metric ex:rdfAvailMetric ;
    qb:dataSet ex:qualityGraph1 .

  # ... more quality triples
}
```

Listing 3: A *Dataset Quality Graph*

4.3 Exploring and Visualising the daQ Metadata

CubeViz is a tool for visualising data cubes. Figure 3 depicts four different CubeViz chart visualisations from computed quality metadata[11].

A *horizontal bar* represents each metric (Figure 3(a)) and shows its value (x-axis) with respect to the dataset (y-axis). Here, the different 'datasets' analysed are actually successive revisions of one dataset. This chart provides a clear view of how the value associated to each one of the measured metrics changes as the dataset evolves. The horizontal layout is appropriate when the range of metric values is wide, and the number of different datasets is relatively small.

Similar to the horizontal bars chart, the *vertical bar chart* (Figure 3(b)) allows the user to compare the values computed for each of the metrics (y-axis), with respect to the dataset (x-axis). In contrast with its horizontal counterpart, this chart is more appropriate when there are many datasets analysed but the range of metric values is not so wide.

In the *radar chart* (Figure 3(c)), the datasets are represented as slices of a circle and the values corresponding to the metrics are depicted as points and lines of a particular color. This chart provides a clear view of how the values of the metric differ from each other for each particular dataset. Furthermore, it allows one to assess the overall quality of a dataset, by showing whether the values of the metrics are concentrated around sections of the circle regarded as 'good' or 'bad'.

The lines plot (Figure 3(d)), lists the different datasets against the values of the metrics. Here, where 'different datasets' are actually different revisions in the evolution of one dataset, this plot provides a comparison of the evolution of the quality of the dataset, with respect to each metric. The lines emphasise the points where the values of the metrics changed noticeably from one version to the next.

4.4 Creating Observations based on Quality Dimensions

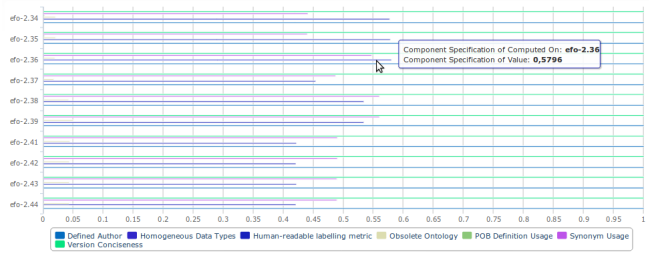
The daQ framework allows the definition of quality metrics in three levels of abstraction: Category – Dimension – Metric. Although the instances have a link between these three levels, we only perform observations on the metric level. Therefore, when visualising and analysing observations, the consumer would only be able to observe the metrics from all categories and dimensions, instead by specific dimension or category. Thanks to the link between the three levels of abstraction, no manual human intervention is required to analyse a set of metrics grouped by a specific quality dimension.

Data Cube slices allow the grouping of observation subsets. Since slices are not intended to represent arbitrary selections in a data cube, `qb:ObservationGroup` should be used. Listing 4 shows a SPARQL CONSTRUCT defining an ObservationGroup, where all observations in the Accessibility dimension are grouped in a constructed `ex:dimObs1` resource.

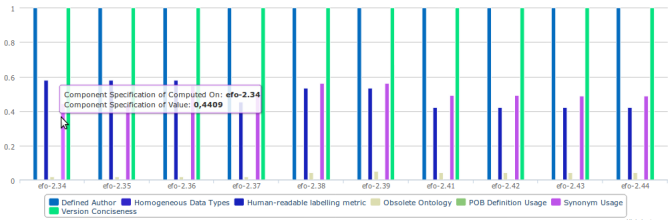
```
CONSTRUCT { ex:dimObs1 a qb:ObservationGroup ;
        qb:observation ?obs .
      }
WHERE
 {
   SELECT DISTINCT ?metricInst ?obs {
   ?dimInst  a dqm:Accessibility .
   ?dimInst  ?prop ?metricInst .
   ?metricInst daq:hasObservation  ?obs .
   ?metricInst a ?metric .

   GRAPH <http://www.diachron-fp7.eu/dqm#> {
```

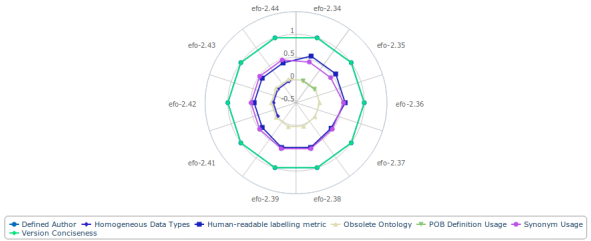
[11]The quality metadata used can be found in `https://raw.githubusercontent.com/diachron/quality/master/src/test/resources/cube_qg.trig`



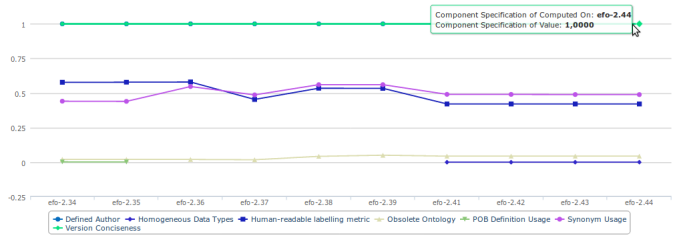
(a) Horizontal Bar Chart



(b) Vertical Bar Chart



(c) Radar Chart



(d) Lines Plot

Figure 3: Visualising Quality Metadata

```
    ?prop rdfs:subPropertyOf  daq:hasMetric .
    ?metric rdfs:subClassOf daq:Metric .
  }
}
}
```

Listing 4: Creating A Data Cube Observation Group using SPARQL

The resulting construct output is shown in Listing 5.

```
ex:dimObs1 a qb:ObservationGroup ;
 qb:observations ex:obs1 , ex:obs2 , ex:obs3 , ex:obs4 .
```

Listing 5: A Data Cube Observation Group.

4.5 The DIACHRON Project

The DIACHRON project ("Managing the Evolution and Preservation of the Data Web") combines several of the use cases mentioned so far. DIACHRON's central cataloguing and archiving hub is intended to host datasets throughout several stages of their life-cycle [3], mainly evolution, archiving, provenance, annotation, citation and data quality. As a part of the DIACHRON project, we are implementing scalable and efficient tools to assess the quality of datasets. A web-based visualisation tool, to be implemented as a CKAN plugin, will

- allow data publishers to perform quality assessment on datasets, which will provide them with quality score metadata and also assist them with fixing quality problems;
- allow data consumers to filter and rank datasets by multiple quality dimensions.

The daQ vocabulary is the core ontology underlying these services. It will help these services to do their jobs, i.e. adding quality metadata to datasets, which in turn is displayed on the web frontend.

5. RELATED WORK

To the best of our knowledge, the Data Quality Management (DQM) vocabulary [12] is the only one comparable to our approach. Fürber et al. propose an OWL vocabulary that primarily represents data requirements, i.e. what quality requirements or rules should be defined for the data. Such rules can be defined by the user herself, and the authors present SPARQL queries that "execute" the definitions of the requirements to compute metrics values. Unlike our daQ model, the DQM defines a number of classes that can be used to represent a data quality rule. Similarly, properties for defining rules and other generic properties such as the rule creator are specified. The daQ model allows for integrating such DQM rule definitions using the *daq:requires* abstract property, but we consider the definition of rules out of daQ's own scope. Also, the proposed daQ vocabulary gives the freedom to the user to define and implement any metrics required for a certain application domain.

Our design approach is inspired by the digital.me Context Ontology (DCON[12]) [2]. Attard et al. present a structured three-level representation of context elements (Aspect-Element-Attributes). The DCON ontology instances are stored as Named Graphs in a user's Personal Information Model. The three levels are abstract concepts, which can be extended to represent different context aspects in a concrete ubiquitous computing situation.

The W3C recommends VoID and the Data Catalog Vocabulary (DCAT [18]) ontologies recommended by the W3C provide metadata vocabulary for describing datasets. The "Vocabulary of Interlinked Datasets" (voID) ontology allows the high-level description of a dataset and its links [1]. On the other hand, DCAT describes datasets in data catalogs, which increase discovery, allow easy interoperability between data catalogs and enable digital preservation. With the daQ ontology, we aim to extend what these two ontologies have managed to achieve for datasets in general to the specific aspect of quality; enabling the discovery of a good quality (fit to use) datasets by providing the facility to "stamp" a dataset with quality metadata.

6. CONCLUDING REMARKS

In this paper we presented an extension to the Dataset Quality Ontology (daQ), an extensible vocabulary to provide quality benchmarking metadata of a linked open dataset to the dataset itself. In Section 2 we presented a number of use cases that motivated our idea. These included analysis of data versions, dataset retrieval,

[12] http://www.semanticdesktop.org/ontologies/dcon/

automatic link identification based on the quality of data entities, and finally the extension of the five star open data scheme by a star for quality. The precise definition of these use cases assisted in the development of the daQ ontology and its Data Cube extension (Section 3).

The ontology is progressing in a fast pace, and further developments to cover the intended use cases are also in the pipeline. The next iteration phase is to further model the daQ ontology to cover the provenance aspect of quality metadata. The development and extension of new concepts to the daQ ontology should ensure that (i) high standards are kept, and (ii) that the ontology is not bloated out of proportion – i.e. keeping with the main idea that the framework is a light weight ontology.

Currently, using daQ, we are in the process of implementing (Section 4) a number of domain-specific and domain-independent metrics, following a survey of linked data quality metrics [22]. Since a quality metadata describes the LOD dataset on which quality was calculated, the daQ framework enables us to create Named Graphs within the dataset itself. We also demonstrated how quality metadata can be visualised using available Data Cube enabled applications such as CubeViz, and how observations can be grouped together automatically using the daQ three level abstract layer.

One of the tools which will support the daQ framework is the DIACHRON platform. This platform will enable consumers to rank and filter datasets according to the quality metadata. Having tools and platforms supporting the daQ will finally allow us to test and evaluate the vocabulary thoroughly, to see whether the daQ (and the quality metadata) itself is of a high quality, i.e. fit for use.

7. ACKNOWLEDGMENTS

This work is supported by the European Commission under the Seventh Framework Program FP7 grant 601043 (http://diachron-fp7.eu).

References

1. Alexander, K. et al. Describing Linked Datasets – On the Design and Usage of voiD, the 'Vocabulary of Interlinked Datasets'. In: *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*. Madrid, Spain, 2009.
2. Attard, J. et al. Ontology-based Situation Recognition for Context-aware Systems. In: *Proceedings of the 9th International Conference on Semantic Systems*. I-SEMANTICS '13. Graz, Austria: ACM, 2013, pp. 113–120. http://doi.acm.org/10.1145/2506182.2506197.
3. Auer, S. et al. Managing the life-cycle of Linked Data with the LOD2 Stack. In: *Proceedings of International Semantic Web Conference (ISWC 2012)*. 22% acceptance rate. 2012. http://iswc2012.semanticweb.org/sites/default/files/76500001.pdf.
4. Berners-Lee, T. et al. Tabulator: Exploring and Analyzing linked data on the Semantic Web. English. In: *Proceedings of the The 3rd International Semantic Web User Interaction Workshop (SWUI06)*. Nov. 2006.
5. Bizer, C. Quality-Driven Information Filtering in the Context of Web-Based Information Systems. PhD thesis. FU Berlin, Mar. 2007. http://www.diss.fu-berlin.de/diss/receive/FUDISS_thesis_000000002736.
6. Carroll, J. J. et al. Named Graphs, Provenance and Trust. In: *Proceedings of the 14th WWW conference*. (Chiba, Japan, 10th–14th May 2005). Ed. by A. Ellis, T. Hagino. ACM Press, 2005, pp. 613–622.
7. Carroll, J. J. et al. Semantic Web Publishing using Named Graphs. In: *ISWC Workshop on Trust, Security, and Reputation on the Semantic Web*. Ed. by J. Golbeck et al. Vol. 127. CEUR Workshop Proceedings. CEUR-WS.org, 9th May 2005. http://dblp.uni-trier.de/db/conf/semweb/iswc2004trust.html#CarrollBHS04.
8. Crosby, P. Quality is Free: The Art of Making Quality Certain. Mentor book. McGraw-Hill, 1979. http://books.google.ie/books?id=bR_LnQEACAAJ.
9. Cyganiak, R., Reynolds, D., Tennison, J. The RDF Data Cube Vocabulary. W3C Recommendation. World Wide Web Consortium (W3C), 16th Jan. 2014. http://www.w3.org/TR/2014/REC-vocab-data-cube-20140116/.
10. Debattista, J., Lange, C., Auer, S. daQ, an Ontology for Dataset Quality Information. In: *Linked Data on the Web (LDOW)*. (Seoul, 8th Apr. 2014). Ed. by C. Bizer et al. 2014. http://events.linkeddata.org/ldow2014/.
11. Flemming, A. Quality Characteristics of Linked Data Publishing Datasources. http://sourceforge.net/apps/mediawiki/trdf/index.php?title=Quality_Criteria_for_Linked_Data_sources. [Online; accessed 13-February-2014]. 2010.
12. Fürber, C., Hepp, M. Towards a Vocabulary for Data Quality Management in Semantic Web Architectures. In: *Proceedings of the 1st International Workshop on Linked Web Data Management*. LWDM '11. Uppsala, Sweden: ACM, 2011, pp. 1–8. http://doi.acm.org/10.1145/1966901.1966903.
13. Harth, A. VisiNav: A system for visual search and navigation on web data. In: Web Semantics: Science, Services and Agents on the World Wide Web 8(4) (2010). Semantic Web Challenge 2009 User Interaction in Semantic Web research, pp. 348–354. http://www.sciencedirect.com/science/article/pii/S1570826810000600.
14. Heim, P., Ziegler, J., Lohmann, S. gFacet: A Browser for the Web of Data. In: *Proceedings of the International Workshop on Interacting with Multimedia Content in the Social Semantic Web (IMC-SSW 2008)*. Vol. 417. CEUR Workshop Proceedings. Aachen, 2008, pp. 49–58. http://CEUR-WS.org/Vol-417.
15. Hogan, A. et al. An empirical survey of Linked Data conformance. In: J. Web Sem. 14 (2012), pp. 14–44.
16. Hogan, A. et al. Searching and browsing Linked Data with SWSE: The Semantic Web Search Engine. In: Web Semantics: Science, Services and Agents on the World Wide Web 9(4) (2011). {JWS} special issue on Semantic Search, pp. 365 – 401. http://www.sciencedirect.com/science/article/pii/S1570826811000473.
17. Juran, J. M. Juran's Quality Control Handbook. 4th ed. McGraw-Hill (Tx), 1974. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0070331766.
18. Maali, F., Erickson, J., Archer, P. Data Catalog Vocabulary (DCAT). W3C Recommendation. World Wide Web Consortium (W3C), 16th Jan. 2014. http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/.
19. Ngonga Ngomo, A.-C., Auer, S. LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In: *Proceedings of IJCAI*. 2011.

20. Pirsig, R. Zen and the Art of Motorcycle Maintenance: An Inquiry Into Values. Essence (Philosophy). Vintage, 1974. http://books.google.ie/books?id=M69poeV1UhoC.

21. Volz, J. et al. Discovering and Maintaining Links on the Web of Data. In: *ISWC*. Vol. 5823. LNCS. Springer, 2009.

22. Zaveri, A. et al. Quality Assessment Methodologies for Linked Open Data (Under Review). In: Semantic Web Journal (2013). This article is still under review. http://www.semantic-web-journal.net/content/quality-assessment-linked-open-data-survey.