Undefined 1 (2015) 1–5
IOS Press



# A comprehensive quality model for Linked Data

**Editor(s):** Name Surname, University, Country
**Solicited review(s):** Name Surname, University, Country
**Open review(s):** Name Surname, University, Country

Filip Radulovic *, Nandana Mihindukulasooriya, Raúl García-Castro and Asunción Gómez-Pérez
*Ontology Engineering Group, Escuela Técnica Superior de Ingenieros Informáticos*
*Universidad Politécnica de Madrid, Spain*

**Abstract.** With the increasing amount of Linked Data published on the Web, the community has recognised the importance of the quality of such data and a number of initiatives have been undertaken to specify and evaluate Linked Data quality. However, these initiatives are characterised by a high diversity in terms of the quality aspects that they address and measure. This leads to difficulties in comparing and benchmarking evaluation results, as well as in selecting the right data source according to certain quality needs. This paper presents a quality model for Linked Data, which provides a unique terminology and reference for Linked Data quality specification and evaluation. The mentioned quality model specifies a set of quality characteristics and quality measures related to Linked Data, together with formulas for the calculation of measures. Furthermore, this paper also presents an extension of the W3C Data Quality Vocabulary that can be used to capture quality information specific to Linked Data, a Linked Data representation of the Linked Data quality model, and a use case in which the benefits of the quality model proposed in this paper are presented in a tool for Linked Data evaluation.

Keywords: Data Quality, Linked Data, Ontology, Quality Model

## 1. Introduction

The Linked Data principles promote publishing data and interlinking them in a machine-readable manner using Web standards. Linked Data, along with the other Semantic Web technologies, allows data to be interlinked and reused across organizational boundaries instead of being data silos used by a single organization. Linked data has several advantages over other data paradigms, namely [1]: (a) global identifiers for data that can be accessed using the Web infrastructure and typed links between data from different applications; (b) the graph-based RDF data model that allows consuming and merging data from different sources without having to do complex structural transformations; and (c) explicit semantics of data expressed in RDF Schema or OWL ontologies which can be aligned and mapped to data models of other applications using techniques such as ontology matching. However, along with these benefits there are a new set of challenges for data quality with respect to aspects such as dereferenceable identifiers, semantic accuracy and consistency. Furthermore, the process of Linked Data generation generally includes data transformation steps, mapping data to several vocabularies or ontologies and fusing data from different data sources, which opens the door to possible data quality issues.

Quality is well recognized as a crucial need across domains (e.g., civil engineering, software), and in order to provide high quality products and services, the specification and evaluation of quality is of high importance [2]. Similarly, data is a pivotal asset in many domains such as medicine, education or government, and the importance of data quality has led to different data quality legislations such as the US Data Quality

*Corresponding author. E-mail: fradulovic@fi.upm.es.

Act [3] or the BCBS 239 Data Quality Mandate [4]. Furthermore, the success of many business processes depends largely on the quality of data [5]. While the literature describes different definitions of data quality, the general notion is that data quality is tightly connected to the use and usefulness of data. The ISO defines data quality as a "key component of the quality and usefulness of information derived from that data" [5], while Juran and Godfrey define it as a "fit for intended use in operations, decision-making, and planning" [6]. Following this discussion, with an increasing amount of data available on the Web as Linked Data, the quality of Linked Data datasets is of high concern.

Various initiatives exist with the common goal of specifying the quality of Linked Data and evaluating Linked Data datasets. Quality assessment in these initiatives, however, is quite diverse since different authors focus on different aspects of Linked Data, on different characteristics (e.g., completeness, licensing, accuracy), and on different measures for these characteristics (e.g., missing links, indication of attribution, semantically incorrect values). Furthermore, some authors have developed methodologies and tools for Linked Data evaluation, which are also characterized with a high diversity in terms of the evaluated characteristics and measures.

Quality models are important for providing consistent terminology and guidance for quality assessment and are the basis for the evaluation of any product or service. This is especially significant for the integration of evaluation results and benchmarking, which is one important aspect of evaluation [7,8], and without a quality model it is sometimes difficult to integrate evaluation results, perform benchmarking, or to select products or services according to their quality. Because of this, the ISO recognized the need for a quality model for data, and produced the ISO 25012 quality model [5]. However, the ISO data quality model can be regarded as very general and, furthermore, it does not include particularities of Linked Data. The W3C has also recognized the need for having a unified ontology for describing data quality and is in the process of producing the W3C Data Quality Vocabulary (DQV)[1] within the W3C Data on the Web Best Practices Working Group. Nevertheless, DQV aims to be a lightweight ontology suitable for any type of data on the web (e.g., CSV, XML, HTML, RDF, etc.), thus it is generic and does not address the specific characteristics of Linked Data. Furthermore, it only provides a base framework for describing quality metrics and measures but does not define concrete quality metrics, which are expected to appear in quality models. Nevertheless, a data quality model could use DQV as the base ontology for representing the quality model elements as Linked Data.

Motivated with the previous discussion, this paper presents a quality model for the evaluation of Linked Data. It is a hierarchical quality model that provides unique terminology and that describes quality elements (i.e., a set of quality characteristics and a set of measures) related to Linked Data. The quality model presented in this paper has been defined relying on the state of the art in Linked Data quality evaluation and specification and extends the ISO 25012 data quality model. Unlike the current state of the art, our quality model formalizes a classification of different types of quality measures, and defines some measures, together with their related formulas that have not been specified in the literature.

This paper also presents an extension of the W3C Data Quality Vocabulary in order to provide means to describe the particularities of Linked Data quality, together with the Linked Data representation of the quality model according to the mentioned extension (RDF instances), with dereferenceable URIs of all the quality model elements. By using these artefacts, it is possible to capture the evaluation results of any particular Linked Data dataset, as well as to make unique references to the evaluated metrics. This can ease interoperability and provide better integration of various evaluation efforts.

The quality model proposed in this paper has been used in a use case of a tool for the evaluation of Linked Data. This paper also presents the mentioned use case and how the quality model has been used in the development of the evaluation tool. Furthermore, the tool developed for the use case uses the DQV ontology and the extension proposed in this paper, together with the Linked Data representation of the quality model, to describe Linked Data evaluation results.

This paper is organized as follows. Section 2 presents related work in the fields of quality modelling, Linked Data quality evaluation and specification, and ontologies for representing Linked Data quality. Section 3 presents the quality model for Linked Data, while Section 4 presents the ontology for capturing the results of Linked Data assessment, as well as the RDF dataset that describes the Linked Data quality model. Section 5 presents a use case in which the Linked Data quality model proposed in this paper has been used in the de-

[1] http://www.w3.org/TR/vocab-dqv/

velopment of a tool for the evaluation of Linked Data datasets. Section 6 discusses the main contributions of this paper and, finally, Section 7 draws some conclusions and ideas for future work.

## 2. Preliminaries

This section describes the related work in the domains of interest of this paper, which include quality modeling, Linked Data quality specification, and ontologies for representing Linked Data quality.

### 2.1. Quality models

A quality model is defined through a set of specific quality characteristics, quality sub-characteristics, quality measures, and through the relationships between these characteristics and measures [5,9,10] and, to this extent, represents a specification of quality-related information. Quality measures defined in a quality model capture some information about quality characteristics and sub-characteristics and, usually, a classification of different types of quality measures is specified in the quality model [11,12]:

- *Base measures* are measures that are a direct output of an evaluation; they can be related to one part of an evaluation (i.e., one test) or to the whole evaluation. An example of a base measure for web browsers can be startup time, memory consumption or number of open tabs in a single test.
- *Derived measures* are measures obtained by combining different base measures. An example of a derived measure for web browsers can be memory consumption per open tab in a single test.
- *Indicators* are measures that are obtained by combining base and/or derived measures (e.g., from a number of tests), and are related to a whole evaluation. An example of an indicator for web browsers can be average startup time, or average memory consumption per tab.

Although the classification of the previously specified types of quality measures should be specified for each quality characteristic in a quality model, in the cases where a simple quality measure can sufficiently describe a sub-characteristic, derived and/or base measures are not necessary and quality indicators can be defined as a direct output of the evaluation. An example of such case can be an indicator that describes multilingual support of a web browser.

In some cases, additional inputs in the evaluation process are required in order to evaluate a specific measure (e.g., page loading time requires a set of specific web pages to load). Furthermore, these inputs are necessary in some cases in order to obtain enough information that enables the evaluation of a measure (e.g., for evaluating domain consistency in a triple of a dataset, it is necessary to obtain information about the ontology used for representing a dataset and the domain of a property found in the observed triple).

Where possible, for each quality measure, relationships should be defined formally (e.g., in terms of formulas) and each indicator is assigned as a measure of some quality sub-characteristic. For example, average startup time measures the browser time behaviour sub-characteristic, which can be defined as a sub-characteristic of the time behaviour characteristic.

Hierarchical structures of quality measures, as defined above, contribute in a better understanding of quality measures and their relationships [13]. Furthermore, especially having in mind information such as formulas, the evaluation of quality measures is more straightforward and an easier task when their complete hierarchy is defined.

Quality models are accepted as a valuable resource in quality assessment and specification and, in this context, they are used as a reference to the quality measures to be evaluated. By providing important details about quality measures, such as definitions, scales or formulas, quality models provide a guidance on which measures are important for evaluation and how to measure them. Various quality models have been described in the literature, both generic ones as domain-specific ones.

When it comes to data quality, the International Organization for Standardization (ISO) recognised the need for a data quality model and produced the ISO 25012 (SQuaRE) data quality model [5]. According to such standard, data quality is “a common prerequisite to all information technology projects".

The ISO 25012 quality model defines fifteen data quality characteristics classified into inherent data quality characteristics and system-dependent data quality characteristics. The standard also recognises that the data quality characteristics can have different priorities in different cases. Furthermore, the standard allows that, depending on the use case, some characteristics can be excluded or that new ones can be added. This has been a common practice in software engineering, where various researchers have developed domain-specific software quality models based

on the generic ISO 25010 software quality model [10] by introducing new quality characteristics and sub-characteristics. In order to extend an existing quality model, researchers usually use methods that are based on a top-down approach, such as the ones presented by Franch and Carvallo [14] or by Behkamal et al. [2], or methods that are based on a bottom-up approach, such as the one described by Radulovic et al. [11]. Furthermore, Dromey suggests that both approaches can be important for building quality models [15].

The top-down approach for extending quality models starts from an existing generic quality model, i.e., adopts an existing quality model and defines new quality characteristics and sub-characteristics. It then continues with the definition of quality measures for measuring these quality sub-characteristics and of the relationships between these measures. On the other hand, a bottom-up approach starts by defining a hierarchy of quality measures and the relationships between them. These relationships are typically defined in terms of the formulas used for the calculation of these measures. Once the quality measures are defined, a hierarchy of quality sub-characteristics and characteristics is constructed, which are then aligned to quality characteristics from an existing generic quality model.

### 2.2. Linked Data quality specification

To the best of our knowledge, there is no clearly defined quality model for Linked Data. However, various efforts over the years have contributed to the understanding and quality specification of Linked Data. These efforts are mostly concentrated on quality evaluation of Linked Data datasets, as well as on the theoretical aspects of Linked Data quality.

Unlike the ISO standards, the literature describes the quality of Linked Data using different terminology. Characteristics of a dataset are called *dimensions* by Zaveri et al. [16], which are assessed with quality *indicators*. Furthermore, according to Bizer and Cyganiak [17], a procedure for measuring a data quality *dimension* is called *indicator, measure*, or *metric*.

Zaveri et al. [16] provide in their work a comprehensive review of the various efforts related to Linked Data quality specification and evaluation, with a comprehensive classification of quality dimensions and metrics found in the literature. It presents 69 metrics grouped into 18 quality dimensions extracted from 30 Linked Data quality related papers published from 2002 to 2014. The conclusion that can be made from their work is that the efforts described are quite diverse in terms of the dimensions evaluated and the calculated measures. Since 2014, Behkamal et al. published a set of 10 metrics to assess 18 quality issues they identified [18]. Besides, Albertoni et al. proposed a metric for estimating the multilingual information gain through a linkset [19].

There are several tools in the literature that support quality assessment of Linked Data [17,20,21,22,23,24,25,26,27,28,29,30,31,32,33]. These tools vary in their scope for assessing quality. On the one hand, there are tools that are focused on assessing quality along one dimension, such as trust (TrustBot [21], Trellis [20], tSPARQL [22]) or interlinking (LinkedQA [27], LiQuate [32]). On the other hand, there are tools that are frameworks for generic quality assessment such as Luzzu [34], RDFUnit [29], or Sieve [28]. Furthermore, there is a set of tools that allows an exploratory inspection of quality issues such as ProLOD [23], LODStats [24], ABSTAT [25], and Loupe [26] that mainly use different statistics and patterns extracted from data. The quality model proposed in this paper can be used by all these tools.

With respect to the state of the art described in this section, the quality model presented in this paper contributes in several ways: i) it provides a unified reference for Linked Data quality, by describing a set of quality characteristics and measures, together with definitions and formulas; ii) it introduces a hierarchy of quality measures as described in the beginning of Section 2.1, relying on the current state of the art and extending it with the mentioned hierarchy in the case of each measure; and iii) it describes important details of each quality measure, such as quality aspects related to Linked Data, units of measurement or measurement scales.

### 2.3. Existing quality specification and assessment ontologies

A number of ontologies, related to quality specification and assessment have been developed in the Semantic Web / Linked Data field to this date. These ontologies represent quality meta-models that can be used for describing quality-related information found in quality models. Next, we give a brief overview of the most relevant ontologies in the context of Linked Data quality and quality modelling in general.

The *Quality Model* ontology[2] (QMO) defines a generic ontology for representing quality models and

[2] http://purl.org/net/QualityModel#

their resources in any particular domain; it can be used as a generic ontology for specifying quality. The main classes of QMO are based on the ISO standards and it also uses the ISO terminology. Apart from the classes for describing quality mesures (base measures, derived measures and indicators) and quality characteristics, QMO provides the means to describe units of measurement and measurement scales for each quality measure. Also, QMO provides a number of properties for describing relationships between the quality measures. In the context of the conceptual model, QMO provides means to describe the data related to the Linked Data quality model, since it is intended to be a general ontology.

The *Evaluation Result* ontology[3] (EVAL) defines a generic ontology for representing results obtained in an evaluation process; it is a ontology for representing the results of a quality assessment and is an extension of QMO. The classes of this ontology provide means for capturing the specific values obtained in an evaluation process and for relating such values to quality measures and evaluated subjects (e.g., a specific dataset). They also provide the possibility to describe measurement scales and units of measurement for the obtained values, as well as inputs in the evaluation. In the context of the conceptual model, EVAL provides means to describe quality values and evaluated datasets.

The *Data Quality Management Vocabulary* (DQM) [35] is an ontology for representing data quality management activities in Semantic Web architectures. The main concepts of this ontology include data quality requirements, i.e., quality-relevant expectations on data and data quality reports with data quality scores. The goal of the DQM ontology is to automate the creation of quality reports based on the data quality requirements defined using the DQM ontology with a data quality score based on each requirement. Further, by using Semantic Web technologies it aims to do automated consistency checking between a set of data quality requirements and also to facilitate the exchange of both data quality requirements and data quality results. Unlike the conceptual model, DQM provides classes that are specifically related to some concrete aspects of quality (e.g., a class for denoting that a property is missing in a dataset).

The Dataset Quality Ontology (daQ) [36] is an ontology for representing the quality of a dataset. The ontology defines the classes related to quality category, dimension, and metric, and several properties to define the relationships between these classes. The classes and properties in daQ are defined as abstract and, therefore, they are not directly used. Instead, the intended use of this ontology implies the creation of specific classes and properties defined as subclasses and sub-properties of those defined in daQ. This means that, unlike in QMO and EVAL, where the elements such as measures and characteristics are defined as instances, when using daQ these elements are mainly defined as classes. In the context of the conceptual model, daQ is equivalent to QMO, with the difference of the usage of different terminology.

The Data Quality Vocabulary[4] (DQV) is an ontology for representing the quality of datasets that is being developed by the W3C. Similarly as daQ, and unlike QMO and EVAL, DQV is an ontology specifically developed having in mind datasets. Currently, DQV provides classes and properties for capturing information about quality categories, dimensions and metrics of a dataset, as well as about quality certificates, standards and provenance related to a dataset. However, at this point in time DQV is still under development and changes to the current design can be expected in the future. In the context of the conceptual model, DQV tends to provide the means for capturing both the details about quality (i.e., characteristics and measures) and about quality values (results of evaluation). Furthermore, although DQV is specifically designed for datasets, it does not provide the means to describe some specific aspects of Linked Data.

With respect to the state of the art described in this section, this paper contributes with the extension of existing ontologies in order to enable capturing information related to Linked Data which is not covered by the existing ontologies, as well as with bringing existing ontologies under unique umbrella by connecting their semantically related concepts.

## 3. Quality model for Linked Data

This section describes a quality model for Linked Data and how it was defined using the bottom-up methodology proposed by Radulovic et al. [11]. The starting point for the definition of the quality model was the state of the art in Linked Data quality assess-

[3] http://purl.org/net/EvaluationResult#

[4] http://www.w3.org/TR/vocab-dqv/

ment and specification, and in particular the work done by Zaveri et al. [16]. Since the quality model presented in this section describes a classification of quality measures (i.e., base measures, derived measures and indicators), we have decided to adopt the terminology as described by the ISO standards.

The work by Zaveri et al. does not specify any base measures, derived measures nor indicators per se, and it does not specify a classification of quality measures, as the mentioned one that has been adopted in our work. Rather, Zavery et al. define metrics which in different cases are related to different types of measures in our classification. In some cases, a metric described by Zaveri et al. appears in our quality model as a base measure, and in these cases we have used this measure in order to define derived measures and/or indicators that do not appear in the work by Zaveri et al. In other cases, a metric described by Zaveri et al. appears in our quality model as an indicator, and in these cases we have defined the base and/or derived measures that are used to calculate these indicators.

### *3.1. Quality model overview*

Data quality is a multifaceted concept and different quality measures in a quality model can be related to different aspects of quality. The Linked Data quality model that we propose encompasses the different aspects of Linked Data quality, as illustrated in Figure 1, and all quality measures are classified according to these aspects.

The aspects of Linked Data quality can be categorized into two main groups: i) aspects related to inherent data quality; and ii) aspects related to the infrastructure that is used for serving the data. For instance, on the one hand, quality characteristics such as the accuracy of facts represented in Linked Data or the completeness of a dataset are intrinsic to the data themselves. On the other hand, quality characteristics such as response time of a Linked Data resource or support for different media types through content negotiation depend more on the capabilities of the server applications and hardware devices that are used to serve the data. The same dataset hosted in different infrastructures and system configurations could have different levels of quality.

The motivation for identifying the aspects associated with Linked Data and annotating the quality measures with the relevant aspect is to help quality evaluators to select the most relevant metrics depending on their quality requirements and to have a better understanding about them. For instance, Linked Data providers will know that any measures related to infrastructure will have to be re-evaluated if the provider changes the Linked Data servers and other publishing infrastructure.

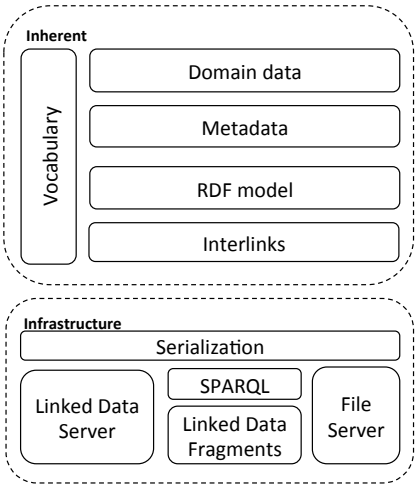


Fig. 1. Aspects of Linked Data quality

The **Domain data** aspect refers to the concrete facts contained in the dataset. The quality of the domain data can be measured with respect to different quality characteristics such as accuracy, completeness, or timeliness. For instance, in a dataset about cities of Spain, a fact such as that the city of Madrid has a population of 3,165,235 inhabitants could exist. This fact can be assessed to verify that it is correct and not outdated (i.e., it reflects the situation in the real world) so that it has sufficient quality for a given use case. Further, if the use case requires the dataset to have information about all the cities in Spain the completeness can be verified by checking the dataset against a list of all Spanish cities from an official dataset from the Spanish government.

The **Metadata** aspect refers to the information that provides the context and additional information about the domain data or conditions on the usage of data. The quality of metadata can be measured with respect to quality characteristics such as compliance or trustworthiness. For instance, for a dataset to be fit for a given use case, the data consumer may require to know about the provenance information, such as the provider of the data and its source, or the license information, so that a consumer can evaluate whether she can legally use the data for a concrete use case in commercial settings.

The **Vocabulary** aspect refers to the selection of vocabulary (ontology) terms representing both domain data and metadata. The ontologies used can be evaluated with respect to quality characteristics such as interoperability, conciseness, or understandability. For instance, if the common standard ontologies such as FOAF, DC Terms, SKOS, PROV that facilitate interoperability are used and if the ontologies used have dereferenceable identifiers with appropriate documentation that increase their understandability.

The **RDF model** aspect refers to different designs that are taken into account when modelling the domain data and metadata as RDF data. The quality of the RDF data model can be measured with respect to quality characteristics such as representational conciseness or the irregular use of RDF features such as collections, containers, or reification. For example, the use of collections without valid properties such as first and rest properties can violate the representational conciseness and affect performance.

The **Interlinks** aspect refers to exposing the RDF data as Linked Data and to linking the data to other relevant data so that consumers can discover more related data with the follow-your-nose approach. The quality of Linked Data interlinking can be measured with respect to quality characteristics such as accessibility or representational conciseness. For instance, in addition to proper RDF modelling, data can be made more useful by applying the Linked Data principles so that entities are named using HTTP URIs, useful information is provided when those URIs are looked up, and the entities are linked to the other related entities.

The **Infrastructure** aspects identify the approaches that are commonly used to expose Linked Data, such as Linked Data servers, SPARQL and Linked Data Fragments endpoints, or RDF dumps in file servers. The quality of the infrastructure can be measured with respect to quality characteristics such as availability, performance, or compliance. Linked Data servers such as Pubby or Elda are used to expose Linked Data as dereferenceable Linked Data resources via the HTTP protocol and they are associated with properties such as response time, throughput, or the different media types supported.

The **Serialization** aspect refers to the representation of RDF data in some RDF serialization format such as Turtle, RDF/XML, JSON-LD, N3, N triples, N quads, or Trig. In the case in which Linked Data are available as bulk download, the serialization could be a compressed archive such as a zip or a tarball archive. The quality of the representation can be measured with respect to quality characteristics such as syntactic accuracy, interoperability, or versatility. For instance, if an RDF document is serialized using RDF/XML the serialized representation of the model should follow all the syntactic rules defined by the RDF/XML Syntax specification [37].

Apart from this categorization in terms of aspects, Linked Data quality encompasses different levels of RDF concepts, including: i) IRIs/Blank nodes/Literals; ii) individual statements (i.e., triples); iii) RDF graphs; and iv) RDF datasets as a whole. These different levels are shown in Figure 2, and each quality measure in the Linked Data quality model is related to one of these levels.

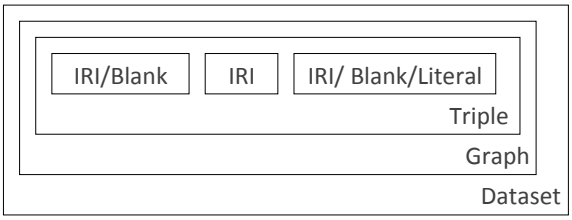


Fig. 2. Data model (RDF) levels

The following sections present how the bottom-up approach has been used in order to define the Linked Data quality model. Due to space reasons we cannot present the quality model definition in complete details; therefore, for illustration purposes we present the outcomes of each step in the bottom-up methodology related to only a subset of the quality model. For each quality measure presented, we emphasize the aspect it measures (Figure 1) and the level on which the measure is calculated (Figure 2).

The complete overview of our quality model can be found at the Linked Data quality model wiki[5].

### *3.2. Identification of base measures*

The first step in building the Linked Data quality model was to identify the base measures. In total, we have defined 89 base measures, out of which 44 directly come from the survey by Zaveri et al. [16], and 45 have been newly introduced in our work. These measures are described in detail in the quality model wiki and are classified according to the quality characteristic that they are related to.

[5] http://delicias.dia.fi.upm.es/LDQM

With respect to the RDF data model levels (Figure 2), each base measure can be related to any of such levels. For example, base measures related to IRIs include:

- *IRI dereferenceability*. Whether an IRI is derefenceable or not. This measure is related to the *Infrastructure* aspect (Figure 1), and possible values for this measure are *true* and *false*.
- *Short IRI*. Whether an IRI is short or not. This measure is related to the *RDF model* aspect, and possible values for this measure are *true* and *false*.

In some cases, base measures can be related to a triple in a graph or a dataset. An example of such a base measure includes:

- *Subject dereferenceability*. Whether a subject in a triple is dereferenceable or not. This measure is related to the *Infrastructure* aspect, and possible values for this measure are *true* and *false*.

Depending on the context of their use in the evaluation, i.e., in the calculation of derived measures or indicators, some base measures can be related both to the IRI or triple level. An example of such a measure is:

- *Subject types*. A list of classes that an instance represented by an IRI (or a subject in a triple) is type of. This measure is related to the *Vocabulary* aspect, and possible values for this measure are any ontology class.

The previous base measures are related to the IRIs or triples in a dataset, and these measures alone can be useful in the process of error correction and data repair. Finally, an example of base measures that are related to a graph or a dataset include:

- *Number of interlinked subjects*. The total number of all subjects in a dataset that are linked.
- *Number of subjects*. The total number of subjects in a dataset.
- *Number of IRIs*. The total number of IRIs in a dataset.
- *Number of triples*. The total number of triples in a dataset.

*Number of interlinked subjects* base measure is related to the *Interlinks* aspect and the rest of the base measures are related to the *Domain data* aspect and the value for each of these measures can be any natural number.

### 3.3. Identification of derived measures

In this step, the previously defined base measures are used in combination with the inputs in the evaluation (e.g., an ontology) in order to define derived measures. Similarly as in the case of base measures, derived measures for Linked Data datasets can be related to different RDF data model levels. In total, 23 different derived measures have been defined, which are described in detail in the quality model wiki, and are classified according to the quality characteristic that they are related to. In total, 6 derived measures come directly from the survey by Zaveri et al., and 17 derived measures have been newly introduced in our work.

By analysing all the defined base measures, we identified several patterns of defining derived measures that will be used by the quality model presented in this paper.

Some patterns are related to the aggregation of base measures in a higher RDF data model level. When a base measure of a lower level such as an IRI is measured, it can be aggregated to come up with derived measures that are associated with a higher level such as triples, RDF graphs or datasets. An example of such aggregation is:

- *Number of dereferenceable IRIs*. The total number of dereferenceable IRIs. This measure is related to the triple, graph or dataset levels, and is defined using *IRI dereferenceability*, which is an IRI level base measure. Furthermore, this derived measure is related to the *Infrastructure* aspect, and possible values for this measure are any natural number.
- *Number of dereferenceable subjects*. The total number of dereferenceable subjects. This measure is related to the graph or dataset levels, and is defined using *Subject dereferenceability*, which is an IRI level base measure. Similarly as in the previous case, this derived measure is related to the *Infrastructure* aspect, and possible values for this measure are any natural number.

Other patterns are related to the interpretation or the combination of base measures that are related to the same RDF data model level. Sometimes, it is possible to combine or interpret various base measures in order to define new derived measures that are on the same RDF data model level as the base measures used for their definition. Examples of such derived measures include:

– *Disjoint classes*. Whether an instance represented with a specific IRI is an instance of disjoint classes. This measure is related to the IRI level and is defined using *Subject types*, an IRI-level base measure. Furthermore, this derived measure is related to the *RDF model* and *Vocabulary* aspects, and possible values for this measure are *true* and *false*. In order to calculate this derived measure, an ontology is needed as an input in the evaluation in order to examine ontology classes and axioms and to obtain information about disjoint clases to be compared with types of the observed instance in a dataset.
– *Domain consistency*. Whether the type of a subject in a specific triple is consistent with the domain of a property of a triple. This measure is related to the triple level, and is defined using *Subject types*, a triple-level base measure. Furthermore, this derived measure is related to the *RDF model* and *Vocabulary* aspects, and possible values for this measure are *true* and *false*. In order to calculate this derived measure, an ontology is needed as an input in the evaluation in order to examine properties and to obtain information about property domain to be compared with the type of the observed triple in a dataset.

### 3.4. Identification of indicators

In this step, we have defined 124 quality indicators by combining base and derived measures, out of which 32 have been newly defined in our work. Usually, indicators are defined using the base or derived measures on a lower data model level, and they are themselves related to the higher data model levels. Similar as in the case of base and derived measures, an indicator can be related to different RDF data model levels. All the indicators are described in detail in the quality model wiki and are classified according to the quality characteristic that they measure.

From the previously specified derived measures, the following indicators were obtained:

– *Average IRI dereferenceability*. The average number of dereferenceable IRIs. This measure can be related to the triple, graph, or to the dataset levels and, furthermore, it is related to the *Infrastructure* aspect.
– *Average subject dereferenceability*. The average number of dereferenceable subjects. This measure can be related to the graph or dataset levels and, furthermore, it is related to the *Infrastructure* aspect.
– *Average disjoint classes*. The average number of instances of disjoint classes. This measure can be related to the graph or dataset levels and, furthermore, it is related to the *RDF model* and *Vocabulary* aspects.
– *Average domain consistency*. The average number of triples in which the subject is consistent with the property domain. This measure can be related to the graph or dataset levels and, furthermore, it is related to the *RDF model* and *Vocabulary* aspects.

In some cases, quality indicators can be derived based only on base measures. An example of such indicators include:

– *Instance interlinking*. The average number of interlinked instances. This measure can be related to the graph or dataset levels and, furthermore, it is related to the *Domain data* aspect.
– *Average short IRIs*. The average number of short IRIs. This measure can be related to the graph or dataset levels and, furthermore, it is related to the *RDF model* aspect.

All previously specified indicators have a ratio scale with values ranging from zero to one hundred, expressed in percentage.

Finally, some indicators have been directly defined without the need for base or derived measures. Examples of such indicators include

– *SPARQL 1.1 support*. Whether a dataset SPARQL endpoint supports the SPARQL 1.1 language. This indicator is related to the dataset level, and to the *Infrastructure* aspect. The possible values for this indicator are *true* and *false*.

The base measures, derived measures, and quality indicators in the Linked Data quality model are based on the results of the survey by Zaveri et al. [16], i.e., on the state of the art in Linked Data quality specification and assessment. However, the quality model proposed in this paper describes a classification of measures in greater detail and it also defines a higher number of measures.

### 3.5. Specification of relationships between measures

After the set of base measures, derived measures and indicators was defined, we have specified the formal

relationships between these measures in terms of the formulas used for their calculation, which is a practice that is not always followed in the current state of the art.

Next, we present the formulas for the measures described in previous sections, in those cases in which a formalization through a formula applies.

The formula for *Short IRI* (1) defines whether an IRI is short with respect to some predefined threshold

$$\text{IRI.length} < \text{threshold} \tag{1}$$

Formulas for *Number of dereferenceable IRIs* (2) and *Number of dereferenceable subjects* (3) calculate the total number of dereferenceable IRIs and subjects in a dataset, respectively.

$$\#\text{ different IRIs where (IRI dereferenceability = true)} \tag{2}$$

$$\#\text{ triples where (Subject dereferenceability = true)} \tag{3}$$

Formulas (4) and (5) calculate the *Disjoint classes* and *Domain consistency* derived measures, respectively.

$$\text{subject types} \not\subseteq \text{disjoint classes} \tag{4}$$

$$\text{subject types} \subseteq \text{property domain} \tag{5}$$

Similarly as in the case of derived measures, the following formulas have been defined for indicators: *Average IRI dereferenceability* (6), *Average subject dereferenceability* (7), *Average disjoint classes* (8), *Average domain consistency* (9), *Instance interlinking* (10), and *Average short IRIs* (11).

$$\frac{\#\text{ dereferenceable IRIs}}{\#\text{ IRIs}} \times 100 \tag{6}$$

$$\frac{\#\text{ dereferenceable subjects}}{\#\text{ subjects}} \times 100 \tag{7}$$

$$\frac{\#\text{ IRIs where (disjoint classes = true)}}{\#\text{ IRIs}} \times 100 \tag{8}$$

$$\frac{\#\text{ triples where (domain consistency = true)}}{\#\text{ triples}} \times 100 \tag{9}$$

$$\frac{\#\text{ interlinked subjects}}{\#\text{ subjects}} \times 100 \tag{10}$$

$$\frac{\#\text{ IRIs where (Short IRI = true)}}{\#\text{ IRIs}} \times 100 \tag{11}$$

### 3.6. Alignment with the quality model

The last two steps of the followed bottom-up method suggest the definition of domain-specific quality subcharacteristics and their alignment with an existing quality model. In the case of Linked Data quality, as shown by Zaveri et al. [16], a large number of measures described in the survey are classified according to various dimensions. Therefore, for the Linked Data quality model we have decided to rely on the classification provided by Zaveri et al. and, starting from this classification and from the quality indicators identified in Section 3.4, we have identified the ISO 25012 quality characteristics that can be measured with the mentioned indicators.

The quality characteristics related to the indicators described in Section 3.4 include:

- *Accessibility*. The degree to which data can be accessed in a specific context of use, particularly by people who need supporting technology or special configuration because of some disability [5]. It can be measured using *Average IRI dereferenceability* and *Average subject dereferenceability*.
- *Availability*. The degree to which data has attributes that enable it to be retrieved by authorized users and/or applications in a specific context of use [5]. It can be measured using *SPARQL 1.1 support*.
- *Completeness*. The degree to which subject data associated with an entity have values for all expected attributes and related entity instances in a specific context of use [5]. It can be measured using *Instance interlinking*.

– *Compliance*. The degree to which data has attributes that adhere to standards, conventions or regulations in force and similar rules relating to data quality in a specific context of use [5]. It can be measured using *Average short IRIs*.
– *Consistency*. The degree to which data has attributes that are free from contradiction and are coherent with other data in a specific context of use [5]. It can be measured using *Average disjoint classes* and *Average domain consistency*.

The Linked Data quality model includes fifteen quality characteristics: accessibility, accuracy, availability, completeness, compliance, confidentiality, consistency, credibility, currentness, efficiency, precision, portability, recoverability, traceability, and understandability. Quality measures defined in the quality model cover twelve of these quality characteristics; as to this date no quality measures have been defined in the state of the art for three quality characteristics. The quality characteristics that do not have any quality measures associated are confidentiality, precision and recoverability.

Figure 3 presents the base measures, derived measures, indicators and quality characteristics presented as an example in this section, together with the references to the formulas described in Section 3.5. For a better visibility, some measures are repeated on the figure, and they are marked with the * sign. Due to space reasons, the inputs in the evaluation (*disjoint classes* and *property domain*) that are used for the calculation of some derived measures (i.e., *Disjoint classes* and *Domain consistency*) are not shown.

Table 1 shows all the quality characteristics identified in the Linked Data quality model, together with the indicators that can be used for their measurement.

## 4. Ontological representation of the quality model

This section discusses how the quality model presented in the previous section can be represented in RDF using existing ontologies described in Section 2.3 and, furthermore, it presents a set of extensions to those existing ontologies so that the quality metrics and their measures can be described with fine-grained details.

### 4.1. Conceptual model

Figure 4 shows the conceptual model of the ontology for representing the Linked Data quality model, using the terminology adopted in this paper. The model describes a hierarchy of quality measures and quality characteristics related to a quality model, with important information such as measurement scales and scoring functions (i.e., the formulas for the calculation of values for a specific quality measure). Each quality measure is calculated using a specific technique (which can be automatic, semiautomatic or manual), it can be subjective or objective, has a specific duration, and can be used for obtaining some other measure. Furthermore, when performing an evaluation, a quality value related to a specific quality measure and an evaluation subject (e.g., a dataset) is obtained.

### 4.2. Extensions to existing ontologies

The ontologies that could be used for the representation of quality specification and assessment of Linked Data, as described in Section 2.3, are either general or related to all types of data. For example, although DQV is a lightweight ontology that can suite the needs of Linked Data, it is defined to fit all types of data and it does not cover some aspects that are specific to Linked Data quality (Section 3). In this section we describe an extension of the existing ontologies, in order to cover the Linked Data specificities which is one of the requirements of our scenario of capturing quality-related information of Linked Data; this extension mainly relies on DQV, since it is expected to become the W3C standard for representing data quality.

Figures 5 and 6 present the proposed extension of the current ontologies for quality representation and assessment, adapted to the domain of Linked Data. The classes already described in the existing ontologies (Section 2.3) are presented in white boxes, together with their namespaces, while the extensions are represented with grey boxes; new properties are marked in bold.

Quality values obtained in an evaluation are represented with the *dqv:QualityMeasure* class, with equivalent classes being *eval:QualityValue* and *daq:Observation*. Quality measures are represented with the *dqv:Metric* class, with equivalent classes being *qmo:QualityMeasure* and *daq:Metric*. Furthermore, QMO classes representing base measures, derived measures and quality indicators are also reused. Quality characteristics are represented with the *dqv:Dimension* class, with equivalent classes being *qmo:QualityCharacteristic* and *daq:Dimension*. Evaluation processes are represented with the *eval:Evaluation* class, while datasets are represented with the *dcat:Dataset* class

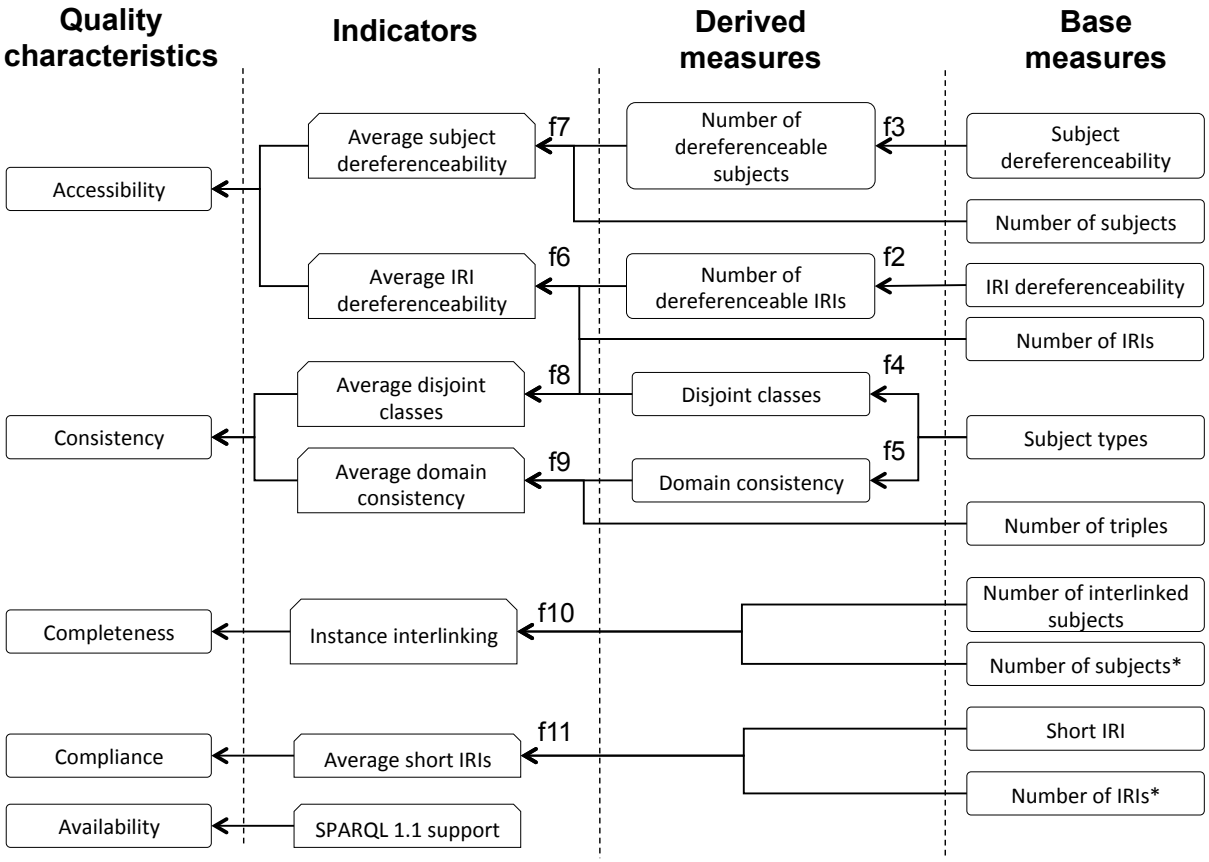


Fig. 3. Subset of the Linked Data quality model

from the well-known DCAT ontology[6]. Furthermore, some classes for representing concepts such as time instants, time intervals, measurement scales, and units of measurement are reused from the W3C Time ontology[7] (*time*) and from the ontology of units of measure[8] (*om*).

By extending the QMO ontology, relationships between quality measures are incorporated into the proposed extensions. QMO defines relationships which denote that one measure can be used for obtaining some other measure, or that increasing the value of one measure implies increasing or decreasing the value of some other measure.

In some cases for certain measures and their scales a *qmo:RankingFunction* property can be specified in an objective way, denoting whether in the case of numerical results obtained for such measures higher or lower values are more desirable (e.g., for precision or number of dereferenceable URIs, it is clear that a higher value is more desirable).

The extensions of the presented ontology tend to cover additional information related to Linked Data quality specification and assessment which can be important for easier interpretation, benchmarking, interchange, and understanding of quality measures and evaluation results. The *QualityAspect* class describes the aspect of Linked Data quality the measure is related to (Figure 1), while *Granularity* class describes the evaluated RDF level (Figure 2). These two concepts are specifically related to Linked Data quality. The motivation for including these two concepts is to allow quality evaluators to select and filter the most relevant and applicable metrics depending on their use case. For instance, if there was a change in the infrastructure or in the transformation process, evaluators can select the aspects that have the most impact from those changes and re-evaluate them. Similarly, depending on whether one is evaluating a single triple, a graph, or a dataset

[6] http://www.w3.org/TR/vocab-dcat/
[7] http://www.w3.org/TR/owl-time/
[8] http://www.wurvoc.org/vocabularies/om-1.8/

Table 1
Linked Data quality model characteristics and indicators.

| **Characteristic (ISO 25012)** | **Indicators** |
|---|---|
| Accessibility | Average IRI dereferenceability, Average subject dereferenceability, Average predicate dereferenceability, Average object dereferenceability |
| Accuracy | Average automatic validation errors, Average crowdsourcing validation errors, Average datatype syntax errors, Average syntactic rules syntax errors, Average RDF pattern errors, Average ill-typed literals, Average datatype compatibility, Average distance-based outliers, Average deviation-based outliers, Average distribution-based outliers, Average triple correctness, Average crowdsourced incorrect triples, Average misspelled literals, Average inaccurate labels, Average correct classification, Average property misuse, Average invalid rules, Average entity mismatch |
| Availability | SPARQL support, SPARQL 1.0 support, SPARQL 1.1 support, RDF dump, Average IRI RDF description, Average content type IRIs, Average content negotiation support, Average accept header support, Average sustainable IRIs, Multiple serialization formats, Multiple languages |
| Completeness | Interlinking degree, Clustering coefficient, Centrality, Linked Data mappings, In-links, Average sameAs linked, Average blank nodes, Number of entities, Number distinct properties, Average undefined classes, Average undefined properties, Average undefined objects, Blank nodes use |
| Compliance | Average correct HTTP redirect, Average LDP GET support, Average LDP PUT support, Machine-readable licence, Human-readable licence, License propagation, Average HTTP IRIs, Average short IRIs, Average IRI uniqueness |
| Consistency | Average stable IRIs, Average number of inconsistent functional dependence subjects, Average disjoint classes, Average misplaced classes, Average misplaced properties, Average misused datatype properties, Average misused object properties, Average deprecated subjects, Average deprecated properties, Average invalid inverse functional values, Average ontology hijacking, Average negative dependent properties, Average geometric violation, Average domain consistency, Average range consistency, Average axiom violations, Schema completeness, Property completeness, Population completeness, Instance interlinking, Average mapped types |
| Credibility | Document digital signature, SPARQL digital signature, Average graph digital signature, Author provenance, Contributors provenance, Publisher provenance, Dataset sources provenance, Dataset ranking, Crowdsourcing relevance, Provenance-based trust, Opinion-based thrust, Social networks trust, Average facts trust, Blacklisted, Authority, Content-based trust, Metadata-based trust, Average one-path trust, Average many-paths trust, Decision network trust, List trust, Publisher trust, Association trust, Average dataset rating |
| Currentness | Dataset freshness, Datasource freshness |
| Efficiency | RDF dump compression, Average slash IRIs, Low latency, High throughput, Response scalability, Average IRI caching, Average RDF primitives |
| Portability | Terms reuse, Vocabulary reuse |
| Traceability | Provenance |
| Understandability | SPARQL service description, Average internal redundant properties, Average external redundant properties, Average label unambiguity, Average elements labelling, Dataset metadata, IRI pattern, Regular expression, SPARQL examples, Vocabulary list, Mailing lists presence, Data interpretability |

she can select the most relevant metrics using the granularity concept.

Some extensions, although could be related to data quality in general, could also carry valuable information for Linked Data. These include the information on how often a measure has to be assessed (*AssessmentFrequency*), whether a measure is dependent on the system (*isSystemDependent*), the period of time during which the result for the measure is valid (*TemporalValidity*), the technique used in the evaluation for assessing a measure (*AssessmentTechnique*), whether an assessment technique is subjective or objective (*isSubjective*), the expected duration of the assessment (*ExpectedDuration*), and whether an evaluation is done automatically, semi-automatically or manually (*AutomationLevel*).

### 4.3. Representing the Linked Data quality model

The ontology presented in the previous section has been implemented in OWL and is available online[9].

The Linked Data quality model presented in this paper has been described in RDF by using the ontology extension. The description is available online[10] and it can be reused by the tools that utilize the quality model presented in this paper, which can bring consistency

[9] http://www.linkeddata.es/ontology/ldq/
[10] http://linkeddata.es/resource/ldqm/

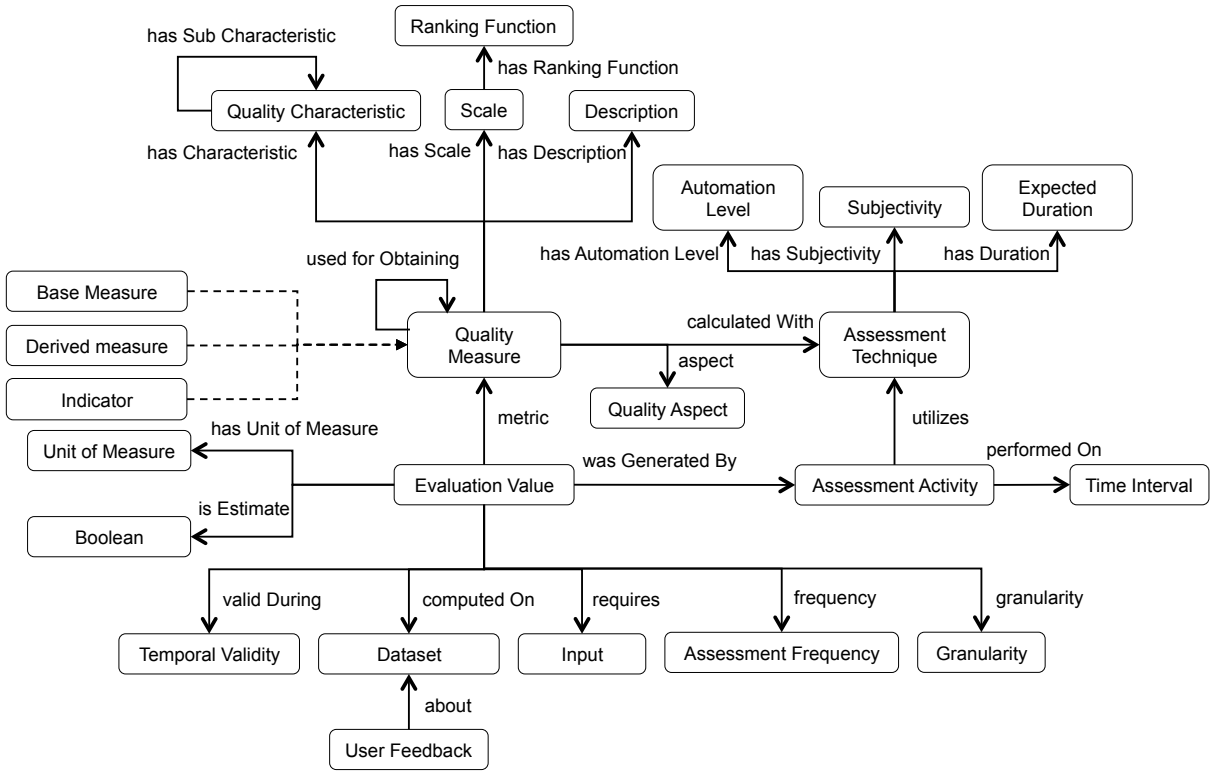


Fig. 4. The conceptual model

in results representation across various evaluation efforts. To this extent, the provided RDF representation of the Linked Data quality model can be significant in the production, interchange, and consumption of quality evaluation data.

## 5. Use Case

This section presents a practical use case of the usefulness of the Linked Data Quality Model discussed in the paper. The section presents a command-line tool, LD Sniffer, which allows users to assess the quality of Linked Data resources in a dataset such as DBpedia. Given a set of resource URIs each identified by an IRI, for example, "http://es.dbpedia.org/resource/Madrid", the tool can dereference the corresponding RDF graph and assess each IRI in the graph using the Linked Data Quality Model. The LD Sniffer tool[11] is an open source project under the Apache 2.0 license.

The motivation of presenting this use case is to discuss the usefulness of the model with relation to how a quality model could help the design and implementation of such a tool, how the quality model could improve the understanding of the metrics with explicit semantics and reduce the opportunity for ambiguity and misinterpretations, and finally how the quality model and its formal representation could improve the comparison of the assessment results generated by tools and be used for further tasks such as selection of Linked Data resources with a sufficient quality or recommendation of Linked Data resources.

### *5.1. Use of the quality model*

The main functionality of the LD Sniffer tool is to generate a quality assessment for a set of given Linked Data resources. Thus, the tool will take a set of URIs of Linked Data resources as the input and generate an output as a set of metrics that provide indications about the quality of the resource. With the aforementioned use case, the first step of the quality assessment tool developer is to identify how the quality of a Linked Data resource can be measured and what metrics can be used to measure quality. The quality model presented in this paper provides the necessary guidance

[11] https://github.com/nandana/ld-sniffer

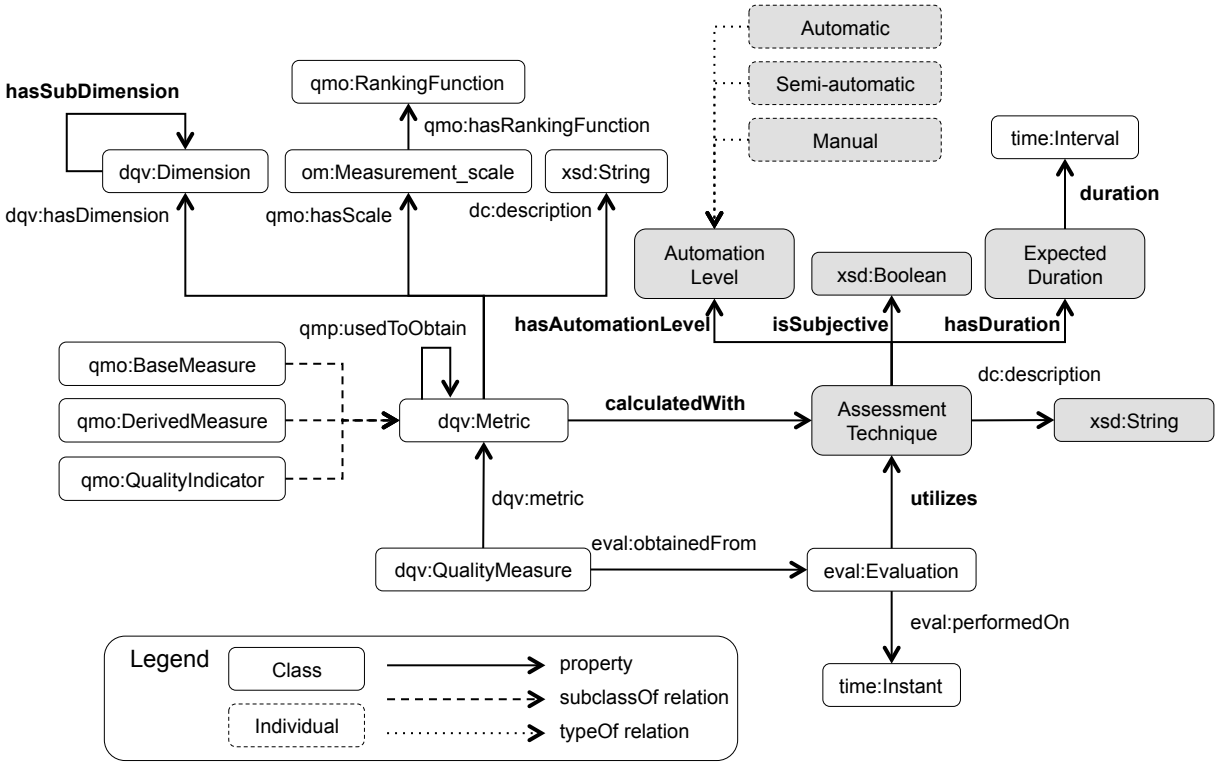


Fig. 5. The ontology extension (I)

on that by providing an overview of the different aspects that have to be taken into account such as domain data, metadata, vocabularies, interlinking, and infrastructure.

Having an understanding about the different quality aspects allows the assessment tool developers to decide and focus on the aspects that they want to assess. For instance, the initial version of LD Sniffer is planned to be a domain-agnostic general purpose quality assessment tool so it will not focus on quality characteristics such as the accuracy or completeness of facts. It will focus on generic aspects that apply to Linked Data and are described in the previous section such as *Metadata*, *RDF model*, *Serialization*, and *Infrastructure*. Later on, the developers of LD Sniffer could decide to add the aspects that they did not consider such as *Domain data* or *Vocabularies* used by linking the tool to a domain knowledge base which can verify the accuracy of the facts or the most appropriate ontologies to be reused. Understanding those different aspects allows the Linked Data tool assessment developers to develop their roadmaps using the quality model as a guidance. In the case of the LD Sniffer tool, the developers are mostly interested in the infrastructure aspect which is highly related to the accessibility dimension.

The second step is to decide which metrics to be used in the quality assessment. The quality model provides a set of metrics that can be used to measure different quality characteristics. In the model, those metrics are described using base measures, derived measures, and indicators. These metrics in the quality model provide insights about which metrics can be implemented by the tool for a given aspect. Furthermore, the base measures provide information to the tool developers on which metrics can be directly evaluated with the subject of the evaluation and the derived measures specify how they can be combined to formulate other measures. Having a Linked Data quality model with a well-defined set of metrics allows the developers to implement those metrics in a standard manner. As the metrics in the quality model are defined as Linked Data with global identifiers and explicit semantics, all tools interpret the metrics in the same way. This allows the LD Sniffer tool to advertise the metrics that it uses in an unambiguous manner. The separation of base measures, derived measures, and indicators also helps the assessment tool developers in

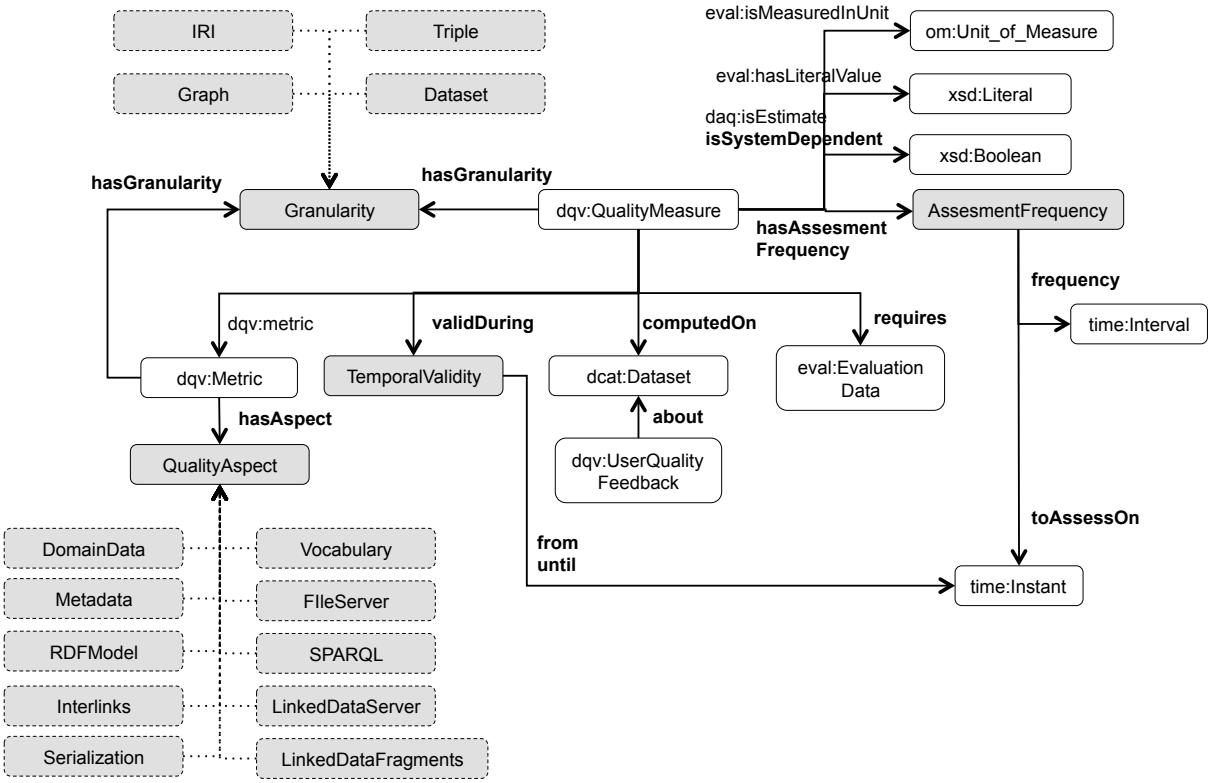


Fig. 6. The ontology extension (II)

how to design the tool. For instance, indicators are important at the time of decision making and, thus, they could be highlighted in the quality assessment results. The Linked Data Quality Model allows finding relevant metrics for a given use case, either via the Linked Data Quality Model wiki[12] or using the its SPARQL endpoint[13] .

Once the set of metrics are defined and implemented, the next step is the representation of quality results. Having global identifiers for the quality metrics helps in representing the quality results. Dataset validation tools that produce evaluation results that are described using the model proposed in this paper and reference to the Linked Dada representation of the quality model have the advantage of easier publication of structured results, as well as of reusability of such results.

Figure 7 shows an example of some evaluation results that are described according to the ontology proposed in this paper (*ldq*), and that reference the RDF representation of the Linked Data quality model (*ldqm*).

### 5.2. Evaluation of DBpedia resources

The proposed quality model and the LD sniffer tool were used to perform an assessment of the accessibility of a subset of DBpedia resources. The details of the evaluation are described using LDQM and are available as dereferenciable Linked Data along with a SPARQL endpoint[14] for querying and as a downloadable RDF dump[15]. The evaluation is focused on the accessibility quality characteristics of the model and uses 8 base measures, 4 derived measures, and 4 indicators as described in the quality model. 1 million IRIs in more than 100K resources were assessed to calculate the aforementioned results.

Figure 8 shows the distribution of the average subject dereferenciablity, average predicate dereferencia-

[12] http://delicias.dia.fi.upm.es/LDQM/index.php/Accessibility

[13] http://linkeddata.es/resource/ldqm/sparql

[14] http://nandana.github.io/ld-sniffer/sparql.html

[15] https://datahub.io/dataset/ldqm-dbpedia-2016

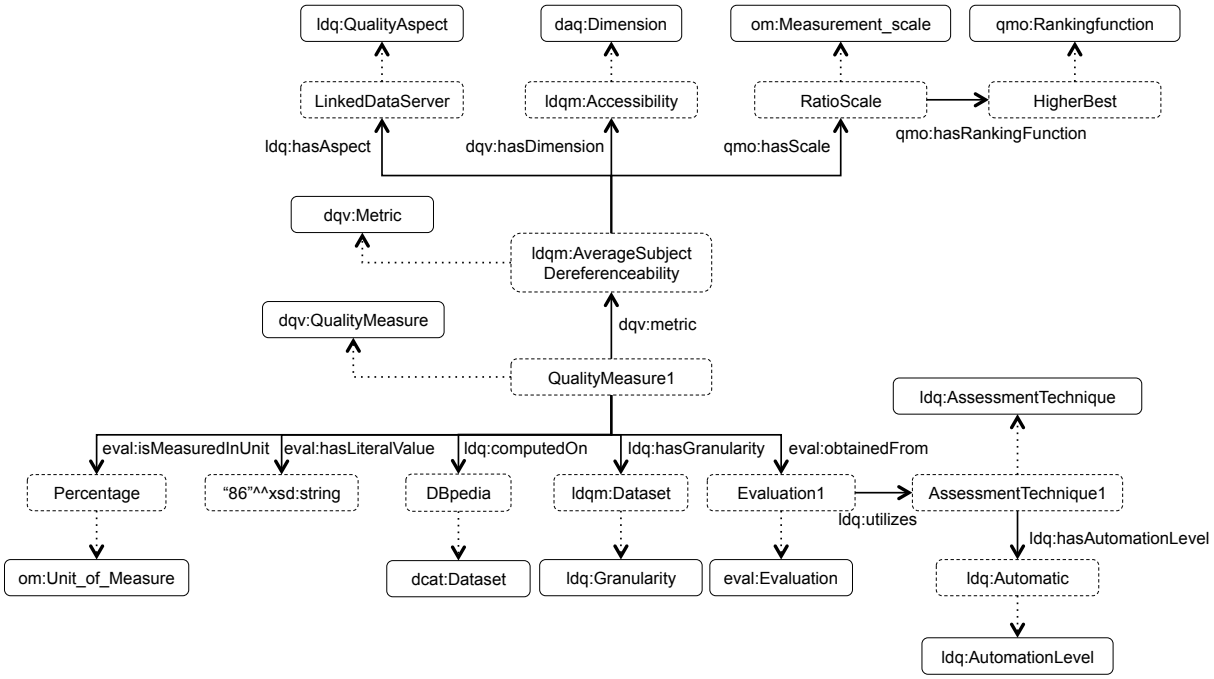


Fig. 7. Example of dataset evaluation results

blity, and average object dereferenciablity measurements of the instances in the DBpedia dataset of the dbo:Person and dbo:Place classes, of the predicates used to describe such instances, and of the objects in those predicates. The X-axis of the line graphs show the average dereferenciablity (rounded to the nearest integer) while the Y axis shows the number of resources with a given average dereferenciablity. In general, both average predicate dereferenciability and average object dereferenciability are commonly in the 90 to 100 range while the average subject dereferenciability varies in a wider range. The main reason for this is the fact that the DBpedia server fails when many requests are sent and returns a '503 Service Unavailable' error.

Figure 9 shows a pie chart with the most common errors (HTTP status codes) returned when dereferencing IRIs found in the content of dbo:Person and dbo:Place Linked Data resources. The most common status codes include failures in the server to provide a valid response such as '500 Internal Server Error', '502 Bad Gateway', or '503 Service Unavailable' and currently non-exiting IRIs with '404 Not Found', or '410 Gone'. More details of the analysis are found in the evaluation results website[16].

The main advantage of using, on the one hand, the Linked Data Quality Model for the evaluation and, on the other hand, its RDF representation for representing the information about this evaluation is that it provides reference for the measures to be evaluated and a large amount of provenance information for understanding how the evaluation was performed. For instance, when a metric such as "average subject dereferenceability"[17] is dereferenced it provides information such as the definition of the metric, its description, the base measures or derived measures used for calculating the given indicator, the scale, the relevant quality characteristic, and the Linked Data aspect. In addition, the base measures such as "number of distinct subjects"[18] provide the technique that was used to calculate the base measure. This additional information helps the quality evaluation result consumers to get a better understanding of the evaluation results and utilize them for making decisions.

[16] http://nandana.github.io/ld-sniffer/

[17] http://linkeddata.es/page/resource/ldqm/QualityIndicator/Averagesubjectdereferenceability

[18] http://linkeddata.es/page/resource/ldqm/BaseMeasure/Numberofdistinctsubjects'

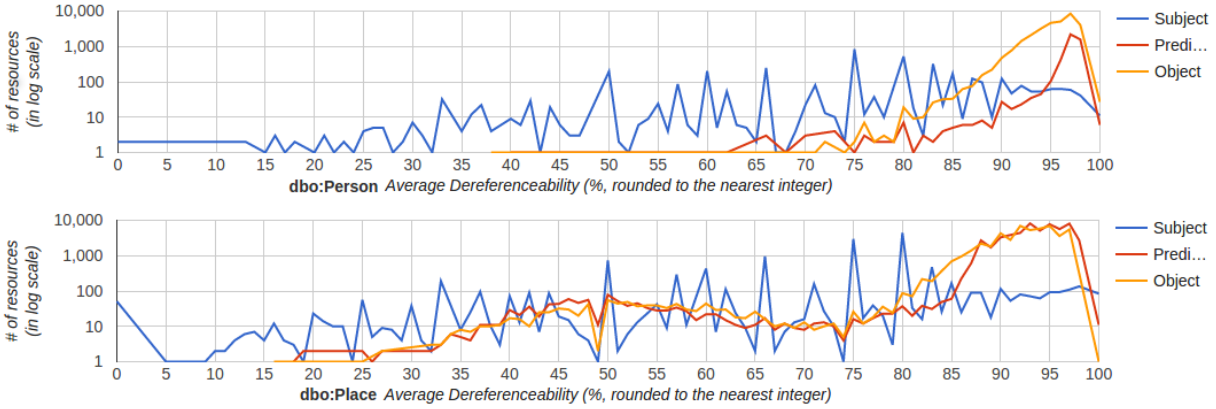


Fig. 8. Average dereferenciability of dbo:Person/dbo:Place instances

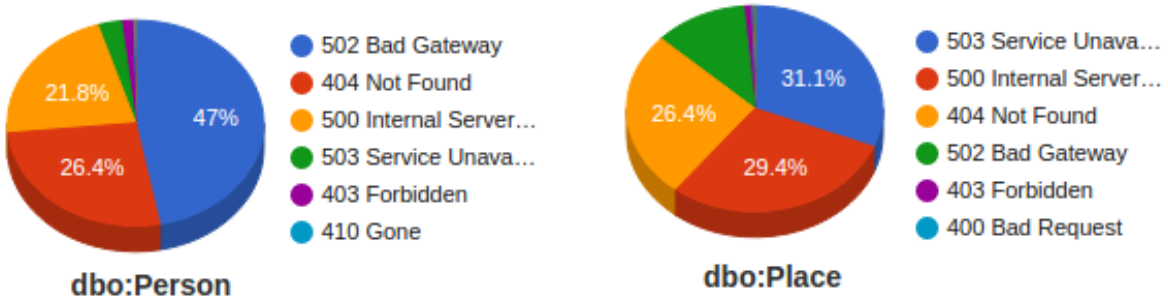


Fig. 9. Distribution of HTTP status codes for non-dereferenciable IRIs

## 6. Discussion

This paper presents a quality model for Linked Data. As there are no exact criteria to be referred to when evaluating quality models [2], this section provides a discussion related to various criteria described in the literature [2,11], as well as to the criteria that we consider as the most important for the Linked Data domain (i.e., consistency).

*Comprehensiveness* tends to describe the coverage of the quality model in terms of quality characteristics and quality measures [2], i.e., whether a quality model is complete with respect to the state of the art efforts in quality assessment.

The quality model presented in this paper is based on the survey by Zaveri et al. [16] which is, arguably, quite comprehensive with respect to the quality characteristics and quality measures that are evaluated in the state of the art. Since the coverage of the quality model presented in this paper is in direct relation with the coverage of the survey by Zaveri et al., it can be assumed that the Linked Data quality model is as comprehensive and complete as the mentioned survey. Furthermore, the hierarchy of quality measures in the Linked Data quality model includes measures that are related to various aspects of Linked Data quality (Figure 1), covering all the defined aspects. To this extent, the Linked Data quality model proposed in this paper is also comprehensive in terms of the described Linked Data quality aspects. Furthermore, the Linked Data quality model covers the quality measures in more details than Zaveri et al., and it also describes a higher number of measures.

*Applicability* tends to describe the cases in which a quality model has been successfully applied in practice.

Since the main purpose of a quality model is to provide guidelines in the evaluation process, as a use case we have developed a tool for Linked Data evaluation (Section 5). Although the mentioned tool does not

cover all the quality measures described in the Linked Data quality model, it has been developed following the quality model guidelines. Furthermore, the evaluation results produced by the tool are in line with the quality model.

Despite the fact that the tool presented in this paper is an initial effort in exploiting the Linked Data quality model, it can be assumed that this quality model has potential to be successfully applied in practice as a reference for Linked Data evaluation since parts of it are already used by existing tools and ontologies to represent their evaluation results.

*Understandability* tends to describe whether a quality model can be easily understood in order to be applied in practice. To this extent, Behkamal et al. argue that an important factor for clear and unambiguous quality models is their hierarchical organization of elements [2], while Bertoa et al. argue that understandability of a quality model is influenced by its structure and organization [13].

The Linked Data quality model presented in this paper is a hierarchical quality model, similarly as those defined by the ISO (e.g., ISO 25010 and ISO 25012). Furthermore, in order to address the specific nature of the domain and to reduce the ambiguity that is identified as a problem in generic quality models [38], we have introduced a hierarchy of quality measures and quality characteristics specific to Linked Data, and we have also provided definitions for all the elements as well as, where applicable, formulas for the calculation of quality measures. To this extent, it can be assumed that the Linked Data quality model presented in this paper is characterized by a high level of understandability.

*Consistency* tends to describe to which extend the elements of a quality model (e.g., quality characteristics and quality measures) are in agreement and compatibility with each other. Radulovic et al. argue that, for a quality model to be consistent, all the formulas in a quality model have to contain only those quality measures that are already defined in such model [11]. Furthermore, it is recommendable that all quality measures on lower levels in a hierarchy (i.e., base measures or derived measures) are used for obtaining quality measures on higher levels (i.e., derived measures or indicators), although in some cases measures on lower levels can be stand-alone and bring important information related to particular evaluation.

The quality model for Linked Data presented in this paper defines the quality measures and quality characteristics that, according to the guide by Radulovic et al., are completely compatible. All the base measures are used for defining derived measures or indicators, all the derived measures are used for defining indicators, and all the formulas contain only those measures that are defined in the quality model. Furthermore, all the indicators defined in our quality model are used for the measurement of quality characteristics, and all the specified measures are used in formulas. In this sense, it can be concluded that the Linked Data quality model presented in this paper is consistent.

## 7. Conclusions and future work

This paper describes a quality model for Linked Data, which extends the ISO 25012 data quality model. Such quality model is a step towards a consistent terminology for Linked Data quality, and it describes a comprehensive set of quality characteristics and measures specific to Linked Data, together with their definitions and formulas. Furthermore, it can serve to Linked Data publishers and producers as a quality requirements checklist.

The Linked Data quality model has been based on the current state of the art in Linked Data quality specification and evaluation, especially on the work by Zaveri et al. [16]. Regardless, it contributes to the state of the art by fomalizing a classification of different types of quality measures (i.e., base measures, derived measures and indicators) which, in the Linked Data field, have not been previously defined, as well as by describing quality measures that have not been described previously.

The quality model proposed in this paper adopts the quality characteristics from the ISO 25012 standard. It includes all the ISO 25012 quality characteristics, although some of these characteristics do not have any measures identified yet in the Linked Data quality evaluation state of the art. However, these gaps can guide the future research on Linked Data evaluation.

This paper also presents a conceptual model and an implementation of the conceptual model into an ontology for capturing the information specifically related to the Linked Data field; this ontology is an extension to existing ontologies for quality specification and assessment. Furthermore, by using the presented ontology, a description in RDF of the Linked Data quality model is also provided. This can help developers of various Linked Data evaluation tools to easily describe their results and reuse the Linked Data quality model,

which can further lead to better reusability and benchmarking of evaluation results.

The Linked Data quality model is described in the wiki, which is a living document that tends to capture new measures that might appear in the literature. Similarly, the RDF description of the quality model is expected to evolve, following the advances in Linked Data quality specification and evaluation.

The quality model presented in this paper has been used in the development of LD Sniffer, an online tool for assessing the quality of Linked Data resources. As discussed in Section 5, the experience has shown that a quality model could help in the design and implementation of an evaluation tool, as well as that the quality model could improve the understanding of measures or even of how a concrete evaluation tool works.

Section 6 discusses the quality model proposed in this paper in terms of various evaluation criteria. However, the quality model has not been thoroughly evaluated, which is an important line of future work. A special focus of the evaluation can be the applicability of the quality model and its usefulness as judged by potential users.

## References


[1] Mihindukulasooriya, N., García-Castro, R., Esteban-Gutiérrez, M.: Linked Data Platform as a novel approach for Enterprise Application Integration. In: Proceedings of the 4th International Workshop on Consuming Linked Data (COLD2013), Sydney, Australia (Oct 2013)

[2] Behkamal, B., Kahani, M., Akbari, M.: Customizing ISO 9126 quality model for evaluation of B2B applications. Information and software technology **51**(3) (2009) 599–609

[3] OMB: Guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of information disseminated by federal agencies. Part IX. Office of Management and Budget. Technical report, Office of Management and Budged (USA) (2002)

[4] Chisholm, M.: The Implementation of Basel Committee BCBS 239: An Industry-Wide Challenge for International Data Management. In: Proceedings of the 2nd International Data and Information Management Conference (IDIMC), Loughborough, UK. (2014)

[5] ISO: ISO/IEC 25012:2008, Software engineering – Systems product Quality Requirements and Evaluation (SQuaRE) – Data quality model. Technical report, International Organization for Standardization (2008)

[6] Juran, J.M., Godfrey, A.B.: Juran's Quality Handbook. Fifth Edition. McGraw-Hill (2000)

[7] Kitchenham, B.: DESMET: A method for evaluating software engineering methods and tools. Technical report, Department of Computer Science, University of Keele, Staffordshire, UK (1996)

[8] García-Castro, R.: Benchmarking Semantic Web technology. Studies on the Semantic Web vol. 3. AKA Verlag – IOS Press. (2010)

[9] Azuma, M.: SQuaRE: The next generation of the ISO/IEC 9126 and 14598 international standards series on software product quality. In: Proceedings of the European Software Control and Metrics Conference (ESCOM). London, UK. (2001) 337–346

[10] ISO: ISO/IEC 25010:2011, Systems and software engineering – Systems and software Quality Requirements and Evaluation (SQuaRE) – System and software quality models. Technical report, International Organization for Standardization (2011)

[11] Radulovic, F., García-Castro, R., Gómez-Pérez, A.: SemQuaRE – An extension of the SQuaRE quality model for the evaluation of semantic technologies. Computer Standards & Interfaces **38** (2015) 101–112

[12] ISO: ISO/IEC 15939:2007, Systems and software engineering – Measurement process. Technical report, International Organization for Standardization (2007)

[13] Bertoa, M.F., Troya, J.M., Vallecillo, A.: Measuring the usability of software components. Journal of Systems and Software **79**(3) (2006) 427–439

[14] Franch, X., Carvallo, J.: Using quality models in software package selection. Software, IEEE **20**(1) (2003) 34–41

[15] Dromey, R.: Software product quality: theory, model, and practice. Software Quality Institute, Brisbane, Australia (1998)

[16] Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: A survey. Semantic Web – Interoperability, Usability, Applicability (2014)

[17] Bizer, C., Cyganiak, R.: Quality-driven information filtering using the WIQA policy framework. Web Semantics: Science, Services and Agents on the World Wide Web **7**(1) (2009) 1–10

[18] Behkamal, B., Kahani, M., Bagheri, E.: Quality Metrics for Linked Open Data. In: Database and Expert Systems Applications, Springer (2015) 144–152

[19] Albertoni, R., De Martino, M., Podesta, P.: A linkset quality metric measuring multilingual gain in SKOS thesauri. In: Proceedings of the 2nd Workshop on Linked Data Quality (LDQ2015), Portorož, Slovenia. (2015)

[20] Gil, Y., Ratnakar, V.: Trusting information sources one citizen at a time. In: Proceedings of the 1st International semantic Web Conference, Sardinia, Italy. Springer (2002) 162–176

[21] Golbeck, J., Parsia, B., Hendler, J.: Trust Networks on the Semantic Web. In: Cooperative Information Agents VII. Volume 2782. Springer Berlin Heidelberg (2003) 238–249

[22] Hartig, O.: Trustworthiness of data on the web. In: Proceedings of the STI Berlin & CSW PhD Workshop, Citeseer (2008)

[23] Böhm, C., Naumann, F., Abedjan, Z., Fenz, D., Grütze, T., Hefenbrock, D., Pohl, M., Sonnabend, D.: Profiling Linked Open Data with ProLOD. In: Proceedings of the 26th International Conference on Data Engineering Workshops (ICDEW), Long Beach, California, USA, IEEE (2010) 175–178

[24] Ermilov, I., Martin, M., Lehmann, J., Auer, S.: Linked Open Data statistics: Collection and exploitation. In: Knowledge Engineering and the Semantic Web. Springer (2013) 242–249

[25] Palmonari, M., Rula, A., Porrini, R., Maurino, A., Spahiu, B., Ferme, V.: ABSTAT: Linked Data Summaries with ABstraction and STATistics. In: The Semantic Web: ESWC 2015 Satellite Events. Springer (2015) 128–132

[26] Mihindukulasooriya, N., Poveda-Villalón, M., García-Castro, R., Gómez-Pérez, A.: Loupe – An Online Tool for Inspecting

Datasets in the Linked Data Cloud. In: Demo at The 14th International Semantic Web Conference, Bethlehem, USA. (2015)

[27] Guéret, C., Groth, P., Stadler, C., Lehmann, J.: Assessing Linked Data mappings using network measures. In: The Semantic Web: Research and Applications. Springer (2012) 87–102

[28] Mendes, P.N., Mühleisen, H., Bizer, C.: Sieve: linked data quality assessment and fusion. In: Proceedings of the 2012 Joint EDBT/ICDT Workshops, ACM (2012) 116–123

[29] Kontokostas, D., Westphal, P., Auer, S., Hellmann, S., Lehmann, J., Cornelissen, R., Zaveri, A.: Test-driven evaluation of Linked Data quality. In: Proceedings of the 23rd international conference on World Wide Web, ACM (2014) 747–758

[30] Feeney, K.C., O'Sullivan, D., Tai, W., Brennan, R.: Improving curated web-data quality with structured harvesting and assessment. International Journal on Semantic Web and Information Systems (IJSWIS) **10**(2) (2014) 35–62

[31] Acosta, M., Zaveri, A., Simperl, E., Kontokostas, D., Auer, S., Lehmann, J.: Crowdsourcing Linked Data quality assessment. In: Proceedings of the 12th International Semantic Web Conference, Sydney, NSW, Australia. Springer (2013) 260–276

[32] Ruckhaus, E., Baldizán, O., Vidal, M.E.: Analyzing Linked Data Quality with LiQuate. In: On the Move to Meaningful Internet Systems: OTM 2013 Workshops, Graz, Austria, Springer (2013) 629–638

[33] Beek, W., Rietveld, L., Bazoobandi, H.R., Wielemaker, J., Schlobach, S.: LOD laundromat: A uniform way of publishing other people's dirty data. In: Proceedings of the 13th International Semantic Web Conference, Riva del Garda, Italy. Springer (2014) 213–228

[34] Debattista, J., Auer, S., Lange, C.: Luzzu – A Framework for Linked Data Quality Assessment. In: Proceedings of IEEE Tenth International Conference on Semantic Computing (ICSC2016). (2016) 124–131

[35] 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, Uppsala, Sweden, ACM (2011) 1–8

[36] Debattista, J., Lange, C., Auer, S.: daQ, an Ontology for Dataset Quality Information. In: Proceedings of the Workshop on Linked Data on the Web, co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014. (2014)

[37] Beckett, D., McBride, B.: RDF/XML syntax specification (revised). W3C recommendation **10** (2004)

[38] Al-Kilidar, H., Cox, K., Kitchenham, B.: The use and usefulness of the ISO/IEC 9126 quality standard. In: Proceedings of the 2005 International Symposium on Empirical Software Engineering, Noosa Heads, Australia. (2005) 126–132