Assessing, Monitoring and Analyzing Linked Data Quality in Public SPARQL Endpoints*

Muhammad Intizar Ali[1], Qaiser Mehmood[1], and Muhamamad Saleem[2]

[1] Insight Centre for Data Analytics,
National University of Ireland, Galway, Ireland
{ali.intizar,qaiser.mehmood}@insight-centre.org
[2] University of Leipzig, Germany
saleem@informatik.uni-leipzig.de

Abstract. In this paper, we propose a domain agnostic and query driven approach to monitor, assess, and analyze quality of the linked data hosted by public SPARQL endpoints. We identified various quality related metrics for linked datasets and used linked data vocabulary to represent quality information. We provide a Linked Data Quality (LDQ) dataset, which is generated after conducting various quality related tests over a few public SPARQL endpoints. Our main goal in this paper is to provide a platform for monitoring, assessing and analyzing linked data quality. Data consumers can also execute various analytical queries over LDQ to analyze quality related metrics of the public SPARQL endpoints. We hope that LDQ will increase data consumer's confidence over public SPARQL endpoints and will support the wide adoption of these datasets in various linked data applications.

1 Introduction

Linking Open Data (LOD) is gaining popularity with every passing day and the amount of data available at LOD is growing rapidly. The LOD cloud contains data originated from hundreds of sources and the number of data sources is continuously increasing[3] [3]. These datasets are accessible through different interfaces such as SPARQL endpoints, triple patterns fragments, RDF datadumps, and HDT files. SPARQL endpoints provide a public interface for querying the underlying RDF data. Provision of access to linked datasets through SPARQL queries not only facilitates an easy access to the datasets, but it also allows data consumers to integrate data from multiple datasets on the fly. Moreover, applications can use these datasets without committing any resources to locally host these large linked datasets. According to SPARQLES (https://sparqles.ai.wu.ac.at/), which is a service to monitor status of public

* This research has been partially supported by Science Foundation Ireland (SFI) under grant No. SFI/12/RC/2289.

[3] LOD Stats: http://stats.lod2.eu/

SPARQL endpoints, there are around 557 sparql endpoints accessible on the Web, last accessed: July 2019).[4].

However, a wide adoption of public SPARQL endpoints is hindered by a number of challenges. Data quality, reliability and quality of service are among the prominent challenges faced by any linked data application using SPARQL endpoints. Limited availability of information related to data quality results into decreasing the confidence and trust of data consumers in public open linked data services. To this end, different monitoring services have been proposed to monitor and evaluate the quality of service features of public SPARQL endpoints. However, in order to evaluate data quality of any dataset usually a deep understanding of the internal structure of the data and domain specific knowledge is required.

In this paper, we propose a domain agnostic and query driven quality monitoring and assessment approach to remotely assess the quality of the linked datasets which are accessible via public SPARQL endpoints. We identified various quality related metrics for linked datasets which can be monitored through various SPARQL queries. Contrary to the existing query driven approaches, we designed a linked data quality (LDQ) dataset, which contains quality profiles of different public SPARQL endpoints generated at various timestamps. Each quality profile holds results of query-driven tests conducted over any given SPARQL endpoint. Initially, we focused on three important aspects of linked data, namely (i) IRI's, (ii) data types, and (iii) data structured-ness (introduced in [6]). Regarding IRI's, we designed tests to evaluate the validity of the IRI's in the linked dataset. We also evaluated dereference-ability of these IRI's. Regarding the data types, we provide a sample test to locate all DateTime literals which are wrongly stored as string data types, and lastly for data structured-ness we computed individual and weighted class coverage to show the coherence or structured-ness of any given dataset. Despite we conducted an evaluation for a limited number of parameters, the LDQ dataset is easily extensible and users can evaluate any quality metric of their choice by designing their own query driven tests and execute them over any SPARQL endpoint. Results of all quality assessment tests are stored as linked data following LDQ vocabulary[5] structure and these results are linked to a quality profile generated for that particular public SPARQL endpoint. Our aim is to provide a central monitoring service which executes quality assessment tests following a pre-defined schedule and it also allows its users to execute on-demand tests. A quality profile of each public SPARQL endpoint will be generated after every planned test and values for different quality metrics will be stored in the quality profile. We host LDQ as a SPARQL endpoint accessible at: `http://srvgal89.deri.ie:8022/sparql`. The open access to public SPARQL endpoints hosting LDQ data facilitates data consumers to directly execute various analytical queries for analyzing quality metrics of any SPARQL endpoint. Users can also analyze historical data to understand quality related evolution by

[4] SPARQLES service is executed periodically to check status of public SPARQL endpoints and the number of available SPARQL endpoints can fluctuate.

[5] Data Quality Vocabulary: `https://www.w3.org/TR/vocab-dqv/`

observing the change pattern of quality metrics over the time. LDQ has potential to increase data consumers confidence over public SPARQL endpoints and hence, can contribute towards the wide adoption of public SPARQL endpoints by linked data applications. We also provide a Web interface to execute test over a limited number of endpoints. We foresee LDQ provided as a service for quality monitoring and attaching the evaluated quality profiles to each dataset (initially only public SPARQL endpoints) listed in the Linked Open Data Cloud.

Structure of the Paper: We position our work in comparison with the state of the art in Section 2. In Section 3, we identify linked data quality metrics and present LDQ data model. Section 4 discusses our quality assessment approach with a list of quality related parameters and their evaluation methods. We discuss on linked data quality monitoring approach and few some evaluation results in Section 5. We conclude our work and discussed future directions in Section 6.

2 Related Work

Different approaches have been proposed for linked data quality assessment over the past [10, 16, 5], which are broadly categorized as (i) automated, (ii) semi-automated, and (iii) manual. Most of these approaches require the involvement of a user with expert domain knowledge of the given dataset under quality inspection. Due to the requirement of domain knowledge, quality assessment tests cannot be generalized for all type of datasets. Test-driven approaches have been proposed for quality assessment of linked datasets and different SPARQL queries are designed to assess the quality of linked data [9]. Similarly, crowdsourcing approaches for linked data quality assessment are also introduced [1]. However, most of these approaches have conducted a one-time quality evaluation. In the dynamic Web environment, linked datasets are also prone to frequent updates, which can potentially change the quality level of the overall datasets after every update. Moreover, linked datasets are gradually increasing and improving at the same time. Hence, one-time quality assessment of any public SPARQL endpoint will not truly reflect the quality assessment of frequently updating linked datasets.

SPARQLES is a monitoring service designed to monitor status of public SPARQL endpoints [4, 18]. This service is executed periodically using various SPARQL queries to monitor four performance metrics of endpoint service namely, (i) Availability, (ii) Performance, (iii) Interoperability, and (iv) Discoverability. Results of the SPARQLES monitoring are accessible at `https://sparqles.ai.wu.ac.at/`. Our proposed work is very closely aligned to SPARQLES except the fact that we are focusing on the quality of the underlying data hosted by the SPARQL endpoint rather than quality of service as monitored by SPARQLES.

Acknowledging the importance of quality measurements of linked open data, a community effort that has led to defining a W3C proposed standard for Data Quality Vocabulary (DQV), accessible at: `https://www.w3.org/TR/vocab-dqv/`. We built our dataset of monitoring linked data quality of public

Data Quality Dimensions	**Definition**
Accessibility	To which extent data is available and accessible.
Amount of Data	The amount of data available is enough to perform the required task.
Believability	Credibility and trustworthiness of a given data set.
Completeness	Data is not missing any values and provides enough information.
Concise Representation	Data is represented in a compact form without any redundancy.
Consistent Representation	Data is having same format.
Ease of Manipulations	To which extent it is easy to manipulate data and apply it to different tasks.
Free of Error	To which extent data is free from errors, correct and reliable.
Interpretability	Language, symbols and units are understandable and definitions are clear.
Objectivity	To the extent the data is unbiased and free from any prejudice or impartiality.
Relevancy	To which extent it is easy to manipulate data and apply it to different tasks,
Reputation	Depending on a task in hand how relevant is the given data.
Security	To the extent access to data is restricted only to the authorised users.
Timeliness	Data is up-to-date and have the latest information related to a given task.
Understandability	To the extent data is easily understood and comprehended.
Value-Addition	The value addition provided by data making its use beneficial.

Table 1. Data Quality Dimensions[11]

SPARQL endpoints using the same vocabulary. A similar approach to represent QoS parameters of public SPARQL endpoints using a QoS data models is presented in [2].

3 Linked Data Quality Metrics and Data Model

In this section, we discuss two important data quality related metrics specifically for linked data quality assessment and present DQV data model which was used for representing and storing values of quality metrics calculated over data hosted by public SPARQL endpoints.

3.1 Linked Data Quality Monitoring

Data quality is a broad term referring to a variety of dimensions and quality check metrics. Pipono et. al. summarised 16 dimensions of data quality. Table 1 provides an overview of data quality dimensions listed in [11]. As it is evident from the given list of dimensions that data quality assessment is heavily

dependent on the domain of data as well as requirements of data manipulation tasks. Zavari et. al. presented a comprehensive overview of linked data quality metrics and added a few additional quality metrics which they believed are more relevant to the linked datasets [19]. These metrics are namely, (i) Interlinking, (ii) Licensing, (iii) Versatility, and (iv) Security.

However, due to the distributed nature of the linked data and mostly availability of open access to this data via SPARQL endpoints, it is not easy to apply quality tests locally. Most of the existing quality testing of linked data require a local replica of complete dataset before evaluating quality metrics. Due to the resource constraints it is not easy to download a complete dataset hosted at a SPARQL endpoint either due to limits on data access imposed by the SPARQL endpoint service or simply due to the large size of the hosted data which makes it hard to download and process a local replica.

3.2 Query-driven Linked Open Data Quality Assessment

SPARQL endpoints follow a distributed service oriented architecture, where different endpoints are accessible using SPARQL query service making it very hard to create a local copy of a dataset containing all data sources due to large size and high level of distribution. Hence, contrary to the existing quality checks over linked data which require a complete local access to the whole dataset, we focused on generic mechanisms to assess data quality of linked data hosted by SPARQL endpoints. We define generic quality assessment SPARQL queries which can be executed by any client capable of dispatching queries to SPARQL endpoints using SPARQL query service. We propose a query based evaluation of quality metrics, which can be executed over any endpoint using SPARQL queries. We identify various data quality parameters for linked datasets and consider only the relevant quality parameters, which can be evaluated by executing SPARQL queries.

A few examples of potential query driven quality metrics assessment are listed below;

- Validity of IRIs can be determined by extracting all IRIs in a dataset hosted at a SPARQL endpoint and then check which percentage of the total IRIs are valid IRIs.
- Fact checking by comparing the answers of same query over multiple endpoints hosting similar information.
- Contradictory information detection by using well-know predicates (e.g. date of birth and date of death) and checking whether the corresponding triples are using valid date-time format and free from contradictions (e.g. date of birth, date of death and age triples are presenting accurate information).
- De-referenceability of IRIs in a dataset can check via SPARQL queries indicating to which extent all the IRIs presented in a dataset are dereferenceable.

It is worth mentioning, that the general categorization of quality parameters provided in this article is not exhaustive but rather an indicative list to showcase

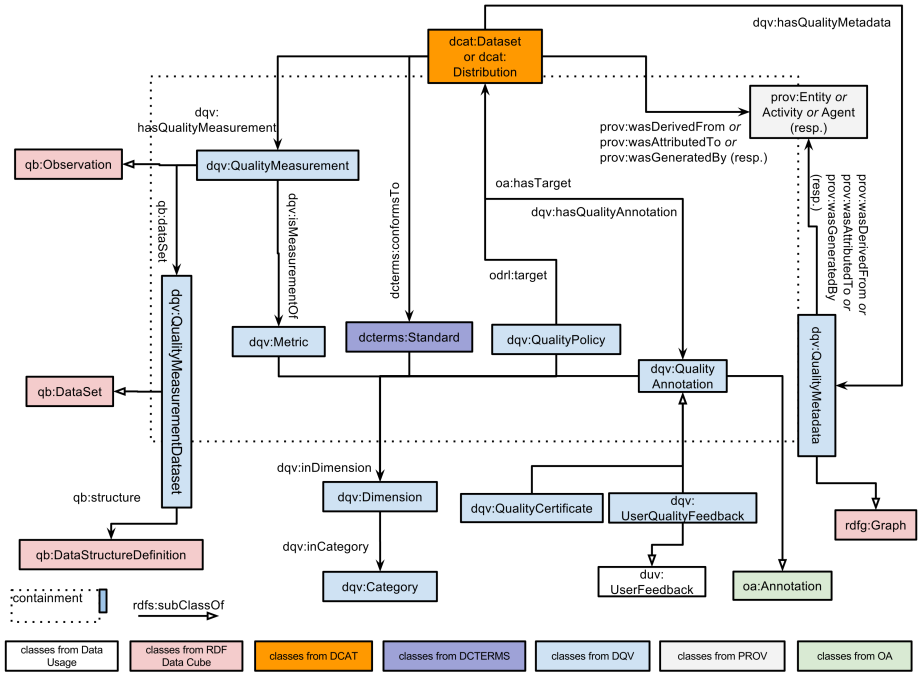

Fig. 1. Linked Data Quality: Data Model (https://www.w3.org/TR/vocab-dqv/)

only relevant quality parameters and their broader categories. The exact categorization of each query-driven test or quality parameter is beyond the scope of this paper. We left this task at the user's discretion to allocate broader category for any of the quality parameters discussed in this paper or even for their own defined quality parameter.

3.3 LDQ Data Model

We used the W3C Data Quality Vocabulary to represent the outcomes of quality evaluation results. Figure 1 gives an abstract overview of the Data Quality vocabulary showing a few relevant classes. LDQ data model is flexible and any number of data quality parameters can be introduced after their proper categorization. Prefix ldq:http://www.insight-centre.org/ldq is the default prefix for all classes and properties starting with ":" symbol in Figure 1. For the most of the dataset, we stick to the classes and prefixes defined within the DQV. The detailed description of the vocabulary can be accessed at the W3C description of DQV accessible at: https://www.w3.org/TR/vocab-dqv/

4 Assessing Linked Data Quality

In order to assess query driven quality of any public SPARQL endpoint, we identified various quality related parameters. This section discuss quality related parameters that are considered in this paper along their assessment methods. Quality parameters, measured in this paper, are mainly categorized in three types, namely, (i) IRI's , (ii) Data Types, (iii) Data Structure. Below we discuss each of these category and their relevant tests.

4.1 IRIs Related Quality Parameters

IRI are one of the key ingredient of linked data and hold a prominent role in the vision and principles of linked data. IRIs related quality parameters indicate to which level any dataset adhere to linked data principles. We consider the following IRI related quality parameters.

IRI Validity: IRI validity refers whether a given IRI is complying to the IRI syntax or not. For example any IRI containing restricted characters (e.g. a space) is not a valid IRI. IRI validity test can be conducted by simply selecting all IRIs and then using pre-defined java *UrlValidator* function to check whether a selected IRI is valid.

IRI Dereference-ability: Dereferencing refers the process of retrieving resource representation. It is an important feature of linked data principles which demands that all IRIs within a link dataset must dereference. It is particularly important for link traversal-based federated SPARQL query processing[7]. In this type of SPARQL federation, the query processing is done through traversing dereference-able IRI's [13]. Quality parameter for linked data can evaluate that how many of the total IRIs are dereference-able. This can be achieved by retrieving the list of all IRIs in the dataset, similar to the IRI validity test, and then follow the *http* path for each IRI to validate whether that particular IRI is dereference-able.

Blank Nodes: Blank nodes are an important feature of linked data, while the number of blank nodes is not necessarily a quality parameter, but a statistical information to showcase the percentage of blank nodes in the linked dataset can definitely indicate the quality of a linked dataset. SPARQL query processing in presence of blank nodes is particularly challenging [8, 17].

4.2 Data Type Related Quality Parameters

These parameters are mainly concerned with the literal values in a linked dataset. Ideally, most of the literals have specific data types announced to indicate which type of data can be stored in that literal. This quality parameter can indicate how correctly data types are defined and whether all literals hold a data value belonging to the right data type.

Date Type Validity: String is a default data type for all literals in linked datasets, unless described otherwise. This leads to possibilities of having values belonging to other data types being stored in string format. A common mistake is to have literal values stored as string instead of the best matching data type for that particular value. A simple date type quality parameter can calculate the total number of all those *xsd:dateTime* values which are wrongly stored as *xsd:String* data type.

4.3 Data Structuredness Related Quality Parameters

These types of quality parameters provide insights related to internal structure of the dataset. Since linked dataset are essentially a graph structure, so these parameters showcase how connected or disconnected is any linked dataset. We discuss few of the structuredness related quality parameters below;

Class Coverage: This metric was introduced in [6] and determines how well the instance data conform to rdf:class (class for short), i.e., how well a specific class is covered by the different instances of that class. The coverage of a class C demented by $Coverage(C)$ is defined as follow:

Definition 1 (Class Coverage). *For a dataset D, let $P(C)$ denote the set of distinct properties having class C and $I(C)$ denote the set of distinct instances having class C. Let $I(p, C)$ denote the number of distinct instances having predicate p and class C. Then, the coverage of the class $CV(C)$ is*

$$CV(C) = \frac{\sum_{\forall p \in P(C)} I(p,C)}{|P(C)| \times |I(C)|}$$

```
SELECT Count(Distinct ?s) as ?occurences
WHERE {
        ?s a <Class name C> .
  ?s <Predicate p> ?o
   }
```

Listing 1. Calculating the number of distinct instances having predicate p and class C denoted by I(p, C)

```
SELECT DISTINCT ?typePred
WHERE {
?s a <Class name C> .
 ?s ?typePred ?o
 }
```

Listing 2. The set of distinct properties having class C denoted by P(C)

```
SELECT Count(DISTINCT ?s)  as ?cnt
WHERE {
?s a <Class name C> .
```

```
?s ?p ?o
}
```

Listing 3. Calculating the number of instances having class C denoted by I(C)

Listings 1, 2, and 3 contain three different SPARQL queries which can be used to evaluate class coverage.

Weighted Class Coverage Definition 1 considers the structuredness of a dataset with respect to a single class. Obviously, a dataset D has instances from multiple classes, with each instance belonging to at least one of these classes (if multiple instantiations are supported). It is possible that dataset D might have a high structuredness for a class C, say CV(C) = 0.8, and a low structuredness for another class C', say CV(C') = 0.15. But then, what is the structuredness of the whole dataset with respect to our class system (set of all classes)? Duan et al. [6] proposed a mechanism to compute this, by considering the weighted sum of the coverage $CV(C)$ of individual classes. In particular, for each class C, the weighted coverage is defined below.

Definition 2 (Weighted Class Coverage). *Taking Definition 1 in to account, the weighted coverage for a class C denoted by $WT(CV(C))$ is calculated using the following formula:*

$$WT(CV(C)) = \frac{|P(C)|+|I(C)|}{\sum_{\forall C' \in D} |P(C')|+|I(C')|}$$

Dataset Structuredness By using Definitions 1, 2, we are now ready to compute the structuredness, hereafter termed as coherence, of a whole dataset D.

Definition 3 (Dataset Structuredness). *The overall structuredness or coherence of a dataset D denoted by $CH(D)$ is define as*

$$CH(D)) = \sum_{\forall C \in D} CV(C) \times WT(CV(C))$$

The dataset structuredness has a direct impact on the query runtimes as well as the result sizes. According to [14], the higher the dataset structuredness, the higher both result sizes and query runtimes of SPARQL queries. This metric is particularly important while designing federated SPARQL query benchmarks [12, 15]. A federated SPARQL querying benchmark should comprise of datasets from multiple domains with varying structuredness values [12].

5 Monitoring & Analyzing Linked Data Quality

In order to monitor the quality of linked data parameters, we defined a variety of query driven and domain agnostic tests which can be executed over linked datasets. We randomly selected 4 public SPARQL endpoints hosting linked datasets from different domains, details of the endpoints and their brief description is presented in Table 2.We conducted different tests on each of these 4 public SPARQL endpoints to monitor their data quality. A simple java program is written to execute SPARQL queries on a remote server. A list of selected

Name	Endpoint URI	Description
DBPedia	http://dbpedia.org/sparql	DBpdeia contains linked data representation of the data extracted from Wikipedia.
Semantic Web Dog Food	http://data.semanticweb.org/sparql	Semantic Web Dog Food contains linked dataset representing publications and attendees record of different conferences and workshops.
Symbolic Dataset	http://symbolicdata.org:8890/sparql	Symbolic data is a dataset designed for profiling, testing and benchmarking Computer Algebra Software (CAS).
LRI Dataset	https://sparql.lri.fr/sparql	LRI is a dataset containing information about the scientists working in a french laboratory.
Open Data	https://data.gov.cz/sparql	This endpoint contains national open data provided by govt. of Czech.
Linked ISPRA	http://dati.isprambiente.it/sparql	This dataset is a compartment of environmental protection information.

Table 2. Public SPARQL Endpoints

SPARQL endpoints was initially provided to the java program together with the list of all possible tests to be executed.

Our main aim for this evaluation was to showcase the feasibility and potential usage of LDQ by evaluating few quality parameters mainly belong to two broad categories of data quality assessment, namely, (i) *Completeness*, and (ii) *Accuracy*. We recommend LDQ users to consider LDQ categories in [19], to design tests for the quality evaluation of their own defined quality parameters. Depending on the nature of the test conducted, either a SPARQL query was able to directly provide the score of quality parameter or in some case additional processing was required after retrieving the SPARQL query results, for example in order to evaluate dereferencing of IRIs, all IRIs were retrieved by a SPARQL query and then each IRIs are tested by java program to locate any description of the resource from the Web. Results of quality tests were annotated following the data model described earlier and directly stored in a locally hosted SPARQL endpoint. We strongly encourage LDQ users to utilize existing LDQ dataset accessible at `http://srvgal89.deri.ie:8022/sparql`.

Listing 4 contains a sample query to access quality profile of Semantic Web Dog Food endpoint, while Listing 5 depicts a sample excerpt of the LDQ dataset.

Table 3 presents values of the different quality parameters assessed after executing these quality assessment tests[6]. We also expect to attract a larger audience who is willing to define their own quality parameters and their data quality assessment tests, in order to facilitate and encourage quality assessment tests design process, we provide source code of LDQ generation at: `https://github.com/qaimeh/LinkedDataQuality`

[6] Details of the tests and source code for test re-execution or reproduce-ability is available at `https://github.com/qaimeh/LinkedDataQuality`

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX dqv: <http://www.w3.org/ns/dqv#>

SELECT DISTINCT ?endpoint ?MeasurementName ?value
FROM <http://linked.data.quality/July-2019>
WHERE {
?endpoint a dcat:Dataset.
?endpoint dcterms:title ?title.
?endpoint dcat:distribution ?endpointDistribution .

?endpointDistribution dqv:hasQualityMeasurement ?measurements.
?measurements dqv:isMeasurementOf ?MeasurementName.
?measurements dqv:value ?value

FILTER (?title ="Semantic Web Dog Food" )

}
```

Listing 4. A Sample Query over LDQ Endpoint

6 Concluding Remarks and Future Directions

In this paper, we present LDQ, a linked data quality monitoring service to assess and analyze data quality of linked datasets. We designed a generic data model to present quality evaluation results for public SPARQL endpoints and showcase the feasibility of our approach by designing two simple quality tests over 5 public SPARQL endpoints. LDQ data model is extensible and users have freedom to define their own quality parameters and design the relevant query driven tests for the assessment of quality parameters. LDQ will serve as a baseline to get a general idea of data quality level of any public SPARQL endpoints, and data consumers can rely on statistics extracted from LDQ before using any public SPARQL endpoint. LDQ monitoring service will act as a central hub for data quality assessment and end-consumers can execute their quality assessment tests. As future directions, we plan to define a comprehensive list of query driven quality assessments tests and execute these tests on the complete list of public SPARQL endpoints available at Datahub. We plan to execute quality assessment tests periodically, which will result into a comprehensive linked data quality dataset and can be used to analyze linked datasets evolution in terms of their quality over the period of time. We also plan to host a linked data quality service for users who are not familiar with SPARQL, users can simply use online service to execute quality tests from a website. We foresee our service being run periodically on all datasets available as SPARQL endpoint and a quality score could be attached to each individual dataset within the whole LOD Cloud.

```
@prefix ldq:<http://insight-centre.org/LDQ#>.
@prefix xsd:<http://www.w3.org/2001/XMLSchema#>.
@prefix void:<http://www.w3.org/TR/void>.
@prefix muo:<http://purl.oclc.org/NET/muo/muo#/>.

:SWDF
    a dcat:Dataset ; dcterms:title "Semantic Web Dog Food" ;
    dcat:distribution :SWDFDistribution ;
    hasQualityMetaData  dqv:QualityMetadataSWDF .

:SWDFDistribution
    a dcat:Distribution ;
    dcat:downloadURL <http://www.scholarlydata.org/dumps/indicators
        /03-02-2018-indicators.nt> ;
    dcterms:title "RDF distribution of dataset" ;
    dcat:mediaType "text/nt" ;  dcat:byteSize "5889"^^xsd:decimal .

:SWDFDistribution
    dqv:hasQualityMeasurement :measurement1 .

dqv:QualityMetadataSWDF
    a dqv:QualityMetadata ;
    prov:generatedAtTime "2015-05-27T02:52:02Z"^^xsd:dateTime ;
    prov:wasGeneratedBy :SWDFQualityChecking .

:SWDFQualityChecking
    a prov:Activity; rdfs:label "The checking of SWDFDatasetDistribution's
        quality"^^xsd:string;
    prov:endedAtTime        "2015-05-27T02:52:02Z"^^xsd:dateTime;
    prov:startedAtTime      "2015-05-27T00:52:02Z"^^xsd:dateTime .

:measurement1
    a dqv:QualityMeasurement ;
    dqv:computedOn :SWDFDistribution ;
    dqv:isMeasurementOf :ntCompletenessMetric ;
    dqv:value "0.5"^^xsd:double ;
    prov:generatedAtTime "2015-05-27T02:52:02Z"^^xsd:dateTime ;
    prov:wasGeneratedBy :SWDFQualityChecking .

:ntCompletenessMetric
    a dqv:Metric ;
    skos:definition "Ratio between the number of objects represented  and
        the number of  objects expected to be represented according to the
         declared dataset scope."@en ;
    dqv:expectedDataType xsd:double ;
    dqv:inDimension :completeness .

#definition of dimensions and metrics
:completeness     a dqv:Dimension ;
    skos:prefLabel "Completeness"@en ;
    skos:definition "Completeness refers to the degree to which all
        required information is present in a particular dataset."@en ;
    dqv:inCategory :intrinsicDimensions .
```

Listing 5. A Sample Excerpt From LDQ Dataset

Name	IR	VI	PV	DI	PD	BN	BS	BO	DT	ST
DBPedia	1950000	1889033	96	1318941	67	55209471	27655447	27554024	0	0.19
SWDF	41700	41416	99	34797	83	37524	28164	9360	428	0.42
SD	41273	40702	98	16286	39	9	6	3	42	0.68
LRI	2047	1438	70	1048	51	421	348	73	1	0.52
Open Data	2048843	871730	42	1859127	90	46749	35369	11380	273	-
ISPRA	598111	597594	99	546609	91	1144	771	373	10907	0.95

Table 3. Quality Parameters Assessment Values (IR=Total IRIs, VI=Valid IRIs, PV= % Valid IRIs, DI=Dereference-able IRIs, PD = % Dereference-able IRIs, BN = Total Blank Nodes, BS=Blank Nodes as Subject, BO= Blank Nodes as Object, DT=Date Time as String, ST= Structuredness, SD = Symbolic Dataset). We were not able to get structuredness value for Open Data SPARQL endpoint due to runtime error.

Acknowledgments

This publication has emanated from research supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289-P2, co-funded by the European Regional Development Fund and Enable SPOKE under Grant Number 16/SP/3804. The work conducted in the University of Leipzig has been supported by the project LIMBO (Grant no. 19F2029I), OPAL (no. 19F2028A), KnowGraphs (no. 860801), and SOLIDE (no. 13N14456)

References

1. M. Acosta, A. Zaveri, E. Simperl, D. Kontokostas, S. Auer, and J. Lehmann. Crowdsourcing linked data quality assessment. In *The Semantic Web–ISWC 2013*, pages 260–276. Springer, 2013.
2. M. I. Ali and A. Mileo. How good is your sparql endpoint? In *On the Move to Meaningful Internet Systems: OTM 2014 Conferences*, pages 491–508. Springer, 2014.
3. C. Bizer, T. Heath, and T. Berners-Lee. Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, pages 205–227, 2009.
4. C. Buil-Aranda, A. Hogan, J. Umbrich, and P.-Y. Vandenbussche. Sparql web-querying infrastructure: Ready for action? In *International Semantic Web Conference*, pages 277–293. Springer, 2013.
5. J. Debattista, C. Lange, and S. Auer. Luzzu-a framework for linked data quality assessment. *arXiv preprint arXiv:1412.3750*, 2014.
6. S. Duan, A. Kementsietsidis, K. Srinivas, and O. Udrea. Apples and oranges: a comparison of rdf benchmarks and real rdf datasets. In *Proceedings of the 2011 ACM SIGMOD International Conference on Management of data*, pages 145–156. ACM, 2011.
7. O. Hartig, C. Bizer, and J.-C. Freytag. Executing sparql queries over the web of linked data. In *International Semantic Web Conference*, pages 293–309. Springer, 2009.

8. D. Hernández, C. Gutierrez, and A. Hogan. Certain answers for sparql with blank nodes. In *International Semantic Web Conference*, pages 337–353. Springer, 2018.
9. D. Kontokostas, P. Westphal, S. Auer, S. Hellmann, J. Lehmann, R. Cornelissen, and A. Zaveri. Test-driven evaluation of linked data quality. In *Proceedings of the 23rd international conference on World Wide Web*, pages 747–758. ACM, 2014.
10. P. N. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. In *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123. ACM, 2012.
11. L. L. Pipino, Y. W. Lee, and R. Y. Wang. Data quality assessment. *Communications of the ACM*, 45(4):211–218, 2002.
12. M. Saleem, A. Hasnain, and A.-C. N. Ngomo. Largerdfbench: a billion triples benchmark for sparql endpoint federation. *Journal of Web Semantics*, 48:85–125, 2018.
13. M. Saleem, Y. Khan, A. Hasnain, I. Ermilov, and A.-C. Ngonga Ngomo. A fine-grained evaluation of sparql endpoint federation systems. *Semantic Web*, 7(5):493–518, 2016.
14. M. Saleem, G. Szárnyas, F. Conrads, S. A. C. Bukhari, Q. Mehmood, and A.-C. Ngonga Ngomo. How representative is a sparql benchmark? an analysis of rdf triplestore benchmarks? In *The World Wide Web Conference*, pages 1623–1633. ACM, 2019.
15. M. Schmidt, O. Görlitz, P. Haase, G. Ladwig, A. Schwarte, and T. Tran. Fedbench: A benchmark suite for federated semantic data query processing. In *International Semantic Web Conference*, pages 585–600. Springer, 2011.
16. A. Schultz, A. Matteini, R. Isele, P. N. Mendes, C. Bizer, and C. Becker. Ldif-a framework for large-scale linked data integration. In *21st International World Wide Web Conference (WWW 2012), Developers Track, Lyon, France*, 2012.
17. A. Stolpe and J. Halvorsen. Distributed query processing in the presence of blank nodes. *Semantic Web*, 8(6):1001–1021, 2017.
18. P.-Y. Vandenbussche, J. Umbrich, L. Matteis, A. Hogan, and C. Buil-Aranda. Sparqles: Monitoring public sparql endpoints. *Semantic web*, 8(6):1049–1065, 2017.
19. A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *Semantic Web*, 7(1):63–93, 2015.