# Quality Assessment of Library Linked Data: a Case Study

Yusniel Hidalgo-Delgado[1](✉), Yoan A. López[1], Juan Pedro Febles Rodríguez[1], and Amed Leiva Mederos[2]

[1] Departamento de Informática, Universidad de las Ciencias Informáticas,
Havana, Cuba
{yhdelgado,yalopez,febles}@uci.cu
[2] Centro de Investigaciones en Informática, Universidad Central "Marta Abreu" de
Las Villa, Santa Clara, Cuba
amed@uclv.edu.cu

**Abstract.** The linked data principles provide an efficient way to interlink resources across several datasets, improving interoperability and discoverability. Several digital libraries around the world are publishing their legacy data from catalogs and authority files following the linked data principles. However, selecting the most suitable datasets for creating links between them is becoming a complex task due to most of them not having the proper data quality. In this paper, we evaluate the quality of data of the LinkedDL dataset. The results are compared to four other datasets in the state-of-the-art. The evaluation showed promising results in the accuracy, consistency, and accessibility metrics.

**Keywords:** Linked data · Digital libraries · Interoperability · Data quality

## 1 Introduction

Libraries play an important role in the visibility and access to scientific production. Currently, libraries are classified into conventional, digital, and hybrid. In the particular case of digital libraries, they collect, store and distribute information on digital media [16]. On the other hand, digital library systems are information systems that support the management of organized collections of digital objects (digital resources) that are oriented to users by providing value-added services [1].

Interoperability and sustainability are keys to realizing the vision of digital libraries that are able to communicate with each other. Interoperability is the ability of a system or a product to work with other systems or products without special effort on the part of the customer. In recent years, several approaches have been proposed to address the problem of semantic interoperability in digital libraries. The studies focus on three fundamental approaches, metadata crosswalk [2,6,11], ontology alignment [12], and linked data [3,18]. Linked data refers

B. Villazón-Terrazas et al. (Eds.): KGSWC 2021, CCIS 1459, pp. 93–108, 2021.
https://doi.org/10.1007/978-3-030-91305-2_8

to a set of principles and best practices for publishing and linking structured data on the Web. Data comes from different sources that can be maintained by organizations with different geographic locations [3].

The publication of data following the linked data principles enhances the discovery and reuse of data in the Web space while solving the semantic interoperability problems between information systems through the use of ontologies as a way of representing knowledge. There are several approaches for publishing library data as linked data, most of them are based on transforming legacy metadata to RDF graphs. Recently, a novel semantic interoperability model called LinkedDL for building linked data-based digital libraries was proposed [10]. In this paper, we asset the quality of library linked data generated by LinkedDL. Preliminary results show an improvement in the quality of linked data generated by LinkedDL in comparison with similar models proposed in the literature.

The paper is organised as follows: in the section Related Work, a brief overview of the quality of linked data is presented. Section Quality Metrics presents the scores for each quality metric evaluated in the LinkedDL dataset. Section Results and Discussions presents a qualitative description of the main results and their practical implications. Finally, the section Conclusions and Future Work presents our final remarks and future research lines.

## 2 Related Work

In recent years, many data publishers have been translating legacy data to linked data without checking the quality of data sources. For this reason, data consumers need to check the quality of linked data to ensure that they are fit for use according to certain quality needs. In this sense, several authors have proposed categories, dimensions, metrics, and tools to measure the quality of the existing linked data on the web of data [4,7–9,14,17,20].

According to Wang et al. [19], a data quality criterion is a particular characteristic of data concerning its quality and can be either subjective or objective. To measure the degree to which a certain data quality criterion is fulfilled for a given linked dataset, each criterion is formalized and expressed in terms of a function with the value range of $[0, 1]$. This function is called the data quality metric. Finally, one or several data quality criteria belongs to a data quality dimension.

A recent survey on quality assessment for linked data found a comprehensive list of 18 quality dimensions and 69 metrics [20]. Additionally, the authors qualitatively analyzed the 30 core approaches and 12 tools using a set of attributes. Most of these quality dimensions were used to evaluate the quality of five large knowledge graphs [9].

In the particular case of quality assessment of linked data in digital libraries, Candela and collaborators adapted existing quality dimensions to the context of digital libraries [4,5]. They carried out an extensive quality assessment study over four digital libraries. This study was taken as a baseline for the quality assessment in this paper. In Table 2, we include results obtained by Candela et al. [4] and compared them with our evaluation results.

# 3 Quality Metrics

To assess the quality of Library Linked Data, we use the LinkedDL[1] dataset. This dataset exposes bibliographic metadata from several scientific journals from Cuba following the Linked Data principles. It contains metadata about Authors, Articles, and Journals. In the case of Journals, links were generated with two other datasets in the web of data: wikidata and ISSN. In the next sections, we detailed each metrics and its corresponding scores in the LinkedDL dataset.

## 3.1 Accuracy

According to Wang et al. [19], the accuracy dimension determines the extent to which data are correct, reliable and certified free of error. The accuracy dimension was evaluated by means of the following four metrics:

*Syntactic Validity of RDF Documents:* syntax errors in RDF can be identified using tools such as the W3C RDF Validator[2]. The metric was originally defined as:

$$m_{synRDF} = \begin{cases} 1 & \text{if all RDF documents are valid} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

At the time of writing this paper, all RDF graphs were validated using the W3C validator, determining that all graphs are found free of syntactic errors. This check will be carried out periodically, taking into account that the graphs are generated regularly incorporating new information.

*Syntactic Validity of Literals:* this metric consists of determining if the literals stored in the RDF graph correspond to the syntax defined for each literal. In digital libraries, it is common to find syntactic patterns in the metadata, such as: names of authors, dates of publication, and identifiers (DOI, ISSN, among others). The syntax of these literals can be checked using regular expressions. The RDF graph $\mathcal{G}$ consists of RDF triples $(s, p, o)$ and a set of literals $\mathcal{L}$.

$$m_{synLit} = \frac{|\{\mathcal{G} \wedge \mathcal{L} \wedge o \text{ is valid}\}|}{|\{\mathcal{G} \wedge \mathcal{L}\}|} \tag{2}$$

The syntactic validity of the literals was checked using regular expressions in the SPARQL query language. The names of the authors, dates, titles, and ISSNs of the journals, among other literals, were checked. In all cases, the results were correct. The following SPARQL query checks and returns the ISSNs of all the journals in the repository:

[1] https://yhdelgado.github.io/linkeddl/.

[2] https://www.w3.org/RDF/Validator/.

```
prefix bibo: <http://purl.org/ontology/bibo/>
SELECT ?issn WHERE{
?s bibo:issn ?issn.
FILTER(regex(?issn,"^\\S{4}\\-\\S{4}$","i"))
}
```

*Semantic Validity of Triples:* this metric consists of evaluating the extent to which an RDF graph $\mathcal{G}$ contains the same values as an RDF graph $\mathcal{S}$ that *a priori* is known to have all its valid triples.

$$m_{semTriple} = \frac{|\{\mathcal{G} \wedge \mathcal{S}\}|}{|\{\mathcal{G}\}|} \tag{3}$$

For the evaluation of the semantic validity of the triples, all the triples that semantically describe the journals included in the repository were selected. To establish the semantic comparison, the existing triples in the semantic description in Wikidata were used. In all cases, the results were correct. For example, the title, ISSN, and URL of the entity *Revista Cubana de Información en Ciencias de la Salud* match both in Wikidata[3] and in the repository generated by LinkedDL[4].

*Duplicate Entities:* this metric consists of calculating the Wikidata link rate with duplicate identifiers, since there can be multiple identifiers in an RDF graph for the same entity. Let $n_w^u$ the number of unique entities linked to Wikidata, and $n_w$ the number of entities linked to Wikidata, then:

$$m_{checkDup} = \frac{n_w^u}{n_w} \tag{4}$$

For the duplicate entities detection, the RDF graph containing links to Wikidata was selected. The existing links are of the type *owl:sameAs*. After executing the following SPARQL query, no duplicate entities were detected in the graph.

```
SELECT ?s (COUNT(?id) AS ?total)
WHERE{?s owl:sameAs ?id}
GROUP BY ?s
HAVING (COUNT(?id)>1)
```

### 3.2 Trustworthiness

Trustworthiness is defined as the degree to which the information is accepted to be correct, true, real and credible [20]. Trustworthiness is evaluated at the following three levels:

*Trustworthiness on the Data Set Level:* the metric is originally defined as shown in Table 1.

[3] http://www.wikidata.org/entity/Q50816707.

[4] https://data.infocientia.com/resource/journal/2307-2113.

**Table 1.** Possible scores according to the metric trustworthiness on the data set level.

| Description | Score |
|---|---|
| Manual data curation, manual data insertion in a closed system | 1 |
| Manual data curation and insertion, both by a community | 0.75 |
| Automated data curation, data insertion by automated knowledge extraction from structured data sources | 0.25 |
| Automated data curation, data insertion by automated knowledge extraction from unstructured data sources | 0 |

The LinkedDL model establishes that the insertion and curation of bibliographic data are carried out by implementing specific wrappers for each data source. Some of these wrappers can insert metadata automatically, as is the case with the implemented wrapper for the OAI-PMH protocol. In other cases, the metadata is inserted manually by experts in library and information science. In both cases, the metadata goes through a manual cataloging and review process, with the aim of guaranteeing their veracity and consistency.

*Trustworthiness on the Statement Level:* this metric assesses whether there is information about the provenance at the instance level. Information about the provenance of the data can be semantically described using a vocabulary or ontology for this purpose, for example, PROV-O. The metric is defined as:

$$m_{fact} = \begin{cases} 1 & \text{provenance on statement level is used} \\ 0.5 & \text{provenance on resource level is used} \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The LinkedDL model does not include information about the provenance of the data at the instance level. Something similar happens with the rest of the digital libraries that are compared in the state of the art.

*Using Unknown and Empty Values:* trustworthiness can be increased by supporting unknown and empty values. These statements require unknown and empty values to be encoded with a different identifier. The metric was originally defined as:

$$m_{NoVal} = \begin{cases} 1 & \text{unknown and empty values are used} \\ 0.5 & \text{either unknown or empty values are used} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

In the LinkedDL model, an automatic conversion is performed from the relational database to the RDF data model. The conversion only takes into account the data that resides in the database, so no triples with unknown or empty values are generated. Something similar happens with the rest of the digital libraries that are compared in the state of the art.

### 3.3 Consistency

Consistency is defined as two or more values that do not conflict with each other [13]. Semantic consistency is the extent to which the repositories use the same values and elements for conveying the same concepts and meanings throughout [15]. Three aspects of consistency are measured as follows:

*Consistency of Schema Restrictions During Insertion of New Statements:* checking the schema restrictions during the insertion of new statements is often done on the user interface in order to avoid inconsistencies. For instance, that the entity to be added has a valid entity type, as expressed by the *rdf:type* property:

$$m_{checkRestr} = \begin{cases} 1 & \text{schema restrictions are checked} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

For the evaluation of this metric, the first 100 results were selected for each of the entities modeled in the graph. For all instances, it was checked using the *rdf:type* property. As a result, in all cases, the entities used this property properly. The following SPARQL query lists the first 100 triples of scientific articles type:

```
PREFIX fabio: <http://purl.org/spar/fabio/>
SELECT ?s ?p ?o WHERE {
?s ?p ?o.
?s rdf:type fabio:JournalArticle.
}
LIMIT 100
```

*Consistency of Statements with Respect to Class Constraints:* this metric measures the extent to which the instance data are consistent with regard to the class restrictions. Let $C$ be the set of all class constraints, defined as $C = \{(c_1, c_2)|(c_1, \text{owl:disjointWith}, c_2) \in \mathcal{G}\}$. Then, let $c_{\mathcal{G}}(e)$ be the set of all classes of instance $e$ in $\mathcal{G}$, defined as $c_{\mathcal{G}}(e) = \{c|(e, \text{rdf:type}, c) \in \mathcal{G}\}$. Then, we can state:

$$m_{conClass} = \frac{|\{(c_1, c_2) \in \mathcal{C}| \neg\exists e : (c_1 \in c_{\mathcal{G}}(e)) \wedge c_2 \in c_{\mathcal{G}}(e))\}|}{|\{(c_1, c_2) \in \mathcal{C}\}|} \tag{8}$$

For the evaluation of this metric, the existing restrictions between the classes used to model the data in the graph are determined. In none of the cases were restrictions of the type *owl:disjointWith* encountered. However, it was found that there is not the same entity modeled with two or more different classes. The following SPARQL query was designed to verify that an entity is not at the same time of type *fabio:Journal* and *fabio:JournalArticle* is shown. No entity with this characteristic was found.

```
PREFIX fabio: <http://purl.org/spar/fabio/>
SELECT COUNT(?entity) as ?total WHERE{
?entity rdf:type fabio:Journal.
?entity rdf:type fabio:JournalArticle.
}
```

*Consistency of Statements with Respect to Relation Constraints:* this metric measures the extent to which the instance data are consistent with the relation restrictions.

$$m_{conRelat} = \frac{1}{n}\sum_{i=1}^{n} m_{conRelat,i}(\mathcal{G}) \tag{9}$$

For the evaluation of this metric, the relationships between classes (ObjectProperty) existing in the graph were determined. For each one of them, the range is determined and it is verified that the instances of classes existing in the graph comply with the range restrictions of the corresponding properties. For example, the *dc:creator* property has the class *fabio:JournalArticle* as the range and the class *foaf:Person* as domain, indicating that a person is the author of a scientific article. To verify this restriction, the following SPARQL query was designed, obtaining that all instances are correct.

```
PREFIX dc: <http://purl.org/dc/elements/1.1/>
SELECT (COUNT(?x) as ?total) ?rangeType WHERE{
?x dc:creator ?o.
?o a ?rangeType
}
GROUP BY ?rangeType
```

### 3.4 Ease of Understanding

The ease of understanding is the degree to which data are understood, readable and clear [19]. In the context of a digital library, this is focused on users and addresses issues such as using textual descriptions and descriptive Uniform Resource Identifiers (URIs). Since most of libraries are local or national, they often provide their content in a single language. The ease of understanding is measured by means of the following four metrics:

*Description of Resources:* Repositories based on semantic web principles may use basic properties (for instance, *rdfs:label* and *rdfs:comment*) to describe resources. Formally, let $\mathcal{P}_{IDesc}$ be the set of relations that contains a label or description and $U_{\mathcal{G}}^{\text{local}}$ be the set of all URIs in $\mathcal{G}$ with local namespace.

$$m_{Descr} = \frac{|\left\{u|u \in U_{\mathcal{G}}^{\text{local}} \wedge \exists(u,p,o) \in \mathcal{G} : p \in \mathcal{P}_{IDesc}\right\}|}{|\left\{u|u \in U_{\mathcal{G}}^{\text{local}}\right\}|} \tag{10}$$

In all RDF graphs generated by the LinkedDL model, the *rdfs:label* properties are used to describe all the resources. In the case of authors, the label stores their full name, in the case of journals, the label stores the title of the journal, and in the case of articles, their title is stored.

*Labels in Multiple Languages:* this metric measures whether labels in additional languages are provided.

$$m_{Lang} = \begin{cases} 1 & \text{labels provided in at least one additional language} \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

The bibliographic metadata published by LinkedDL were harvested from Cuban journals that use the OAI-PMH protocol for the exchange of metadata. Although this protocol supports the export of metadata in several languages, Cuban journals only publish metadata in Spanish, so RDF graphs currently do not have labels in more than one language.

*Understandable RDF Serialization:* this metric measures the use of alternative encodings that are more understandable for humans than RDF/XML format, such as N-Triples, N3 and Turtle.

$$m_{uSer} = \begin{cases} 1 & \text{other RDF serializations than RDF/XML format available} \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

All the RDF graphs generated by the model are serialized in several formats, such as N3, TTL, HDT, and XML. The graphs and their different serializations can be downloaded from the project website.

*Self-describing URIs:* self-descriptive URIs contain a readable description of the entity rather than identifiers, and they help users to understand the resource.

$$m_{mURI} = \begin{cases} 1 & \text{self-describing URIs always used} \\ 0.5 & \text{self-describing URIs partly used} \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

During the URI design stage for the construction of the RDF graphs generated by the model, best practices existing in the literature were adopted. It was taken into account that the URIs contain some term or keyword that reflects the type of the entity it describes. For example, the URIs that describe journals have the structure https://data.infocientia.com/resource/journal/2307-2113. In all of them, the term *journal* appears to denote that the URI identifies a journal in the context of the RDF graph. On the other hand, universal identifiers were used, in this case, the ISSN. This ensures that, regardless of whether or not the semantic description of the entity changes, the URI that identifies it never changes, because the ISSN of a journal does not change regularly.

### 3.5 Interoperability

According to Färber et al. [9], the interoperability is calculated using the following metrics:

*Avoiding Blank Nodes and RDF Reification:* this metric allows to evaluate if there are blank nodes and triples *rdf:Statement* in the RDF graph. The equation that defines the metric is shown below.

$$m_{reif} = \begin{cases} 1 & \text{no blank nodes and no reification} \\ 0.5 & \text{either no blank nodes or noreification} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

The generated RDF graph was modeled with ontologies and vocabularies developed by third parties, avoiding the reification of the graph by using *rdf:Statement*, considering that it is a non-recommended practice in these cases. To check the existence or not of blank nodes in the generated RDF graph, the following SPARQL query was designed and executed, obtaining false in all the results. The query uses the *isBlank* function of the SPARQL language, which returns true if there are blank nodes in the RDF graph and false otherwise. Based on the definition of the metric, the resulting value after the evaluation is 1.

```
SELECT ?s ?p ?o ?blankTest
WHERE {
?s ?p ?o.
BIND(isBlank(?o) as ?blankTest)
}
```

*Provisioning of Several Serialization Formats:* In the interoperability process, the format (s) in which the data is exchanged between information systems play an important role. This metric assesses the support offered by the digital library to serialize the RDF graphs in one or more interchange formats. The equation that defines the metric is shown below.

$$m_{iSerial} = \begin{cases} 1 & \text{RDF/XML and further formats are supported} \\ 0.5 & \text{only RDF/XML is supported} \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

In the proposed approach, the use of the Virtuoso triplestore was adopted. This tool supports the publication of the RDF graphs using a SPARQL Endpoint to perform queries on the stored RDF graphs. Virtuoso supports several RDF graph serialization formats, such as: RDF/XML, Turtle, JSON-LD and N-Triples, all of which are W3C standards. Based on the definition of the metric, the resulting value after the evaluation is 1.

*Using External Vocabulary:* this metric evaluates the use of ontologies and external vocabularies by dividing the number of triples whose predicate uses an external vocabulary by the total number of triples existing in the RDF graph. The equation that defines the metric is shown below.

$$m_{extVoc} = \frac{|\{(s,p,o)|(s,p,o) \in \mathcal{G} \wedge p \in \mathcal{P}_{\mathcal{G}}\}|}{|\{(s,p,o) \in \mathcal{G}\}|} \tag{16}$$

where $\mathcal{P}_{\mathcal{G}}$ is the set of external properties to the graph $\mathcal{G}$.

To evaluate the metric, two SPARQL queries were designed that obtain the existing classes and properties in the RDF graph. Three classes and 14 relationships were obtained. Both the classes and the properties obtained belong to ontologies and vocabularies designed by third-parties and reused in the solution proposal. Based on the definition of the metric, the resulting value after the evaluation is 1. The corresponding SPARQL queries are shown below.

```
SELECT DISTINCT ?type
WHERE {
?subject a ?type.
}

SELECT DISTINCT ?property
WHERE {
?subject ?property ?object.
}
```

*Interoperability of Proprietary Vocabulary:* this metric calculates the fraction of classes and properties with at least one equivalence link to classes and properties in external vocabularies. The equivalences can be declared through the properties *owl:sameAs*, *owl:equivalentClass*, *rdfs:subPropertyOf* or *rdfs:subClassOf*. Let $\mathcal{P}_{eq}$ $=\{$owl:sameAs; owl:equivalenClass; rdfs:subPropertyOf; rdfs:subClassOf$\}$ and $U_{\mathcal{G}}^{\text{ext}}$ consists of all URIs in $U_g$ which are external to the graph $\mathcal{G}$, we can state:

$$m_{propVoc} = \frac{|\left\{x \in \mathcal{P}'_{\mathcal{G}} \cup C_{\mathcal{G}} | \exists (x,p,o) \in \mathcal{G} : (p \in \mathcal{P}_{eq} \wedge (o \in U \wedge o \in U_{\mathcal{G}}^{\text{ext}}))\right\}|}{|\mathcal{P}_{\mathcal{G}} \cup C_{\mathcal{G}}|} \quad (17)$$

Taking into account the maturity of the existing ontologies in the state-of-the-art and their wide adoption by the producers of linked data in digital libraries, we decided to adopt and reuse existing ontologies and vocabularies in the solution proposal. The reuse of ontologies and vocabularies constitutes one of the recommendations of the scientific community that contributes to increasing interoperability between information systems. Considering that the proposed solution does not use its own ontology, it is decided not to calculate the metric $m_{propVoc}$.

### 3.6 Accessibility

Accessibility is the extent to which data are available or easily and quickly retrievable [19]. Accessibility requires the data to be available through SPARQL endpoints and RDF dumps. SPARQL endpoints also allow the execution of federated queries across different data sets, enhancing and increasing the visibility of the LOD. The accessibility involves a variety of criteria as follows:

*Dereferencing Possibility of Resources:* dereferencing of resources is based on URIs that are resolvable by means of HTTP requests, returning useful and valid information. The dereferencing of resources is successful when an RDF document is returned and the HTTP status code is 200. This metric assesses for a set of

URIs whether dereferencing of resources is successful. Let $\mathcal{U}_g$ be a set of URIs, we can state:

$$m_{Deref} = \frac{|Dereferencable(\mathcal{U}_g)|}{|(\mathcal{U}_g)|} \tag{18}$$

To evaluate this metric, the first 500 existing resources were selected in the RDF graph generated by the LinkedDL model. Then, an HTTP request was made to each of the URIs that identify the resources, obtaining the 200 status code in all cases. In this way, it is verified that all resources are dereferenced. The RDF graph is published using the Linked Data Fragments server[5], a tool that guarantees the dereferencing process of all existing resources in the graph. This run this metric evaluation, we designed a python script available at Google Colab[6].

*Availability of the Digital Library:* This metric assesses the availability of the digital library in terms of uptime. It can be measured using a URI and a monitoring service over a period of time. Let $S_r$ be the number of successful requests and $T_r$ be the total number of requests, then:

$$m_{Avai} = \frac{S_r}{T_r} \tag{19}$$

To evaluate this metric, the Linked Data Fragments server that publishes the RDF graph was monitored for a period of seven days. Every five minutes an HTTP request was made to the server, storing the status code obtained. In all cases, the status code was 200, evidencing the high availability of the service.

*Availability of a Public SPARQL Endpoint:* This metric indicates the existence of a publicly available SPARQL endpoint.

$$m_{SPARQL} = \begin{cases} 1 & \text{SPARQL endpoint publicly available} \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

For storing the RDF graph generated by the LinkedDL model, the Virtuoso Open Source server is used. This server provides a public SPARQL endpoint, which can be consulted at https://data.infocientia.com/sparql.

*Provisioning of an RDF Export:* In addition to the SPARQL endpoint, an RDF data export can be provided to download the whole data set.

$$m_{Export} = \begin{cases} 1 & \text{RDF export available} \\ 0 & \text{otherwise} \end{cases} \tag{21}$$

A web page has been designed and published in order to guarantee the reuse of the RDF graphs. The RDF graphs generated by the project can be downloaded

[5] https://linkeddatafragments.org/.

[6] Google Colab.

from https://yhdelgado.github.io/linkeddl/. This page contains an overview of the dataset, as well as links to files in several formats, such as N-Triples, Turtle, HDT, and RDF/XML.

*Support of Content Negotiation:* This metric assesses the consistency between the RDF serialization format requested (RDF/XML, N3, Turtle and N-Triples) and that which is returned.

$$m_{Negot} = \begin{cases} 1 & \text{content negotiation supported and correct content types returned} \\ 0.5 & \text{content negotiation supported but wrong content types returned} \\ 0 & \text{otherwise} \end{cases} \tag{22}$$

Content negotiation is provided by Linked Data Fragments server, however only HTML and RDF/XML formats are supported.

*Linking HTML Sites to RDF Serializations:* HTML pages can be linked to RDF serializations by adding a tag to the HTML header with the pattern <*link rel* = *'alternate' type* = *'content type' href* = *'URL'*>.

$$m_{HTMLRDF} = \begin{cases} 1 & \text{autodiscovery pattern used at least once} \\ 0 & \text{otherwise} \end{cases} \tag{23}$$

This feature is not currently supported on the Linked Data Fragments server used for publishing RDF graphs.

*Provisioning of Repository Metadata:* The repository can be described using VoID. This metric indicates whether a machine-readable metadata about the data set is available.

$$m_{Meta} = \begin{cases} 1 & \text{machine-readable metadata available} \\ 0 & \text{otherwise} \end{cases} \tag{24}$$

The repository generated by LinkedDL supports the description of it using the VoID vocabulary. The metadata describing the repository can be downloaded at https://data.infocientia.com/void.ttl#linkeddl. This feature is essential for the discovery and reuse of published RDF graphs.

### 3.7 Interlinking

Interlinking is the extent to which entities that represent the same concept are linked to each other, be it within or between two or more data sources. The interlinking dimension measures the number and validity of external links as follows:

*Interlinking via owl:sameAs:* This score is obtained as the rate of instances having at least one owl:sameAs triple pointing to an external resource. Let $\mathcal{I}_\mathcal{G}$ be the set of instances in $\mathcal{G}$ we can state:

$$m_{Inst} = \frac{|\left\{x \in \mathcal{I}_\mathcal{G} | \exists \left\{x, sameAs, y\right\} \in g \wedge y \in U_\mathcal{G}^{\text{ext}}\right\}|}{|\mathcal{I}_g|} \tag{25}$$

To evaluate this metric, the total number of existing instances in the RDF graph is first calculated. For this, the following SPARQL query was used:

```
SELECT (COUNT(*) as ?Instances) WHERE { ?s rdf:type ?o}
```

After executing the query, a total of 25657 instances was obtained at the time of the evaluation. Then, the total number of instances that have links to external data sources was calculated, in all cases the instances have the property *owl:sameAs*. To determine the number of instances, the following SPARQL query was executed:

```
SELECT (COUNT(*) as ?Instances) WHERE { ?s owl:sameAs ?o}
```

After executing the query, 27 instances were obtained, so the metric $m_{Inst}$ gets a value of 0.001. The value of the metric is low, so that in successive iterations the graph will be enriched with links to other existing data sources on the data web.

*Validity of External URIs:* Linking to external resources can lead to invalid links. Given a list of URIs, this criterion checks if there is a timeout or error. Let $\mathcal{A}$ be the set of external URIs, then:

$$m_{URIs} = \frac{|\left\{x \in \mathcal{A} \wedge x \text{ is resolvable}\right\}|}{|\mathcal{A}|} \tag{26}$$

To calculate the value of the metric, an HTTP request was executed on the 27 links to external data sources existing in the graph. In all cases, the status code 200 was obtained.

## 4 Results and Discussion

The Accuracy dimension achieves a high score in all the repositories evaluated. In the particular case of the duplicate entities metric, the LinkedDL dataset achieves the highest score, which means that no duplicate entities were detected in the repository concerning existing entities in Wikidata.

The Trustworthiness dimension is not very high in any of the repositories evaluated. However, the LinkedDL dataset scores higher in the library level trustworthiness metric. This is because the existing repositories in the state-of-the-art perform the automatic conversion to linked data from the legacy data sources without previously being reviewed. In the case of the LinkedDL dataset, the metadata is reviewed before and after conversion.

The Consistency dimension obtains the highest score in all the metrics evaluated in the LinkedDL dataset. In the particular case of the consistency metric of the schema restrictions during the insertion of new instances, a higher value is obtained than the rest of the repositories. This is because a check of the schema restrictions is performed during the insertion of new instances into the repository.

The Interoperability dimension obtained the same scores as the evaluated repositories, except for the interoperability of the proprietary vocabulary metric. Similar scores were obtained in the Accessibility dimension, where values similar to those existing in the state-of-the-art were obtained, except the availability of the repository metric. In the case of the linking HTML sites to RDF serializations metric, the only repository that implements this functionality is the British National Bibliography (BNB).

Assessing the quality of linked data is becoming very hard, due to the size of the knowledge base and the lack of automatic tools to measure some dimensions and metrics, among other factors. Also, the majority of the used dimensions and metrics are focused on evaluating the syntaxis and semantic of the linked data generated. However, the research community must define new metrics for assessing the usability of the linked data in real-world scenarios.

**Table 2.** Comparative table of five datasets. BNE: Biblioteca Nacional de España; BNF: Bibliothèque nationale de France; BNB: British National Bibliography; BVMC: Biblioteca Virtual Miguel de Cervantes. Partial results from [4]

| Dimension | Metric | BNE | BNF | BNB | BVMC | LinkedDL |
|---|---|---|---|---|---|---|
| Accuracy | Syntactic validity of RDF documents | 1 | 1 | 1 | 1 | 1 |
| | Syntactic validity of literals | 1 | 1 | 0.99 | 1 | 1 |
| | Semantic validity of triples | 1 | 1 | 1 | 1 | 1 |
| | Check of duplicate entities | 0.99 | 0.99 | 0 | 0.96 | **1** |
| Trustworthiness | On library level | 0.25 | 0.25 | 0.25 | 0.25 | **0.75** |
| | On statement level | 0 | 0 | 0 | 0 | 0 |
| | Using unknown and empty values | 0 | 0 | 0 | 0 | 0 |
| Consistency | Consistency of schema restrictions during insertion of new statements | 0 | 0 | 0 | 0 | **1** |
| | Consistency of statements with respect to class constraints | 1 | 1 | 1 | 1 | 1 |
| | Consistency of statements with respect to relation constraints | 0.98 | 1 | 1 | 1 | 1 |
| Ease of understanding | Description of resources | 0.93 | 0.91 | 0.89 | 0.92 | **1** |
| | Labels in multiple languages | 0 | 1 | 0 | 0 | 0 |
| | Understandable RDF serialization | 1 | 1 | 1 | 1 | 1 |
| | Self-describing URIs | 1 | 1 | 0 | 1 | 1 |
| Interoperability | Avoiding blank nodes and RDF reification | 1 | 1 | 1 | 1 | 1 |
| | Provisioning of several serialization formats | 1 | 1 | 1 | 1 | 1 |
| | Using external vocabulary | 0.53 | 0.69 | 0.90 | 1 | 1 |
| | Interoperability of proprietary vocabulary | 0.81 | 0.85 | 0.35 | 1 | 0 |
| Accessibility | Dereferencing possibility of resources | 1 | 1 | 1 | 1 | 1 |
| | Availability of the repository | 0.86 | 0.99 | 1 | 0.99 | **1** |
| | Availability of a public SPARQL endpoint | 1 | 1 | 1 | 1 | 1 |
| | Provisioning of an RDF export | 1 | 1 | 1 | 0 | **1** |
| | Support of content negotiation | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 |
| | Linking HTML sites to RDF serializations | 0 | 0 | 1 | 0 | 0 |
| | Provisioning of metadata | 0 | 0 | 1 | 1 | 1 |
| Interlinking | Interlinking via *owl:sameAs* | 0.07 | 0.39 | 0.17 | 0.04 | 0.001 |
| | Validity of external URIs | 1 | 1 | 1 | 1 | 1 |

## 5 Conclusions and Future Work

Assessing the quality of linked data is becoming very hard, due to the size of the knowledge base and the lack of automatic tools to measure some dimensions and metrics, among other factors. In this paper, we evaluated the quality of data of the LinkedDL dataset. The comparison with similar approaches in the literature shows that the LinkedDL dataset improves those reported in the state-of-the-art in terms of accuracy, consistency, and accessibility metrics. This assessment is useful for data consumers that need to enrich their collections based on accuracy, consistency, and accessibility metrics. Future work includes the improvement of the LinkedDL dataset, taking into account the quality metrics evaluated with low scores, and the formalization of new quality metrics for assessing the usability of the linked data generated in real-world scenarios.

**Acknowledgments.** We thank the anonymous reviewers for their careful reading of our manuscript and their many insightful comments and suggestions.

## References

1. Agosti, M., Ferro, N., Silvello, G.: Digital library interoperability at high level of abstraction. Future Gener. Comput. Syst. **55**, 129–146 (2016). https://doi.org/10.1016/j.future.2015.09.020. http://www.sciencedirect.com/science/article/pii/S0167739X15003003
2. Barroso, I., Hartmann, N., Ribeiro, C.: Metadata crosswalk for a museum collection in a thematic digital library. J. Libr. Metadata **15**(1), 36–49 (2015). https://doi.org/10.1080/19386389.2015.1011025
3. Berners-Lee, T.: Linked Data - Design Issues (2006). https://www.w3.org/DesignIssues/LinkedData.html
4. Candela, G., Escobar, P., Carrasco, R.C., Marco-Such, M.: Evaluating the quality of linked open data in digital libraries. J. Inf. Sci., 0165551520930951 (2020). https://doi.org/10.1177/0165551520930951
5. Candela, G., Escobar, P., Sáez, M.D., Marco-Such, M.: A shape expression approach for assessing the quality of linked open data in libraries. In: Semantic Web Preprint(Preprint), pp. 1–21. IOS Press, January 2021. https://doi.org/10.3233/SW-210441. https://content.iospress.com/articles/semantic-web/sw210441
6. Chen, Y.N.: A RDF-based approach to metadata crosswalk for semantic interoperability at the data element level. Library Hi Tech **33**(2), 175–194 (2015). https://doi.org/10.1108/LHT-08-2014-0078
7. Debattista, J., Auer, S., Lange, C.: Luzzu-A methodology and framework for linked data quality assessment. J. Data Inf. Qual. **8**(1), 4:1–4:32 (2016). https://doi.org/10.1145/2992786
8. Debattista, J., Lange, C., Auer, S., Cortis, D.: Evaluating the quality of the LOD cloud: an empirical investigation. Semant. Web **9**(6), 859–901 (2018). https://doi.org/10.3233/SW-180306. https://content.iospress.com/articles/semantic-web/sw306
9. Färber, M., Bartscherer, F., Menne, C., Rettinger, A.: Linked data quality of DBpedia, freebase, OpenCyc, Wikidata, and YAGO. Semant. Web **9**(1), 77–129 (2018). https://doi.org/10.3233/SW-170275

10. Hidalgo-Delgado, Y., Xu, B., Mariño-Molerio, A.J., Febles-Rodríguez, J.P., Leiva-Mederos, A.A.: A linked data-based semantic interoperability framework for digital libraries. Revista Cubana de Ciencias Informáticas **13**(1), 14–30 (2019). https://rcci.uci.cu/?journal=rcci&page=article&op=view&path
11. Khan, N.A., Shafi, S., Rizvi, S.Z.: Metadata crosswalks as a way towards interoperability. In: Encyclopedia of Information Science and Technology, 3rd edn., pp. 1834–1842. Data Mining and Databases. IGI Global (2015). https://www.igi-global.com/chapter/metadata-crosswalks-as-a-way-towards-interoperability/112589
12. Martín, A., León, C., López, A.: Enhancing semantic interoperability in digital library by applying intelligent techniques. In: SAI Intelligent Systems Conference, pp. 904–911. IEEE (2015)
13. Mecella, M., Scannapieco, M., Virgillito, A., Baldoni, R., Catarci, T., Batini, C.: Managing data quality in cooperative information systems. In: Meersman, R., Tari, Z. (eds.) OTM 2002. LNCS, vol. 2519, pp. 486–502. Springer, Heidelberg (2002). https://doi.org/10.1007/3-540-36124-3_28
14. Radulovic, F., Mihindukulasooriya, N., García-Castro, R., Gómez-Pérez, A.: A comprehensive quality model for linked data. Semant. Web **9**(1), 3–24 (2018). https://doi.org/10.3233/SW-170267. https://content.iospress.com/articles/semantic-web/sw267
15. Shreeves, S.L., Knutson, E.M., Stvilia, B., Palmer, C.L., Twidale, M.B., Cole, T.W.: Is quality metadata shareable metadata? The implications of local metadata practices for federated collections. In: Proceedings of the Twelfth National Conference of the Association of College and Research Libraries. Association of College and Research Libraries (2005). https://www.ideals.illinois.edu/handle/2142/145. Accepted 2006-10-19T21:25:14Z
16. Singh, T., Sharma, A.: Research work and changing dimensions of digital library. In: Emerging Trends and Technologies in Libraries and Information Services, pp. 39–42. IEEE (2015)
17. Tallerås, K.: Quality of linked bibliographic data: the models, vocabularies, and links of data sets published by four national libraries. J. Libr. Metadata **17**(2), 126–155 (2017). https://doi.org/10.1080/19386389.2017.1355166
18. Villazón-Terrazas, B., Vilches-Blázquez, L.M., Corcho, O., Gómez-Pérez, A.: Methodological guidelines for publishing government linked data. In: Wood, D. (ed.) Linking Government Data, pp. 27–49. Springer, New York (2011). https://doi.org/10.1007/978-1-4614-1767-5_2. http://www.w3.org/TR/ld-bp/
19. Wang, R.Y., Strong, D.M.: Beyond accuracy: what data quality means to data consumers. J. Manage. Inf. Syst. **12**(4), 5–33 (1996). https://doi.org/10.1080/07421222.1996.11518099. https://www.tandfonline.com/doi/abs/10.1080/07421222.1996.11518099
20. Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S.: Quality assessment for linked data: a survey. Semant. Web **7**(1), 63–93 (2016). http://content.iospress.com/articles/semantic-web/sw175