Semantic Web 0 (2017) 1–0
IOS Press

Evaluating the Quality of the LOD Cloud: An Empirical Investigation

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

Jeremy Debattista [a,b,*], Christoph Lange [b] and Sören Auer [b] and Dominic Cortis [c,d]
[a] *ADAPT Centre, School of Computer Science and Statistics, Trinity College Dublin, Ireland*
E-mail: jerdebattista@gmail.com
[b] *Fraunhofer IAIS, Schloss Birlinghoven, 53754 Sankt Augustin, Germany*
E-mail: langec@iai.uni-bonn.de, auer@iai.uni-bonn.de
[c] *Department of Mathematics, University of Leicester, College of Science, University Rd, Leicester LE1 7RH UK*
[d] *Faculty of Economics, Management and Accountancy, University of Malta, Msida MSD 2080, Malta*
E-mail: dc156@leicester.ac.uk

Abstract. The increasing adoption of the Linked Data principles brought with it an unprecedented dimension to the Web, transforming the traditional Web of Documents to a vibrant information ecosystem, also known as the Web of Data. This transformation, however, does not come without any pain points. Similar to the Web of Documents, the Web of Data is heterogenous in terms of the various domains it covers. The diversity of the Web of Data is also reflected in its quality. Data quality impacts the *fitness for use* of the data for the application at hand, and choosing the right dataset is often a challenge for data consumers. In this quantitative empirical survey, we analyse 130 datasets ($\approx$ 3.7 billion quads), extracted from the latest Linked Open Data Cloud using 27 Linked Data quality metrics, and provide insights into the current quality conformance. Furthermore, we publish the quality metadata for each assessed dataset as Linked Data, using the Dataset Quality Vocabulary (daQ). This metadata is then used by data consumers to search and filter possible datasets based on different quality criteria. Thereafter, based on our empirical study, we present an aggregated view of the Linked Data quality in general. Finally, using the results obtained from the quality assessment empirical study, we use the Principal Component Analysis (PCA) test in order to identify the key quality indicators that can give us sufficient information about a dataset's quality. In other words, the PCA helps us identify the non-informative metrics.

Keywords: Data Quality, Linked Data, Empirical Study, Data Quality Survey

1. Introduction

Since its inception, the *Linked Open Data (LOD) Cloud* [48] has been a point of reference to the Linked Data community, comprising a number of linked datasets crawled on the Web of Data or added to the *LODCloud* group in the *datahub.io* registry[1]. The maintainers provide a set of criteria for the inclusion of a dataset within the LOD Cloud; more specifically, datasets should be published according to the Linked Data principles as defined in [11]. The Linked Data principles, closely related to the five star scheme for publishing open data[2], can be summarised as *publishing structured, interlinked data, in non-proprietary formats, using URIs*.

This widespread and rapid adoption of the Linked Data principles has brought an unprecedented dimen-

*Corresponding author. E-mail: jerdebattista@gmail.com
[1] `https://datahub.io/group/lodcloud`

[2] See `http://5stardata.info` and [11]

sion on the Web, contributing to the transformation of the Web of Documents to a Web of Data. Thanks to links between the data, one can jump from one source to another in order to retrieve more complete information and answers. Similarly to the Web of Documents, these sources, heterogeneous with regard to their domain, have highly varying quality [27]. Document quality is often only subjectively assessable, and indirect measures such as Page Rank and HITS (hubs and authorities), which calculate the importance of a document vis-à-vis the Web (via links), give a good indication of whether a document is of good quality or a good authoritative source. In a parallel situation, resources in the Web of Data are not simply text (or other HTML components such as tables, images) and links. For LOD datasets, indirect link related quality measures are much less meaningful, (since linked datasets are prone to link spamming [24]) but at the same time a number of other more direct quality indicators exist.

Linked Data resources are usually complex structures encompassing some existing thing (an object in the real world), giving it semantics (i.e. meaning) and possibly linking to other resources, that both machines and humans can understand. According to the editors of the W3C Data on the Web Best Practices document,

> *"data quality can affect the potentiality of the application that uses data, as a consequence, its inclusion in the data publishing and consumption pipelines is of primary importance."* – [34, §9.5]

Making data quality more transparent and easy-to-access is a key factor for the wider penetration of Linked Data and semantic technologies. In this study, the research question we aim to answer is:

What is the quality of existing Data on the Web?

To answer this question, we perform a large scale evaluation of Linked Data quality in terms of data size, domain and quality indicator coverage. More specifically we assess and quantify the quality of 130 datasets ($\approx$ 3.7 billion triples) in the Linked Open Data Cloud over a number of quality indicators, as classified in [55]. Furthermore, such an investigation leads to other insights, such as identifying which of the assessed metrics are the most informative to describe the quality of a linked dataset (cf. Section 6.2).

Using Luzzu [17], a quality assessment framework for Linked Data, and a number of quality metrics (including some probabilistic approximation metrics), this study produces a quality metadata graph for each assessed dataset (publicly available for consumption as Linked Data resources), represented in terms of the Dataset Quality Vocabulary (daQ) [18][3]. The benefits of these metadata graphs are two-fold: (1) humans can understand the quality of a dataset better, using ranking or visualisation tools, thus making more informed decisions prior to using a dataset; and (2) machines can automatically process the quality metadata of a dataset.

The remainder of this article is structured as follows. We first discuss related work regarding analysis of various aspects of Linked Data (Section 2). In Section 3 we perform a primary investigation towards the *openness* of the Linked Open Data, followed by the dataset acquisition description in Section 4. Following the data acquisition process, in Section 5 we assess and discuss the quality of these datasets against 27 metrics related to four different quality categories as described in [55]. We then use the assessment results in order to identify the non-informative quality metrics in Section 6, followed by the conclusions in Section 7.

2. Studying the Quality of the Data on the Web

Empirical studies encourage stakeholders to engage in further discussions and enable them to improve the current state of the discussed topic, in this case of the quality of linked datasets. The main contribution of this study is a large-scale analysis of the quality of linked open datasets. We assess various data dumps, SPARQL endpoints that are available and portrayed in the 2014 LOD Cloud snapshot over 27 quality metrics related to different inherent and extrinsic aspects of Linked Data. In this section, we review literature that analyse the quality of various aspects of Linked Data, as a prequel to the large-scale analysis described in this article.

Schmachtenberg et al. [49] crawled the Web of Data in order to present the 2014 version of the LOD cloud diagram. Each crawled dataset was categorised in a topical domain, whose categorisation was then

[3] Luzzu's underlying semantic framework uses daQ, which has existed prior to the more recent Data Quality Vocabulary (DQV) recommended by the W3C [2]. But since properties and classes from both daQ and DQV are marked as equivalent, it is easy for reasoners to transform instances from one vocabulary to another.

used in one of our metrics, *re-use of existing terms* (Metric IO1). Furthermore, during this study, the authors also analysed how different best practices were adopted in the crawled linked datasets. Some of these best practices overlap with the quality metrics presented in our study, including best practices related to the adoption of vocabularies and metadata. In our work, we use the 2014 version of the LOD cloud to better understand the quality of the Web of Data.[4]

Throughout the years, a number of researchers in Linked Data quality have come up with numerous quality metrics that were consolidated in a systematic survey by Zaveri et al. in [55]. The authors of this systematic survey group 69 different quality metrics in 18 dimensions and four categories: *Accessibility*, *Intrinsic*, *Contextual*, and *Representational*. Our empirical investigation towards the quality of the Web of Data complements the survey undertaken in [55], with 27 out of the 69 described metrics being implemented and assessed over a number of datasets (cf. Section 5.1).

In [28], Hogan et al. crawled and assessed the quality of around 12 million RDF statements. The main aim was to discuss common problems found in RDF datasets, and possible solutions. More specifically, this work aimed at uncovering errors related to accessibility, reasoning, syntactical and non-authoritative contributions. The authors also provided suggestions on how publishers can improve their data, so that the consumers can find "higher quality" datasets.

In a follow up article [30], Hogan et al. conducted a larger empirical study on Linked Data conformance, with around 1 billion quads (i.e. triples + graph identifier) assessed. The aim of this study was primarily to define a number of quality metrics from various best practices and guidelines, and to assess the level of conformance of the assessed datasets against these metrics. The quality metrics considered in our work overlap with seven metrics defined in [30]: (i) avoiding blank nodes; (ii) keeping URIs short; (iii) avoiding prolix features; (iv) re-using existing terms; (v) dereferenceability of resources; (vi) usage of external URIs; and (vii) human-readable metadata. The metrics in our assessment are similar to those in [30], with some adjustments as we explain in Section 5. Apart from a larger set of quality metrics in this article, one must point out that two corpuses are different, where in [30] the authors used data crawled from the Web of Data. Nevertheless, the conclusions from [30] are more or less the same, four years later, that publishers might forgo certain quality guidelines as they might be impractical. This can be seen from the distribution of quality metric values amongst the datasets, in both studies.

Buil-Aranda et al. [4] conducted a number of long-term experiments, mostly related to availability quality (extrinsic) metrics on around 480 SPARQL endpoints. The authors report that only one third of the endpoints have descriptive metadata such as VoID and service descriptions[5], whilst the query response performance varies widely from one endpoint to another. Our experiments confirm the performance variation and show that no single solution is available for streaming all triples directly from the endpoint (cf. Section 5.6). The authors also propose SPARQLES[6], a tool for monitoring the availability of public SPARQL endpoints (among other tests). With SPARQLES, consumers can make informed decisions more easily on whether a certain SPARQL endpoint is reliable and suitable for the task at hand.

In a recent study Assaf et al. [6] shed light on the quality of the metadata of datasets available in the Linked Open Data Cloud. This metadata was used in our dataset acquisition process. In [6], the metadata is checked for general, access, ownership and provenance information. The authors concluded, that metadata quality is in a bad condition. More specifically, licensing and accessibility metadata contains noisy data, resulting in incorrect information. We discuss the quality of LOD Cloud metadata in more detail in Section 3.

Suominen and Mader [51], define a number of quality metrics in order to assess SKOS vocabularies with the aim of identifying their re-use in applications. The assessment is based on three categories: (i) labelling and documentation; (ii) structural issues (e.g. class disjointness issues); and (iii) Linked Data issues (e.g. invalid URIs). The authors reported that most of their representative SKOS vocabularies contained structural errors, and presented a set of correction algorithms to address such issues. These issues discussed in [51] are also relevant to linked datasets, which we also discuss in this article, however in Suominen and Mader's article these intrinsic and extrinsic aspects are discussed in light of Linked Data vocabularies, more specifically SKOS-driven vocabularies.

[4] A more recent version of the LOD cloud was released on 20 February 2017, i.e., after we had conducted our study.

[5] http://www.w3.org/TR/sparql11-service-description/

[6] http://sparqles.ai.wu.ac.at/

In [22], Giménez-García et al. focus on dataset reuse to assess the trust of Linked Datasets. The authors use LOD Laundromat data dumps in order to compute PageRank values on datasets and rank these datasets based on their *trustworthiness* value. Results show that popular datasets such as DBpedia and Geonames feature in the top 10, however their approach also captures services such as `purl.org`, which hosts multiple datasets. In this paper we assess trustworthiness from a different aspect, by analysing provenance information of a dataset.

Meusel and Paulheim in [38] analyse a number of quality issues in schema.org for Microdata and compare the results with the findings using linked datasets in [28]. The metrics studied in [38] include: (i) usage of undefined types and properties; (ii) misuse of datatype or object properties; (iii) violation of datatypes; and (iv) incorrect usage of property domain and range. The analysis performed in [38] shows that the Microdata formats adopting schema.org are prone to be less problematic with regards to undefined elements than LOD, however fares worse in the other three metrics against the LOD results analysed in [28]. In this paper we will discuss these metrics and assess them against the 2014 version of the LOD cloud.

3. 'O'penness in the Linked Open Data Cloud

Having metadata as part of a published dataset is the first step in putting a dataset on the open data map (thus encouraging discoverability [45]), as it is generally the first access point for consumers who wish to use the published data. Metadata ensures that it complies with best practices by making it self-descriptive [26, §5.5]. Therefore, 'doing metadata right' is a must for any kind of published open data. In a holistic assessment of open government data initiatives, Attard et al. [7] describe a number of initiatives that had the aim to assess the quality of metadata. This shows further the importance metadata is given in open data.

Open Data, in terms of the *Open Definition* should be able to

> *... be freely used, modified, and shared by anyone for any purpose* [41].

More specifically, open data should [41, §1]:

1. have a defined open license or status – having a license is the only way to define boundaries between the publisher and the consumer (who can also re-publish the data without worrying about using the data improperly);
2. be accessible, i.e. in the case of Linked Open Data a dataset should have some entry point such as a data dump or SPARQL endpoint (preferably referred to in dataset metadata defined by standard vocabularies);
3. be machine readable, if possible interoperable, for example, by using RDF;
4. have an open format.

Drawing parallels with Linked Open Data, Berners-Lee proposed the five-star open data principles, in which the first three stars are similar to the principles of the Open Definition, whilst the last two are more related to the Linked Data principles, i.e. (4th star) using URIs to identify things, and (5th star) linking between the published data and external data [11].

Heath and Bizer [26] provide a checklist for Linked Data publishing, which includes the provision of provenance metadata, licensing metadata, and dataset level metadata in terms of standard vocabularies such as VoID [3] or DCAT [36]. Schemas like DCAT and VoID enable metadata description in a semantically interoperable format and can be exchanged between various agents. Currently, there are further schema initiatives, including the Dataset Quality Vocabulary (daQ) [18] and the compatible W3C Data Quality Vocabulary (DQV) [2] to represent quality metadata for datasets, and the W3C Data Usage Vocabulary (DUV) [35] to describe various factors of a dataset such as citation and feedback from a human consumer perspective.

Our study is based on the LOD Cloud snapshot that was taken in 2014, containing about 188 million crawled triples[7]. Metadata description of these datasets can be retrieved easily from the Linked Data catalog[8] published together with the 2014 snapshot. Whilst the API of the CKAN data management system, which drives *datahub.io* and other data catalogs, includes a metadata export functionality in terms of DCAT [36], metadata of new datasets imported to the catalog is generally manually added as a textual description and thus prone to errors such as inconsistency and duplication. For example, in the *formats* tags[9], we find a vari-

[7] This number was taken from `http://data.dws.informatik.uni-mannheim.de/lodcloud/2014/ISWC-RDB/`, although the actual number of triples in the referred datasets is larger.

[8] `https://datahub.io/group/lodcloud`

[9] For this analysis we use `https://datahub.io/group/lodcloud` catalogue.

ety of tags referring to the same format (the number in brackets refer to the number of datasets tagged):

- application/rdf+xml (17); application/rdf xml (4);
- api/sparql (368); sparql (4);
- text/turtle (75); ttl (10); rdf/turtle (7); turtle (2);

We find also a number of tags that we could not match with an appropriate format or else tags with formats of a proprietary nature, for example:

- RDF (187) [possibly application/rdf+xml, but this had to be verified manually];
- xhtml, rdf/xml, turtle (2) [this is one tag with three possible formats];
- example/* (2);
- mapping/twc-conversion (5)

Having a variety of formats in such metadata would hinder the potential re-use of datasets by automated agents as they would not be able to decipher the type of data in question automatically. In order to follow the best Linked Open Data practices, such metadata should be standardised and interoperable between different machines, for example, the use of ontologies such as the *Media Types as Linked Data ontology* [44] should be considered in order to standardise the metadata practices between datasets within a catalog.

3.1. LOD Cloud Datasets' Accessibility

In order to identify which datasets had some kind of access point, an initial experiment was performed on the 2014 LOD Cloud snapshot[10,11]. The LOD Cloud snapshot has a total of 569 datasets. According to the metadata provided in the DataHub, only around 42% of them (239 datasets) had a possible[12] Linked Data access point, i.e., a data dump URI, SPARQL endpoint, or a VoID dataset description. From these 239 datasets, 50% had multiple access points, 33 datasets only had a data dump defined, 74 had a SPARQL endpoint, whilst 13 datasets had just a VoID description URI defined. Figure 1 depicts datasets from the LOD Cloud snapshot that are actually accessible.

[10] `http://lod-cloud.net/versions/2014-08-30/lod-cloud.svg`

[11] These initial experiments were performed in December 2015, prior to the actual quality assessments. This was part of the data acquisition process which is described in Section 4.

[12] We added a validation stage which is described in Section 4.

3.2. LOD Cloud Datasets' Licenses and Rights

Licences are the heart of Open Data. They define whether third parties can re-use data or otherwise, and to what extent. In Linked Open Data, one would expect that such licenses are machine readable using predicates such as `dct:license`, `dct:rights` and `cc:licence`, and possibly also in a human readable format (e.g. within `dc:description`). Such license specification should also be included in a dataset's metadata. Another initial experiment was performed on the LOD Cloud snapshot to check how many datasets metadata provide some kind of machine readable license in their datahub.io metadata. For retrieving machine readable licenses we use the DCAT metadata attached to each dataset in the LOD cloud snapshot and look for one of the following basic graph patterns:

1. *?ds dcterms:license ?license .*
2. *?ds dcterms:rights ?rights .*

In total, only 40.42% (230 in total) of all datasets represented in the current LOD Cloud snapshot have some kind of license (or rights) defined in a semantic manner. This is higher than the 9.96% reported by Schmachtenberg et al. in [49], where the authors searched for triples with the dataset itself as the subject and a predicate containing the string "licen" or "rights". In Table 1, we list the licenses used within the LOD Cloud snapshot, with the Creative Commons Attribution License (*cc-by*) being used the most (93 instances), followed by the Creative Commons Attribution Share-Alike License (*cc-by-sa*; 47 instances) and the Creative Commons Attribution Non-Commercial V2.0 License (*cc-by-nc 2.0*; 31 instances). In spirit of the Open Data definition described in the introduction, the *cc-by-nc 2.0* license is deemed as a non-conformant[13] license since it does not support some of the definition's principles, more specifically the principle that Open Data could be re-used for any purpose, including commercial purposes [41, §2.1.8]. It was noted that 7 out of 9 machine-readable licenses URIs used in the dataset's metadata were non-semantic resources, meaning that they cannot be dereferenced to an RDF description. In Linked Data, publishers of such metadata should re-use RDF resources, such as Creative Commons[14] [1] and RDF License[15] [46]

[13] `http://opendefinition.org/licenses/nonconformant/`

[14] `https://creativecommons.org/ns`

[15] `http://purl.org/NET/rdflicense/`

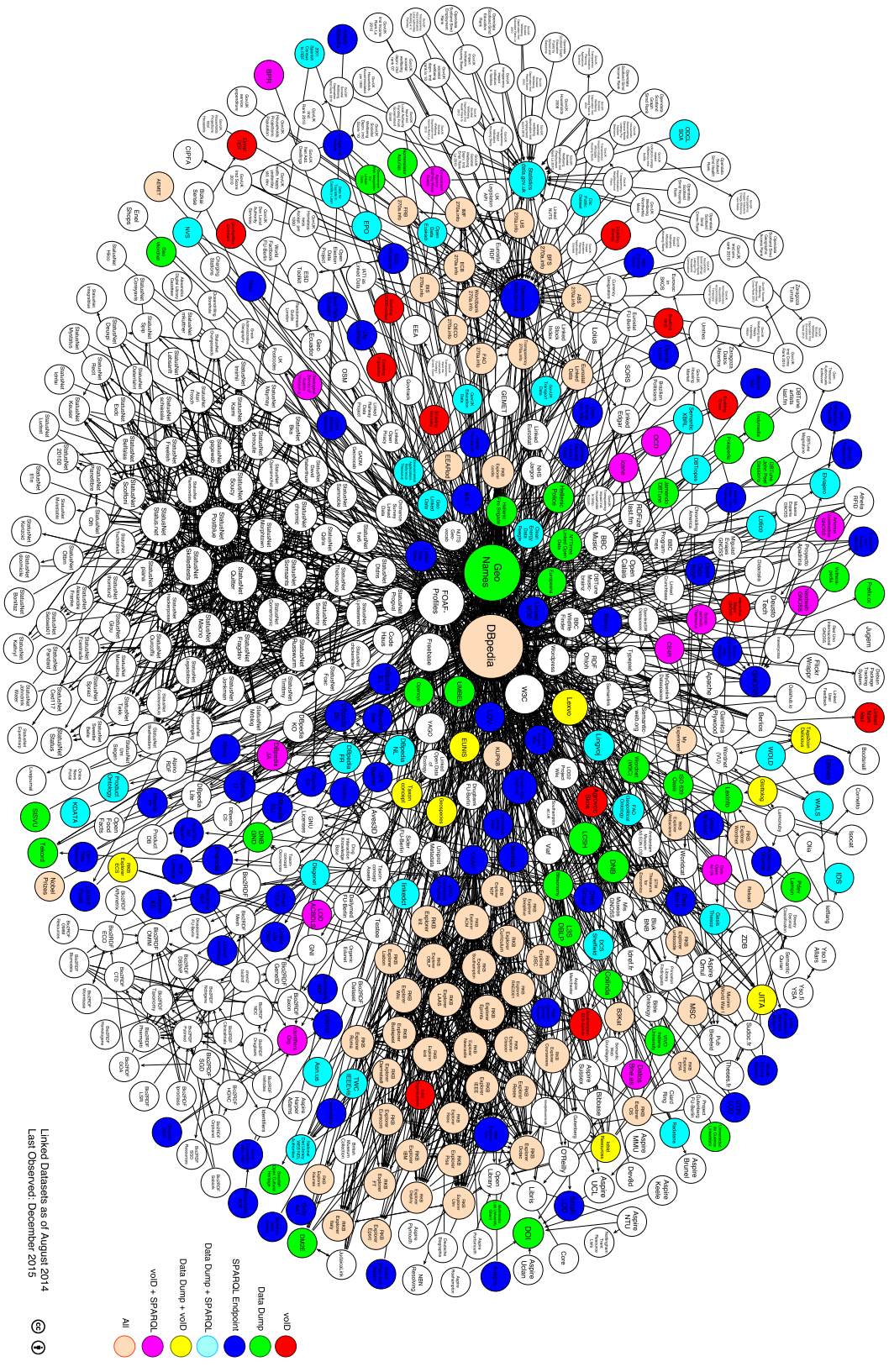

Fig. 1.: Coloring the LOD Cloud Datasets with various Access Methods (Data Dump, VoID, SPARQL Endpoint, or a combination)

A number of data publishers declared the datasets' license (and subsequent rights description) in a human readable manner in the textual description, for example `https://datahub.io/dataset/uniprot`. A regular expression[16] that captures *license* or *copyright* and one of *under*, *grant*, or *right* was performed on all metadata descriptions in order to identify possible license definitions on a dataset. 13 datasets had this kind of human readable license declaration (results displayed in brackets in Table 1). This second experiment identified 5 new licenses used in the LOD Cloud snapshot, two of which (Creative Commons Attribution-NonCommercial-ShareAlike V3.0 and Project Gutenberg License) are non-conformant[17] to open data. Figure 2 shows the datasets with a declared license.

3.3. The LOD Cloud Snapshot and its Future

From our preliminary investigation on the available metadata, we have identified that approximately less than half of the datasets should not be part of the Linked **Open** Data cloud, as they do not satisfy the properties of *Open Data*. Furthermore, the Web of Data, unlike the LOD Cloud snapshot, is volatile. Datasets on the Web, although undesirable, are unpredictable, and thus features, more specifically access points, might not be available on the cloud at all times. Changes in datasets themselves, for example the addition of new external links, could also change the shape of the LOD Cloud as we know it. Such dynamics of the Web of Data are described further in [31] where the authors presented the Dynamic Linked Data Observatory[18], from which a comprehensive analysis over 29 weeks was conducted. Their study show that around 60% of the data(sets) did not change, 5% went offline, whilst the rest had changes in the datasets itself. SPARQLES[19], a tool monitoring the availability of public SPARQL endpoints (amongst other tests), shows that only around 45% were available (from a total of 549 publicly available endpoints monitored) at the time of study[20]. Downtime can be caused by various issues, such as network failures or high server load. Availability statistics, provided by SPARQLES, show that as at November 2015, 181 endpoints (around 32% from 549 endpoints) have a $\geq$ 99% uptime. In April 2015, this number stood at 242, therefore over a period of 6 months, at least 11% of these endpoints became less reliable. Overall, 239 endpoints (around 44% – as at November 2015) are the least reliable, having an uptime of < 5%. In the future, if the LOD Cloud snapshot is to represent the state of the Web of Data, these dynamics should also be considered. Thus, ideally the LOD Cloud snapshot is dynamically updated as datasets are added, die and change.

4. Dataset Acquisition Process

In this section we detail the process for identifying possible datasets that are used for the empirical study. Our main goal was to automate the whole process, whilst retrieving as many datasets as possible. The metadata of the 2014 LOD Cloud was taken as the primary corpus for this study. Each dataset in the LOD Cloud, grouped by their fully qualified domain name (FQDN)[21], has a corresponding generated DCAT metadata entry in the datahub.io portal. Metadata descriptions of these datasets can be easily retrieved from the catalog's Linked Data interface.

4.1. Identifying Datasets' Access Points

For this initial experiment we retrieved the distribution resources (from the `dcat:distribution` property) defined in the dataset metadata (`dcat:Dataset`), in order to identify the media types and corresponding URLs where the dataset is made available for consumption. We aimed to identify the **data dump** (containing all triples of the dataset), a **SPARQL endpoint** description, and a **VoID description** for each dataset. Ideally, a dataset description provides all three resources as this would enable agents to consume the dataset using different processors. Figure 1 shows the LOD Cloud indicating the retrieved datasets and their respective (meta) data access methods, whilst Figure 3 shows an overview of the marking and subsequent retrieval process of the LOD datasets used for the assessment.

[16] `.*(licensed?|copyrighte?d?).*(under|grante?d?|rights?).*`

[17] As identified by OpenDefinition.org `http://opendefinition.org/licenses/nonconformant/`

[18] `http://swse.deri.org/dyldo/`

[19] `http://sparqles.ai.wu.ac.at/`

[20] As of 2nd March 2016

[21] A fully qualified domain name (FQDN) is the complete name for a specific host, for example `de.dbpedia.org` is the FQDN for the German version of DBpedia, whilst `pt.dbpedia.org` is the Portuguese version of DBpedia.

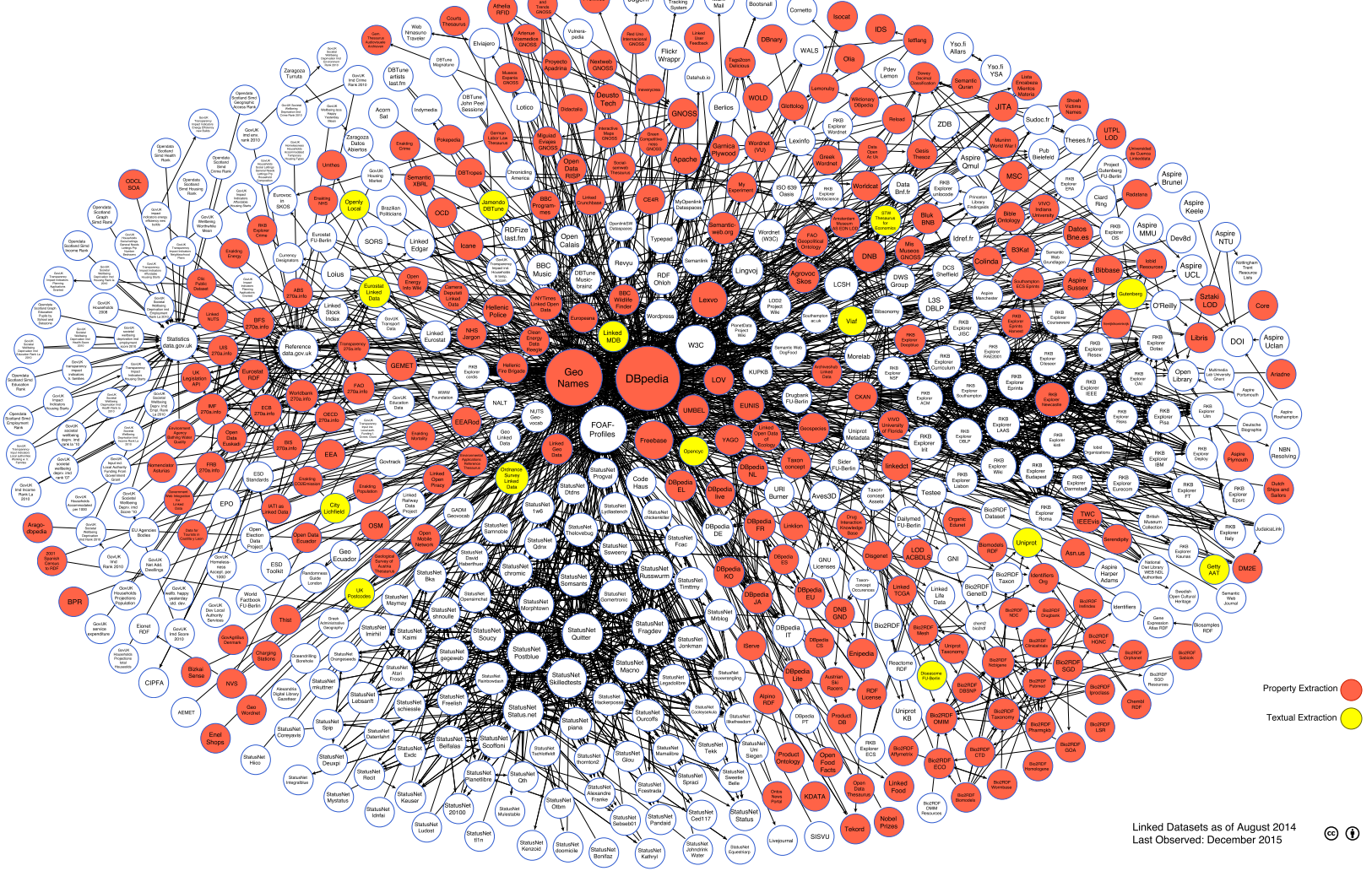

Fig. 2.: Coloring the LOD Cloud Datasets with Licence Availability extracted either via machine readable properties or using regular expressions from textual descriptions.

License Used	Type of License	URL Used	Semantic Resource	Frequency
Creative Commons Attribution License	Requires Attribution	http://www.opendefinition.org/licenses/cc-by	✗	95 (2 Human-readable)
Creative Commons Attribution Share-Alike License	Requires Attribution and Share Alike	http://www.opendefinition.org/licenses/cc-by-sa	✗	48 (1 Human-readable)
Creative Commons Attribution Non-Commercial V2.0 License	Requires Attribution but dataset cannot be used for commercial purposes. This license is a non-conformant license for open data.	http://creativecommons.org/licenses/by-nc/2.0/	✓	32 (1 Human-readable)
Creative Commons CC Zero License	Public domain waiving all rights on the data	http://www.opendefinition.org/licenses/cc-zero	✗	31 (1 Human-readable)
Open Database License	Requires Attribution and Share Alike	http://www.opendefinition.org/licenses/odc-odbl	✗	9
Open Government License for Public Sector Information	Requires Attribution. License can only be used by third parties licensed by the UK Government	http://reference.data.gov.uk/id/open-government-licence	✓	6
Open Data Commons Public Domain Dedication and Licence	Public domain waiving all rights on the data	http://www.opendefinition.org/licenses/odc-pddl	✗	6 (1 Human-readable)
Open Data Commons Attribution License	Requires Attribution	http://www.opendefinition.org/licenses/odc-by	✗	5
GNU Free Documentation License	Share Alike	http://www.opendefinition.org/licenses/gfdl	✗	4
Creative Commons Attribution-NonCommercial-ShareAlike V3.0	Requires Attribution and Share Alike but dataset cannot be used for commercial purposes	–	-	2 (2 Human-readable)
OS Open Data License	Requires Attribution and Share Alike	–	–	2 (2 Human-readable)
Eurostat Policy	Requires Attribution	–	–	1 (1 Human-readable)
Project Gutenberg License	Restricts Commercial Use	–	–	1 (1 Human-readable)
Creative Commons Attribution-NoDerivs License	Does not allow work to be re-used in derivative works	–	-	1 (1 Human-readable)

Table 1

List of licenses used in the metadata, extracted by machine readable properties and from human readable descriptions (figures in brackets).

With regard to data dumps, we looked for media types that are generally associated with the Semantic Web, such as `application/rdf+xml` (which is the minimal requirement for any linked dataset [26, §5.1]) and `text/turtle`. In pursuance of acquiring the largest possible linked dataset coverage, we identified other possible wrongly tagged media types (e.g. `rdf`) and added them to our script[22].

Similarly, for SPARQL endpoints we looked at those distribution resources with an `api/sparql` media type. If the dataset had no SPARQL distribution defined, we probed for availability of a SPARQL endpoint by accessing the path `/sparql` at the fully qualified domain name. Having such a canonical endpoint path is a common practice. In fact, 69.58% of endpoints registered in SPARQLES end with the path `/sparql`. If a SPARQL endpoint is available, we perform a simple *ASK* query to check whether the endpoint responds to queries. A similar SPARQL endpoint retrieval process is described in [42].

VoID descriptions were retrieved from media types containing `void` in their value. Typical media types included `meta/void`. Similar to SPARQL endpoints, if a VoID description is not available as part of the distribution, we look for the metadata in the `/.well-known/void` path of the FQDN, as recommended in [3, §7.2], following the RFC 5785 [40] practices. The VoID metadata is checked for a `void:Dataset`, in order to retrieve possible data dumps (via the `void:dataDump` property) or access the SPARQL endpoint (via the `void:sparqlEndpoint` property).

Following this methodology the acquired dataset collection has a number of known bias factors:

[22] All experiments can be replicated by downloading the scripts available on GitHub: `https://github.com/jerdeb/lodqa`

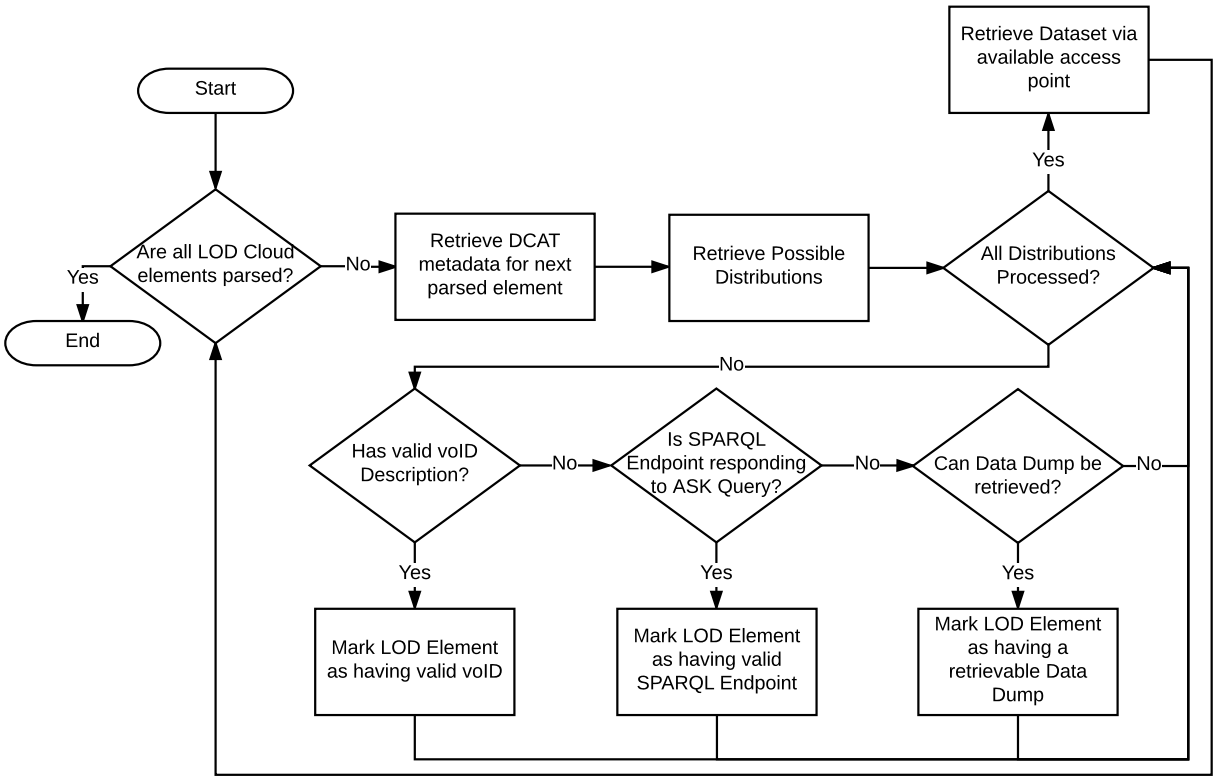

Fig. 3. A high-level flowchart depicting the marking and retrieval process of datasets from the LOD Cloud.

– the harvesting of datasets from the LOD Cloud was performed in December 2015 and the download of the data dumps between December 2015 and February 2016, thus the quality assessment of these datasets reflects the dumps available at the time of download (this does not apply to SPARQL endpoints);
– the downloaded data dumps cover a wide range of tagged media types (also considering incorrect tags), but our assessment is limited to the following: `application/rdf+xml`, `text/turtle`, `application/x-ntriples`, `application/x-nquads`, `text/n3`, `rdf`, `text/rdf+n3`, `rdf/turtle`;
– distributions with `example` in their title were ignored even though they had a correct media type, as we are only interested in having complete datasets (where possible) for our large-scale quality assessment;
– SPARQL endpoints that did not respond to the *ASK* query were considered unavailable and thus not included in the follow-up assessment.

The downloaded data dumps require some data preparation prior to assessment. Each dataset might have multiple distributions, some defining different sub-datasets, others defining the same dataset with different media types (for different serialisations). All dumps in these distributions are downloaded, and then converted to n-quads, merged, sorted, and cleaned by removing duplicate quads. All datasets are identified using their fully qualified domain name. Figure 4 illustrates the a summary of the datasets' access points identified during the acquisition process.

5. Quality Assessment

This study complements the work undertaken in the survey by Zaveri et al. [55] and the work that the survey refers to, by analysing the quality of a collection of LOD Cloud datasets against a number of metrics classified in the mentioned survey. In general, the assessment is done locally, meaning that no dereferencing of external resources is done, unless required by the quality metric. For each data quality metric

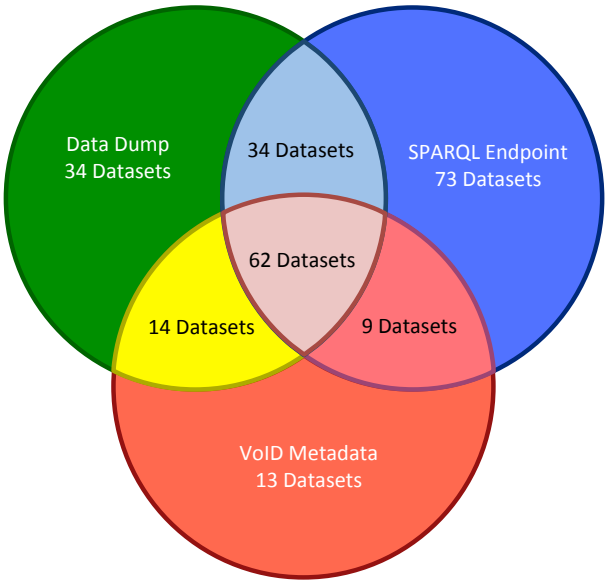

Fig. 4. A Venn Diagram illustrating a summary of the datasets' access points.

we plot a box-and-whiskers chart to summarise metric values and display them on a single graph. Furthermore, with the box-and-whiskers plot, we describe the *sample*'s spread of quality values amongst the LOD Cloud datasets. During the assessment we also collect a sample of the quality problems found during assessment, in order to describe typical problems found in LOD datasets. For each assessed metric we discuss the mean, median, and standard deviation in order to describe the quality values of each dataset collectively in a statistical manner. These descriptive statistical measures present different point of views about the data. The mean and median values are used to find the central values of a set of numeric values, in our case, the quality values of a particular metric over all assessed datasets. The standard deviation value is calculated in order to measure the spread of the data, allowing us to measure the diversity of conformance to a particular metric of the assessed datasets.

5.1. Choice of Data Quality Metrics

In this empirical study we assess the datasets against 27 quality metrics out of 69 metrics described in [55] and related literature, and two additional and novel quality metrics describing provenance information, that were identified from the recent W3C Data on the Web Best Practices guidelines [34]. The majority of the 27 metrics are objective metrics, that is, the metrics' results will not be influenced by the assessor's opinion. For the only subjective metric in this study (re-use of existing terms, cf. Metric IO1), we used the LOD Cloud category classification as the basis of our classification in order to limit any bias. The rest of the metrics identified in [55] were either difficult to implement because of the lack of evidence, for example no algorithm presented or brief description of the metric, or were subjective, meaning that it would be difficult to replicate this study. However, in the future we plan to implement more quality metrics described in [55] in a form of a Linked Data quality assessment as a service.

Since the assessed datasets come from a variety of domains, a certain quality metric might not be relevant, hence some datasets might fare poorly for these particular metrics. Following the overall quality assessment, we provided a service[23] where data consumers can use the generated quality metadata to rank, filter or visualise datasets' quality, based on their choice of metrics.

The choice of generic quality metrics was based solely on the classification in [55]. Nonetheless, there is no study confirming the usefulness of such metrics, and whether or not these quality metrics are informative in a generic assessment such as in this study. Moreover, some metrics may be highly correlated with others and hence provide no additional information. In order to examine this phenomenon, following the quality assessment of the datasets, we statistically analyse the assessment results in order to determine which of the chosen quality metrics are key quality indicators.

5.2. Representational Category

In this section we look at metrics related to the design of data, or in other words: how well the data is represented in terms of common best practices and guidelines. Zaveri et al. [55] categorised a number of metrics in this category within the four dimensions *Representational Conciseness*, *Interoperability*, *Interpretability* and *Versatility*. In Table 2 we list the metrics that are assessed in this category, together with a summary of assessment results showing the mean value (μ), median value (Q_2), and standard deviation (σ_s).

(RC1) Keeping URIs short

Classified in the representational-conciseness dimension, this metric observes the length of URIs. In the Cool URIs document [47], the editors remarked

[23] https://w3id.org/lodquator

	Metric Name	μ	Q_2	σ_s
RC1	Keeping URIs Short	84.07%	97.92%	24.90%
RC2	Minimal Usage of RDF Data Structures	99.44%	100%	2.86%
IO1	Re-use of Existing Terms	34.01%	24.00%	29.10%
IN3	Usage of Undefined Classes and Properties	54.48%	53.33%	32.18%
IN4	Usage of Blank Nodes	96.01%	100%	12.15%
V1	Different Serialisation Formats	0.18 formats	0 formats	0.17 formats
V2	Usage of Multiple Languages	1.72 languages	1 language	2.71 languages

Table 2

List of metrics assessed in the Representational Category together with the assessed mean value (μ), median value (Q_2), and standard deviation (σ_s)

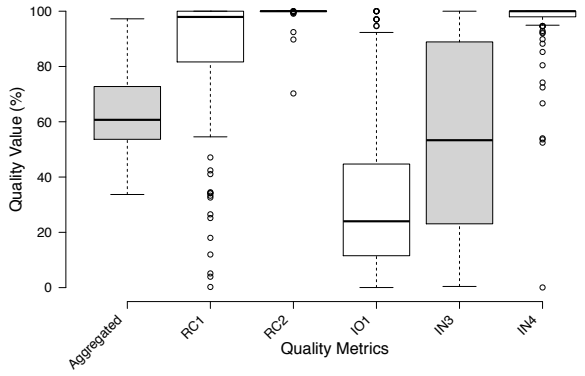

Fig. 5. Representational category box plot excluding Different Serialisation Formats and Usage of Multiple Languages metric are excluded, but included in the aggregated result.

that apart from providing descriptions for people and machines, the best URIs are *simple*, *stable*, and *manageable*.

This metric focuses on the *simplicity* aspect of this definition, where by *simplicity* the editors of the same document mean that having short and mnemonic URIs are easier for humans to remember (e.g. `http://danbri.org/foaf` vs. `https://w3id.org/lodquator/resource/826514e9-e34a-40a2-bc8d-9e6b8bd54770`), whilst serving the purpose of being machine processable. Hogan et al. [30] remarked that short URIs have other benefits such as allowing for smaller sized datasets and indexes.

Metric Computation: The metric computation is based on the W3C best practices for URIs, where the editor suggests that a URI should not be longer than 80 characters [53, §1.1]. Furthermore, URIs with appended parameters are considered as bad, irrelevant of their length. The metric can be quantified as follows:

$$RC1(D) := \frac{size(\bar{u} = \{u \mid (len(u) \leqslant 80) \wedge ('?' \notin u)\})}{size(dlc(D))}$$

where $\bar{u}$ is a set of URIs that have a length (defined by *len*) of 80 or less and are not parameterised (URI contains no '?') and $dlc(D)$ is the set of possible data-level constants in the dataset D being assessed. A data-level constant is defined by [30] as the subject or the object of a quad, when the predicate is not `rdf:type`. Therefore, this metric value measures the ratio of short URIs in a dataset.

Discussion: A box plot with the quality values is illustrated in Figure 5. The box plot for this metric (RC1) suggests that publishers tend to have quite different inclinations on how long the URI identifiers should be. The sample over the LOD Cloud is centred at a median of 97.92% with a standard deviation of 24.90%. A number of outliers (around 13% of the datasets), were detected. These are datasets that scored lower than 54.55%. We also notice that the population is negatively skewed (i.e the median is closer to the third quartile). The median quality value across the assessed datasets is around 84.07%, with 69% of the datasets scoring more than 90%, and 28% of the datasets scoring 100%. From the sample problem report we extracted during the assessment, 14.65% of the URIs were parameterised whilst the rest were URIs longer than 80 characters. This metric has two drawbacks. First, our metric takes into consideration external URIs, however, we acknowledge that the length of such external URIs cannot be influenced by the datasets' publisher. A solution for this is that the metric looks only at locally minted URIs. The second drawback of this metric is the lack of discriminative power, since URIs with 80 characters are adequate, whilst longer ones are deemed to be non-conformant with the guidelines set in [53, §1.1]. In order to avoid this discriminatory power problem, Hogan et al. [30, §5.1 – Issue IV] calculate the metric value based on the mean length of the URIs in a dataset, rather than relying on the 80 character limit, promoting those datasets that have short URIs.

The editor of the Common HTTP Implementation guidelines [53] states that 80 character limit is not a technical limitation but rather a practical goal one should pursue. There might be various reasons for pub-

lishers to use longer URIs. For example, URIs can comprise some structure, such as a directory scheme. Moreover, in a recent article by Szász et al. [52], the authors discussed the idea of self-unfolding semantic URIs, where such URIs follow a specific pattern and template that would result into a set of RDF triples "on-demand". These self-unfolding URIs might be longer than the 80 character limit, but the benefits of such URIs include frequent complex insertion in RDF triple stores, for example in sensor network streams. Therefore, we deem that this 80 character practical limit suggested in [53, §1.1] has its own limitations and could affect the quality value of datasets where publishers cannot do without lengthy URIs, for example having resources with a multi-level domain name such as in the case of university URIs.

(RC2) Minimal Usage of RDF Data Structures

The usage of RDF data structure features, more specifically reification, containers, and collections, is discouraged due to their syntactic/semantic complexity. Despite the fact that a number of efforts were made in order to facilitate the use of such data structures (e.g. the introduction of property paths[24] in SPARQL 1.1 allows the retrieval of all members in an `rdf:List` with one graph pattern: `{?s rdf:rest*/rdf:first ?o .}` – this was not possible in SPARQL 1.0), these are still more complicated to handle. In [26, §2.4.1.2], the author discourages the use of RDF reification since it is "rather cumbersome to query with the SPARQL query language". Furthermore, the authors argue that if set ordering is not required, collections and containers are best avoided. In RDF, these data structures are typically described using blank nodes, which is another discouraged practice (cf. Metric IN4). In [30, §5.3 – Issue VIII], Hogan et al. explain the various issues, such as scalability and lack of semantics, that these features bring about.

Metric Computation: This metric detects the use of standard RDF data structure features. More specifically, this metric checks quads as suggested in [30, §5.3 – Issue VIII]:

- if the predicate is `rdf:type` and the object is one of `rdf:Statement`, `rdf:Alt`, `rdf:Bag`, `rdf:Seq`, `rdf:Container`, or `rdf:List`;
- if the predicate is one of `rdf:subject`, `rdf:predicate`, `rdf:object`, `rdfs:member`, `rdf:first`, `rdf:rest`, or `rdf:_'[0-9]+'`.

The value of this metric can be quantified as follows:

$$RC2(D) := 1.0 - \frac{size(RCC(D))}{size(quads(D))}$$

where $RCC(D)$ is the set of quads from dataset D that satisfy the above conditions, and $quads(D)$ is the set of all quads in dataset D.

Discussion: Similar to the findings of Hogan et al. [30], most publishers do not use RDF data structures. In our assessment 87.2% of the publishers use none, compared to the 78.7% reported by Hogan et al. This is reflected in the short box plot illustration for this metric (RC2 – Figure 5), with the interquartile ranges and whiskers all being close to 100%. The mean quality value of this metric is 99.44% and the calculated standard deviation is $\sigma_s = 2.86\%$ (median value is 100). The σ_s value confirms our findings that most publishers try to minimise the use of such undesired RDF features, with 97% of the datasets ranking within the standard deviation (i.e. having a quality value of at least 96.369%). Similar to Metric RC1, a relatively small number of outliers (around 12% of the datasets) were detected.

Nonetheless, this metric punishes datasets where ordering is essential. For example, the dataset with the lowest quality value (70.25%) for this metric is `http://bibsonomy.org`, which contains resources on scientific papers amongst others. Upon further inspection of this dataset, we found that the dataset's publisher used `rdf:Seq` and `rdf:Bag` to represent the editors and authors of publications in their order, which is essential in this domain. In general, the RDF collections were the most common issue (95.23%), followed by RDF containers (3.09%) and RDF reification (1.67%).

(IO1) Re-use of Existing Terms

Vocabulary re-use is widely advocated. For instance, Bizer and Heath [13] argue that re-using terms from known vocabularies makes it easier for applications to process Linked Data, thus increasing interoperability between agents. Schemas for different domains are meanwhile publicly available; also via registries such as the *Linked Open Vocabulary* (LOV)

[24] `http://www.w3.org/TR/sparql11-query/#propertypaths`

portal[25]. Together with W3C recommendation vocabularies such as *SKOS*, schemas such as *FOAF*, *Dublin Core*, and *SIOC*, amongst others, have become de-facto standards with more than 15% of the LOD datasets using at least one of these vocabularies [49]. Furthermore, the W3C is striving to create standardised cross-domain vocabularies, such as *DCAT* for dataset metadata, and *PROV-O* for provenance, amongst others. Zaveri et al. [55] classify this metric under the interoperability dimension, and focus on the overlap between the dataset in question and its overlap with recommended vocabularies [30, §5.3 – Issue IX].

Metric Computation: This metric assesses if a dataset re-uses relevant terms in a particular domain. More specifically, each dataset is tagged with the domain as classified by the LOD Cloud, for example, the Lexvo dataset is tagged as *linguistics*. The LOV API is then queried with 'linguistics' and all schemas given by the service are used. In particular, this metric checks if a property or a class (in case the predicate is `rdf:type`) used in a triple refers to an existing term in another vocabulary. Since the metric depends on the domain of the dataset, for this experiment all LOD Cloud datasets were tagged according to their identified domain in the cloud itself (each dataset in the LOD cloud is tagged with one domain, for example, DBTropes[26] is tagged with the label *media*). During the initialisation of the metric, the LOV API[27] is invoked to obtain the vocabularies available with the respective tag. Furthermore, based on the usage study conducted in [49], we included the following vocabularies by default for all datasets: RDF, RDFS, FOAF, DCMI Terms, OWL, GEO, SIOC, SKOS, VoID, DCAT.

We identify overlapping classes and properties in the same manner as defined in [30, §5.3 – Issue IX], with the set of known vocabularies generated from LOV. The metric counts the number of external classes and properties (from external vocabularies identified by LOV) for a particular domain:

$$IO1(D) := \frac{size(\overline{cl_{exs}}) + size(\overline{pr_{exs}})}{size(class(D)) + size(prop(D))}$$

$$\overline{cl_{exs}} := \overline{v_c} \cap class(D)$$

$$\overline{pr_{exs}} := \overline{v_p} \cap prop(D)$$

where $class(D)$ is the set of classes in the assessed dataset D, appearing in the object position with predicate `rdf:type` excluding blank nodes. The set $prop(D)$ defines the set of terms appearing at the predicate position of the quads in the dataset D, excluding `rdf:type`. $\overline{v_c}$ and $\overline{v_p}$ are the sets of **all** classes and properties respectively, gathered from the identified external vocabularies for the particular dataset.

Discussion: The box plot for this metric (IO1 – Figure 5) is comparatively (against the other metrics in this category) long and positively skewed, suggesting that most values are small with some larger values. This also suggests that there is a lack of conformity on the principle of re-use; only few publishers rely actively on the re-using vocabularies ($\approx$ 10% of datasets have a quality value of > 90%), with 8.8% of the datasets being outliers in this case as they have a quality value larger than 92.32% (i.e. the upper whisker value). The sample is centred at a median of 24.00% with a standard deviation of 29.10%, and a mean value of 34.01%, indicating low overall re-use. One possibility is the fact that publishers (such as DBpedia) use local terms and properties with few external properties (e.g. `rdfs:label`). Our values are comparable to those in Hogan et al. [30, §5.3 – Issue IX], where the authors also suggest that the amount of re-used terms and properties in their sample is widely distributed.

With the pre-defined tags associated to each dataset, we ensured that each dataset is assessed solely based on its domain, relying on the LOV service to provide us with relevant public vocabularies. This means that our assessment might have either missed some vocabularies, or expected datasets to use terms from a vocabulary which has been overlooked by the publishers. Our current implementation of this metric does not take into account the best practice of introducing user-defined terms by linking them *existing* terms using predicates such as `owl:sameAs`, `owl:equivalentClass`, or `owl:equivalentProperty`; such terms are not currently recognised as valid re-used existing terms.

(IN3) Usage of Undefined Classes and Properties

The invalid usage of undefined classes and properties metric is classified under the interpretability dimension [55], targetting the technical representation of the data itself. Using classes and properties without a formal definition is undesirable, as agents would not be able to understand how the data should be interpreted, for example, during reasoning. Errors lead-

[25] `http://lov.okfn.org/`

[26] `http://dbtropes.org`

[27] `http://lov.okfn.org/dataset/lov/api/v2/vocabulary/search`

ing to such invalid usage include capitalisation errors (e.g. `foaf:person` vs. `foaf:Person`), syntactic errors (e.g. `foaf:img` vs. `foaf:image`), and schema dereferencability issues.

Metric Computation: This metric measures the number of undefined classes and properties in the assessed dataset:

$$IN3(D) := 1.0 - \frac{size(\overline{cl_{undef}}) + size(\overline{pr_{undef}})}{size(class(D)) + size(prop(D))}$$

$$\overline{cl_{undef}} := \{c \in class(D) \mid \{(c, \texttt{rdf:type}, C) \mid C \in \mathcal{C}\} \cap V(ns(c)) = \emptyset\}$$

$$\overline{pr_{undef}} := \{p \in prop(D) \mid \{(p, \texttt{rdf:type}, P) \mid P \in \mathcal{P}\} \cap V(ns(p)) = \emptyset\}$$

An undefined class is a term c that is used as a class in the dataset, as defined above for Metric IO1, but that is not defined to be an instance of a class type ($\mathcal{C} := \{\texttt{rdfs:Class}, \texttt{owl:Class}\}$) in the vocabulary $V(ns(c))$ retrieved by dereferencing its namespace URI.[28] Similarly, an undefined property p is not defined as being of type `rdf:Property`, `owl:ObjectProperty`, `owl:DatatypeProperty`, `owl:AnnotationProperty`, or `owl:OntologyProperty` in the vocabulary $V(ns(p))$. If a term is not dereferenceable, it is considered undefined as well.

Discussion: The box plot for this quality metric (IN3 – Figure 5) covers a range of 99.58%. This suggests that data publishers are using a wide range of defined and undefined classes and properties. Furthermore, the quality value is centred at a median of 53.33% with a standard deviation value of 32.18%, whilst the mean quality value is 54.48%.

A higher value means that publishers were using fewer undefined terms in their dataset. From our assessment, 30.80% of properties used were undefined. Some of the undefined terms were possibly previously defined. For example, for the rkbexplorer datasets, the publishers use terms from the `aktors.org` namespace, which now resolves to a personal blog. As another common pattern of undefined terms, we noticed the use of the wrong namespace for a term that did exist in a similar namespace, for example, `rdfs:Property` as opposed to `rdf:Property`. Other datasets had schemas that were unavailable during the assessment, thus resulting in undefined terms.

(IN4) Usage of Blank Nodes

Blank nodes are undesirable in Linked Data because they cannot be externally referenced, which conflicts with the two Linked Data best practices interlinking and re-using. In simple terms, the scope of blank nodes is *"limited to the document in which they appear"* [26]. Moreover, the existence of blank nodes can cause a number of problems during Linked Data consumption and when performing certain tasks, such as deciding whether two RDF graphs are isomorphic.

Metric Computation: This metric assesses the usage of blank nodes within the subjects and objects. The metric value is assessed as suggested in [30, §5.1 – Issue I]:

$$IN4(D) := \frac{size(dlc(D) \backslash bn(D))}{size(dlc(D))}$$

where $dlc(D)$ is the set of data-level constants in dataset D and $bn(D)$ is the set of blank nodes in D.

Discussion: The box plot (IN4) illustrated in Figure 5, is relatively short, suggesting that most data publishers agree to avoid blank nodes. The median value is 100%, the standard deviation is 12.15%, whilst the mean quality metric value of 96.01% confirms the generally high conformance with this metric.

Whilst the majority of data publishers use blank nodes sparsely or not at all (around 85% of the datasets score higher than 94.93%, which is the lower whisker limit), there is a number of datasets marked as outliers consequently stretching the standard deviation. In particular, the `prefix.cc` dataset uses blank nodes in almost every triple. This dataset affected the standard deviation significantly, which otherwise would be considerably lower than in [30, §5.1 – Issue I]. One should note that the corpus in [30] contained FOAF profiles, which traditionally, according to Hogan et al. [30], may contain many blank nodes. In certain situations, the usage of blank nodes is due to RDF data structure features and OWL axioms, as these structures and axioms use blank nodes as the encoding, though in general avoiding them means that resources in a dataset are more likely to be re-used for linking.

(V1) Different Serialisation Formats

An RDF data model can be serialised using a variety of formats, including RDF/XML, RDFa, Turtle, N-Triples, N-Quads, and JSON-LD. For example, Web applications prefer the JSON-LD format, rather

[28] In cases where slash URIs are used, dereferencing the namespace does not necessarily resolve the schema; therefore the term is used to resolve the term *itself*.

than having to use some parser, as the JavaScript environment handles JSON data internally. The different characteristics of each serialisation brings about different pros and cons, as described in [26, §2.4.2]. The rationale of this metric is to assess whether various consumption methods are supported. Ensuring that a dataset is available in multiple serialisation formats facilitates its use. The metric is classified under the versatility dimension [55].

Metric Computation: This metric checks whether a dataset has multiple serialisation formats defined in its metadata, by verifying that multiple quads having `void:feature` as a predicate exist in the assessed dataset. The `void:feature` predicate is used to express the technical features of a dataset, such as the serialisation formats the dataset is available in [3, §2.6]. Data publishers can serialise their data in up to 23 different formats[29]. The metric can be quantified as follows:

$$V1(D) := size(features(D))$$

where $features(D)$ is the set of valid dataset serialisation features identified by the object in a triple **subject** $\times$ **void:feature** $\times$ **object**.

Discussion: In most cases, the publishers did not define any serialisation format in the metadata of their datasets. In total, only nine datasets had a serialisation format following our guideline. The standard deviation is 0.71 formats whilst the mean value is 0.18 formats per dataset. In total we had one dataset with four different serialisation formats, five datasets (four of which had a pay-level domain of `psi.enakting.org`) had three different serialisation formats listed, two datasets had two formats listed, whilst one dataset had only one serialisation format listed. The most common formats where N-Triples, RDFa, RDF/XML, Turtle, N-Quads, and SPARQL query results in XML format.

A dataset serialised in different formats, widens possible uses in different scenarios. In order to encourage multiple format serialisation, tools such as *Raptor*[30] or *Serd*[31] provide command line functions that transform (bulk) data into various serialisations. One drawback is that using different serialisations (for data dumps) takes up more storage resources. Regarding the generation of VoID metadata, generators and approaches such as [15], help publishers to create VoID descriptions.

(V2) Usage of Multiple Languages

Catering for multiple languages ensures that the dataset reaches a wider global audience. For example, a dataset with literals having only a Maltese language tag is not suitable for Chinese speaking users. On the other hand, if the dataset has literals in both Maltese and Chinese, then the dataset is likely to be re-used more. A plain (textual) literal string can be combined with a language tag (e.g. `@mt`) Furthermore, the Data on the Web Best Practices document suggests that locale parameters should be provided in metadata:

> *"making the language explicit allows users to determine how readily they can work with the data and may enable automated translation services."* – [34]

Language tags also allow agents to express linguistic or text-based information better, for example, providing better localisation. The usage of multiple languages metric is also classified by Zaveri et al. under the versatility dimension [55].

Metric Computation: This metric checks the number of languages a dataset supports. Specifically, the metric checks whether the data (in this case string literals) is *evenly* available in different languages. For example if resource R_1 has two labels with different language tags and R_2, R_3, R_4, R_5 have only one label with one language tag each, then on average the dataset has resources defined in 1.2 languages, which is rounded down to one language.

$$\begin{aligned} \bar{l} :=& \{s \mid \exists o. \\ & o = \texttt{"lexical form"}@lang \\ & \wedge (s,p,o) \in D\} \\ triples_{lang} :=& \{t \mid t.o \in literal(D) \wedge hasLangTag(t.o)\} \\ V2(D) :=& round\left(\frac{size(triples_{lang})}{size(\bar{l})}\right) \end{aligned}$$

where $triples_{lang}$ is the set of all triples in the dataset D whose object is a literal with some language tag, and $\bar{l}$ is the set of unique resources having at least one property with a literal object that has a language tag.

Discussion: In most cases, publishers describe their textual literals using only one language ($\approx 83\%$ of the datasets). One possible reason is that publishers target a particular audience, or do not have the resources

[29] `http://www.w3.org/ns/formats/`
[30] `http://librdf.org/raptor/`
[31] `https://drobilla.net/software/serd/`

to create multilingual datasets. The mean value for this metric is 1.72 languages, whilst the standard deviation is 2.71 languages (median: 1 language). Overall, this metric shows positive skewness as the standard deviation is almost twice of the mean and median values. In total, we had four datasets having fifteen different languages (these had a pay-level domain of `psi.enakting.org`), languages including Arabic, Greek, English, Italian and Spanish. Furthermore, resources where described in thirteen, eight and four different languages, in three different datasets. Finally, eleven datasets had at least resources described in two different languages for human consumption.

Aggregated Results

Figure 5 shows a box plot illustration of the aggregated quality value compared with the category's metrics (V1 and V2 are missing as the quality value are integers used to count things, whilst the rest are float values between 0.0 and 1.0). The overall aggregated box plot shows a population which is slightly skewed to the left, with a median of 60.70% conformance. This shows that there is more variety amongst higher quality values amongst the sample. Nevertheless, the standard deviation is 14.50%, which indicates a moderate distribution, whilst the mean score is 63.60%.

5.3. Contextual Category

According to Zaveri et al. [55], the contextual category groups those dimensions and metrics that are highly dependent on the task at hand. The dimensions classified in this category deal with (i) *relevancy* of a dataset vis-à-vis the task at hand, (ii) degree of data correctness and credibility, i.e. the *trustworthiness* of the dataset, (iii) *understandability* of the data in terms of human comprehensibility and ambiguity, and (iv) *timeliness* of data. In this article, we introduce a new dimension, *provenance*, which for quality purposes we define as *the provision of information regarding the origin of the dataset and of the resources within the dataset itself*. The provenance metrics we propose are similar to those classified under the *trustworthiness* dimension. Furthermore, in this category, we only tackle three metrics related to *understandability*. In Table 3 we list the metrics that are assessed in this category, together with a summary of assessment results showing the mean value (μ), median value (Q_2), and standard deviation (σ_s).

	Metric Name	μ	Q_2	σ_s
P1	Provision of Basic Provenance Information	12.78%	0%	32.89%
P2	Traceability of the Data	2.17%	0%	10.06%
U1	Human Readable Labelling and Comments	43.76%	33.33%	40.93%
U3	Presence of URI Regular Expression	7.75%	0%	32.18%
U5	Indication of Used Vocabularies	2.71%	0%	10.62%

Table 3

List of metrics assessed in the Contextual Category together with the assessed mean value (μ), median value (Q_2), and standard deviation (σ_s).

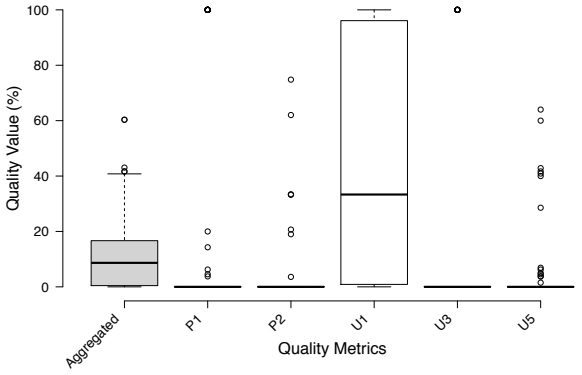

Fig. 6. Contextual category box plot. Outliers are represented by dots.

(P1) Provision of Basic Provenance Information

Data provenance is considered as one of the main assets in a Linked Data.

> *"Data provenance becomes particularly important when data is shared between collaborators who might not have direct contact with one another either due to proximity or because the published data outlives the lifespan of the data provider projects or organisations."* – [34, §9.4]

The importance of data provenance lies in the fact that consumers need to understand where the data comes from and by whom it was produced. In this way, consumers can identify whether for example they can trust the integrity and credibility of the dataset.

Metric Computation: At the very least, a dataset should have a `dc:creator` or `dc:publisher` within their VoID or DCAT metadata. We focus on searching for triples with the predicates `dc:creator`

or `dc:publisher` in every resource pertaining to `void:Dataset` or `dcat:Dataset`. The metric can be formally defined as follows:

$$P1(D) := \frac{\sum_{d \in \overline{ds}(D)} basic(d)}{size(\overline{ds}(D))}$$

where $\overline{ds}(D)$ is the set of resources having a type of `void:Dataset` or `dcat:Dataset` in the assessed dataset D, whilst $basic(d)$ is a function that returns '1' if $d \in \overline{ds}$ has a triple corresponding to $\boldsymbol{d} \times$ (**dc:creator** $\|$ **dc:publisher**) $\times$ **object**.

Discussion: A box plot with the quality values for the contextual dimension metric is given in Figure 6. The box plot for this metric (P1) is, qualitatively speaking, positively skewed, suggesting that most of the sampled datasets contain no basic provenance information in their VoID or DCAT metadata (when available). Nonetheless, this metric has a number of outliers, amounting to around 16.27% of the datasets. From this 16.27%, 71% of the datasets have a quality value of 100%. The σ_s value stands around 32.89% (median 0%), whilst the mean is 12.78%. One possibility of this low score was due to the fact that the metrics did not consider `dcterms:creator` and its equivalent `foaf:maker`. Considering these two predicates in this metric might give higher overall values.

Publishers might add basic provenance triples directly in a dataset rather than in the metadata, which is a drawback in terms of *"understand(ing) the meaning of data"* [34], as the provenance will be unknown to an automated agent looking for this information within the metadata before consuming the actual data. For example, `europeana.eu` attaches a `dc:creator` to every resource (which is understandable considering that it is a dataset about cultural heritage) rather to specific metadata resources. Hence, we encourage publishers to use dataset profiles specifications such as DCAT-AP[32] or editors such as the VoID Editor[33].

(P2) Traceability of the Data

In the Data on the Web Best Practices recommendation, the editors note that

> *"consumers need to know the origin or history of the published data, [...], data published should include or link to provenance information"* – [34, §9.4]

Different publishers might contribute to the same dataset, by publishing within the same namespace. Therefore, it is important that consumers can track the origin of each piece of data/resource in a dataset. This provenance metadata can be described using the PROV-O ontology [33]. PROV-O allows the identification of agents, entities and activities. An agent represents the owner, or the responsible person for an activity or entity. An entity is "a physical, digital, conceptual, or other kind of thing with some fixed aspects [and] may be real or imaginary" [33], for example weather information from Delhi. An activity "occurs over a period of time and acts upon or with entities" [33].

Metric Computation: This metric checks whether each resource has provenance information related to the origin of data. With regard to the quality metric survey in [55], this metric can be related to the "trustworthiness of statements". More specifically, this metric checks for entities with the following characteristics:

- Identification of an *agent* of an *entity* (quads having a predicate `prov:wasAttributedTo`);
- Identification of *activities* in an *entity* (quads having a predicate `prov:wasGeneratedBy`);
  1. Identification of a *data source* in an *activity* (quads having a predicate `prov:used`);
  2. Identification of an *agent* in an *activity* (quads having a predicate `prov:wasAssociatedWith` and/or `prov:actedOnBehalfOf`);

In order to avoid bias, an agent and an activity in an entity are both given a weight of 0.5. Similarly, data source and agent (in an activity) are also given a weight of 0.5. Then, the metric can be computed as follows:

$$P2(D) := \frac{\sum_{e \in prov(D)} val(e)}{size(prov(D))}$$

where $prov(D)$ is the set of entities as described above, whilst $val(e)$ is the quantified weighted value of the entity.

Discussion: Similar to Metric P1, this metric (P2 - Figure 6) is also very positively skewed. Unlike Metric P1, the granularity level of the metadata in this case

[32] `https://joinup.ec.europa.eu/asset/dcat_application_profile/description`

[33] `http://voideditor.cs.man.ac.uk`. List of other VoID editors and generators: `http://semanticweb.org/wiki/VoID.html#Generators_.26_Editors`

can even reach a triple level. This means that the size of the overall dataset can grow very large, therefore publishers might not be willing to trade-off size for better metadata coverage. In fact, we noticed that there is only one publisher (`270a.info` datasets) with metadata enabling users to identify the origin of data. The overall median value is 0%. The standard deviation stands around 10.06%, whilst the mean is 2.17%.

The practice of tracking the origin of data is often ignored by data publishers, possibly for a myriad for reasons, such as the inflating the size of the dataset, or modelling issues. We suggest that publishers add provenance information on the activities undertaken when creating resources in their dataset, and possibly separating this metadata from the data itself by using named graphs.

(U1) Human Readable Labelling and Comments

Data on the Web is meant to be exposed to both humans and machines. Therefore, a human information consumer should be able to comprehend and understand a Linked Data resource. Apart from human understandability, labels and comments can be used in various applications, such as keyword-based and natural-language based search [20]. A Linked Data application is dependent on labels and comments provided with each resource, as the application itself is not yet intelligent enough to try to map a resource to its real-world description. Labels can possibly be extracted from a human readable URI, e.g. extracting the fragment 'Dublin' from `http://dbpedia.org/resource/Dublin`.

Heath and Bizer suggest that predicates such as `rdfs:label`, `foaf:name`, `skos:prefLabel`, or `dcterms:title` should be used to label resources as they are widely supported by Linked Data applications, whilst `dcterms:description` and `rdfs:comment` should be used for a textual description of a resource [26]. Nevertheless, there are a number of vocabularies having terms to describe human readable labels and comments[34]. The authors in [20] study the usage of labels in the Web of Data[35], and reported the occurrence of the various predicates used for resource labelling. [55] classify this metric under the *understandability* dimension.

Metric Computation: The aim of this metric is to calculate a dataset's completeness in terms of human-readable labels and descriptions. The metric measures the percentage of local entities that have a label or a description. More specifically, each resource should have one (or more) of the following predicates, extracted from the top 50 vocabularies used in the LOD Cloud [49]:

- `rdfs:label`;
- `rdfs:comment`;
- `dcterms:title`;
- `dcterms:description`;
- `dcterms:alternative`;
- `skos:altLabel`;
- `skos:prefLabel`;
- `skos:note`;
- `powder-s:text`;
- `skosxl:altLabel`;
- `skosxl:hiddenLabel`;
- `skosxl:prefLabel`;
- `skosxl:literalForm`;
- `schema:name`;
- `schema:description`;
- `schema:alternateName`;
- `foaf:name`.

A Linked Data resource is a *thing of interest*, or in a more practical sense, a set of triples that have the same subject URI. The metric can be computed as follows:

$$U1(D) := \frac{size(\{s \in ent \mid \exists p \in desc.(s,p,o) \in D\})}{size(ent)}$$

$$ent(D) := \{s \mid (s,p,o) \in D\}$$

where *ent* is the set of distinct subject URIs in the assessed dataset D, p is the predicate of a triple, and *desc* is the set of predefined predicates that define a label or description.

Discussion: The box plot for this quality metric (U1 – Figure 6) is relatively tall, showing a uniform distribution. This suggests that data publishers follow varying practices with regard to the conformance of human-readable labels and comments. Furthermore, one must keep in mind that similar to the metric related to the usage of multiple languages (Metric V2), one might not need to label or textually describe a resource. For example, in data cube[36] datasets (such as

[34] A simple search on LOV resulted into 346 terms for labels (12 of which tagged as W3C recommendations) and 150 terms for comments (1 being tagged as a W3C recommendation).

[35] They used the corpus of the 2010 Billion Triples Challenge (`http://challenge.semanticweb.org/`)

[36] `http://www.w3.org/TR/vocab-data-cube/`

the assessed `270a.info` datasets), publishers might not add human readable labels and descriptions to each generated observation since the aim of these observations is to report the measured values of some particular dimension. Moreover, publishers might also use other non-standard schemas to describe resources in a human readable fashion.

The quality value is centred on 33.33% with a standard deviation of 40.93%, whilst the mean quality value is 43.76%. This quality metric displays the highest dispersion from all contextual metrics as evidenced by a wide bar. Moreover, around 29.29% of the assessed datasets have a completeness value of more than 90%, whilst in total around 43% of the datasets have a value of more than 50%. This metric is similar to the one presented in Hogan et al. [30, §5.3 – Issue XI], however, in this assessment we analysed a larger variation than in that of the study in 2012. We can also draw parallels between our assessment results and the results presented in [20], as both assessments show that the community needs to work harder to ensure the completeness of human readable labels and descriptions in Linked (Open) Datasets.

(U3) Presence of URI Regular Expression

One of the main purposes of the Web of Data is to be queried and explored. Structural metadata enables consumers to understand the underlying structure of a dataset. Having a regular expression defining the URI structure of a dataset enables agents to interpret resources better, for example, extracting fragments of a resource URI such as its local name, or querying a dataset to retrieve local resources according to the specified URI structure. The presence of URI regular expression metric is classified under the *understandability* dimension [55].

Metric Computation: This metric checks for the identification of a URI regular expression in the dataset's metadata, and can be quantified as follows:

$$U3(D) := \begin{cases} 1.0 & \text{if has pattern} \\ 0.0 & \text{otherwise} \end{cases}$$

where by *has pattern*, the metric is looking for a triple **subject** $\times$ **void:uriRegexPattern** $\times$ **object** in the assessed dataset.

Discussion: This metric reports 100% if the assessed dataset has a URI regular expression pattern defined. Our assessment showed that only ten of the datasets had such an expression, giving a total mean value of 7.75%, and a standard deviation of 26.84%. The box plot for the metric U3 in Figure 6, illustrates this positively skewed quality indicator.

(U5) Indication of Used Vocabularies

Vocabularies play an important role in the structure of a dataset, since one or more of these vocabularies describe the dataset's resources. Similar to Metric U3, indicating the vocabularies used is part of the structural metadata of a dataset. Knowing the vocabularies used in a dataset, a human consumer can query the data. This metric is also classified under the *understandability* dimension [55].

Metric Computation: This metric checks whether vocabularies used in the datasets, either in the predicate position or in the object position if the predicate is `rdf:type`, are indicated in the dataset's metadata, specifically using the `void:vocabulary` predicate. The RDF, RDFS and OWL vocabularies are not taken into account in this metric. This metric value can be computed as follows:

$$U5(D) := \frac{size(vocabularies(D))}{size(\{ns(t) | t \in class(D) \cup prop(D)\})}$$

where $vocabularies(D)$ is the set of vocabularies, identified by the object in a triple **subject** $\times$ **void:vocabulary** $\times$ **object**. The metric's value represents the ratio of the defined vocabularies in the dataset's VoID description vs. the actual vocabularies used in a dataset, identified by the unique namespaces of the terms, i.e. classes ($class(D)$) and properties ($prop(D)$).

Discussion: Similar to most of the contextual metrics, the box plot for this metric (U5) is, also positively skewed, suggesting that most of the population datasets have no indication of the vocabularies used. Despite having a median value is 0%, this metric has a number of outliers, amounting to around 11% of the population dataset. These outliers pushed the standard deviation to 10.62%, whilst the mean is 2.71%. This metric might also have values that are > 100%, especially when a lot of vocabularies are declared, but not all of them are used in the dataset's triples. Nonetheless, we have not encountered this during our assessment.

From our assessment, around 2,800 different (not unique) vocabularies were used throughout the assessed dataset, whilst only 128 (around 4%) vocabularies were identified by the `void:vocabulary` predicate. Moreover, only 63 of those 128 defined vocabularies (around 63%, and around 2% of the total number

of vocabularies used) were actually used in the dataset. This means that around 37% of the defined vocabularies were not used in their respective datasets. Using VoID generators as part of their publishing methods (mentioned in Metric P1), such issues can be easily rectified by the publishers.

Aggregated Results

Figure 6 shows a box plot illustration of the aggregated quality value compared with the category's metrics. The overall aggregated box plot shows a population with a median of 8.66%. The standard deviation is 13.84%, whilst the mean score is 13.04%. Five datasets from the whole population are "positive outliers" (since their overall quality value in this category is superior to rest of the population). These quality scores shed light on the real problems related to the contextual category. More worrying is the fact that provenance information is not given the same importance as other quality metrics. Data consumers might look at such provenance information to make informed decisions on whether to trust a particular dataset or data publisher prior to using a dataset. Lacking such information might make it hard for data consumers to re-use and adopt some dataset.

5.4. Intrinsic Category

Defined as *"independent of the user's context"* [55], the intrinsic category quality indicators are related to *correctness* and *coherence* of the data. Zaveri et al. [55] classified metrics according to the following dimensions:

1. *syntactic validity* – the conformance of an RDF graph with the RDF standard;
2. *semantic accuracy* – the correctness degree of the represented values with regard to the real world;
3. *consistency* – the level of coherence in a dataset with respect to the knowledge it represents and inference mechanisms;
4. *conciseness* – the degree of redundancy in a dataset; and
5. *completeness* – the extent to which data is complete with respect to the real world.

In this section we assess a metric related to the *conciseness* dimension, seven metrics related to the *consistency* dimension, and one metric from the *syntactic validity* dimension. No metrics were assessed for the other two dimensions mentioned in [55], as they would have required a different experiment setup. For example, for the *completeness* dimensions, we would require to assess the datasets according to their domain. Furthermore, since most metrics in this category rely on external vocabularies, datasets should not be penalised for external vocabularies that are offline. In Table 4 we list the metrics that are assessed in this category, together with a summary of assessment results showing the mean value (μ), median value (Q_2), and standard deviation (σ_s).

	Metric Name	μ	Q_2	σ_s
CN2	Extensional Conciseness	92.04%	99.34%	13.22%
CS1	Entities as Members of Disjoint Classes	100%	100%	0%
CS2	Misplaced Classes or Properties	99.99%	100%	0.01%
CS3	Misused OWL Datatype or Object Properties	98.88%	100%	5.17%
CS4	Usage of Deprecated Classes or Properties	99.97%	100%	0.23%
CS5	Valid Usage of the Inverse Functional Property	96.98%	100%	12.29%
CS6	Ontology Hijacking	93.64%	100%	19.99%
CS9	Usage of Incorrect Domain or Range Datatypes	60.11%	57.14%	13.43%
SV3	Compatible Datatype	96.80%	100%	14.16%

Table 4

List of metrics assessed in the Intrinsic Category together with the assessed mean value (μ), median value (Q_2), and standard deviation (σ_s).

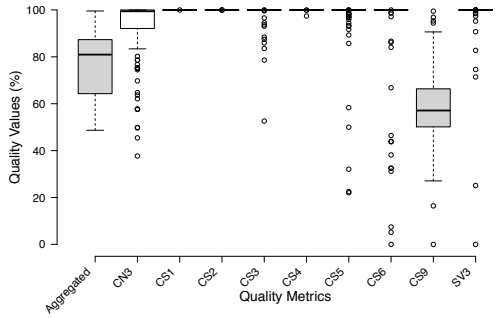

Fig. 7. Intrinsic category box plot. Outliers are represented by dots.

(CN2) Extensional Conciseness

In [14], Bleiholder and Naumann define the conciseness metric as *"measure(ing) the uniqueness of object representations"*. Undoubtedly, from a database point of view, data redundancy causes a dataset to be

large. This issue might not be that significant anymore because of large storage devices, or distributed storage. Moreover, data redundancy can be challenging in terms of data curation. For example, a data curator has to ensure that all "replicated" resources are updated accordingly. Nonetheless, data redundancy is not always a bad thing, for example, such redundancies can lead to improvements in query rewriting in Ontology-based Data Access, although it should be avoided if the publisher does not understand how to maximise its utility [54].

At the Linked Data level, a linked dataset is concise if there are no redundant instances [37]. By redundancy, Mendes et al. [37] mean that there are two local instances with different identifiers but with the same set of properties and corresponding data values. The extensional conciseness metric is classified under the conciseness dimension in [55].

Metric Computation: The extensional conciseness metric checks for redundant resources in the assessed dataset, and thus measures the number of unique instances found in the dataset. We follow the definition of Mendes et al. [37] in order to identify where two instances with a different URI are the same or not. In [19, §5.2], we showed that a naïve implementation of this metric leads to great time complexity, therefore we suggest the use of Bloom Filters [10] as an approximation technique. Using the Bloom filter for identifying possible duplicate instances during the assessment process, we quantify this metric as:

$$CN2(D) := 1.0 - \frac{size(r_{bf})}{size(\overline{ent})}$$

$$r_{bf} := \{r \in \overline{ent} \mid isSet(hash(r))\}$$

where, *hash* is the hashing function used in the bloom filters, and *isSet* is the function that checks if the resource was seen or not before. r is a resource whose hash bits might have been set before, thus indicating a possible duplicate resource. This metric is discussed in more detail in our previous publication [19, §5.2].

Discussion: Our assessment estimated that overall the assessed datasets had a mean of around 7.6% redundant resources. Nevertheless, this does not mean that there is low redundancy on the whole Web of Data, since the standard deviation stands at 13.22% (median 99.34%), which suggests a moderately varied quality value overall. Around 13% of the datasets had a quality value less than the lower whisker, i.e. 78.55%. The range of quality values, including outliers, is 62.31%.

For this estimate value, we used 13 filters with a size of 5,500,000 each, ensuring efficient runtime with a low loss in precision (cf. [19, §6]). Around 76% of the datasets scored a value of 90% or more, meaning that the level of redundancy in these datasets is on the low side. Publishers should keep redundancy at a low level, and ensure that identical resources are not recurrent throughout the dataset. This can be done by creating `owl:sameAs` links between identical resources, without repeating property-value triples.

(CS1) Entities as Members of Disjoint Classes

The Web Ontology Language (OWL) extends the RDFS expressivity by modelling primitives that are otherwise difficult to express in RDFS. Generally, the OWL axioms deal with restrictions that can be placed on an otherwise open world assumption. On the other hand, incorrect usage of OWL features results in inconsistencies and thus jeopardizes reasoning.

The `owl:disjointWith` property is used to *"guarantee(s) that an individual that is a member of one class cannot simultaneously be an instance of a specified other class"* [50, §5.3]. One of the most popular examples of disjoint classes can be found in the FOAF vocabulary, where `foaf:Person` and `foaf:Document` are defined disjoint, which means that the resource `HarryPotter` (as an example) cannot be both a person and a document.[37] This metric is classified under the consistency dimension in [55].

Metric Computation: Metric CS1 checks for disjointness between types in multi-typed resources. Moreover, each assessed type has its super-types inferred[38] in order to check explicitly declared disjointness also between parent classes. Along these lines we quantify this metric as follows:

$$CS1(D) := 1.0 - \frac{\sum_{r \in \overline{ent}} hasDisjointTypes(r)}{size(\overline{ent})}$$

$$hasDisjointTypes(r) := \begin{cases} 1 & \overline{r_{dis}} \neq \emptyset \\ 0 & otherwise \end{cases}$$

[37] Instead, one would have to distinguish between a person `HarryPotterCharacter` and a document `HarryPotterNovel`.

[38] In this article, more specifically the consistency metrics, when we talk about inferencing we refer to the materialisation of the type hierarchy. No other reasoning mechanism is used.

$$\overline{r_{dis}} := \{t \in \overline{types}(r) \mid \exists t' \in \overline{types}(r). \; t \textbf{ owl:disjointWith } t'\}$$

where $\overline{types}(r)$ is the set of the types a resource is a member of and their inferred supertypes (**r rdf:type/rdfs:subClassOf* t**).

Discussion: The assessment shows that, in the presence of the simple inferencing explained above, almost all of the assessed datasets observe the `owl:disjointWith` property and their entities do not violate this property's restriction. In total around 98% of the datasets score a value of 100% for this metric, whilst the other two datasets score a value of more than 99.9%, therefore still considered as of high quality. The mean quality value for this metric is 100%, whilst the standard sample deviation (σ_s) is 0% (median is 100). Such low values in OWL inconsistencies were also reported in [28], were the authors attribute inconsistency problems caused by various incompatible exporters, such as FOAF exporters.

(CS2) Misplaced Classes or Properties

RDF Schema provides property-centric mechanisms for defining classes (`rdfs:Class`) and properties (`rdf:Property`) in vocabularies [16]. This means that:

> *"instead of defining a class in terms of the properties its instances may have, RDF Schema describes properties in terms of the classes of resource to which they apply."* – [16, §2]

The RDF data model is represented by a *triple form* (**subject** × **predicate** × **object**), where the predicate is expected to be a property that describes a resource in the subject position and its value in the object position. On the other hand, a class URI defining the type of a resource is usually in the object position when `rdf:type` is in the predicate position. The RDF data model is flexible allowing *any* resource URI to be in the predicate position. Therefore, whilst in OWL this practice is prohibited (unless OWL 2 punning is used), the data model does not prohibit publishers to have a defined class in the *predicate* position and a property in the *object* position, but this could cause problems when agents are interpreting the data. In this metric triples having the OWL axioms `owl:equivalentProperty` or `owl:inverseOf` are excluded from the assessment, as these require a property to be in the *object* position. This metric is classified under the consistency dimension in [55].

Metric Computation: The misplaced classes or properties metric assesses the datasets' statements in order to check the correct usage of classes and properties. More specifically, this quality indicator checks if the assessed dataset has defined classes placed in the triple's predicate and defined properties in the object position. We quantify this metric as follows:

$$CS2(D) := 1.0 - \frac{size(\overline{c_{misp}}) + size(\overline{p_{misp}})}{size(quads(D))}$$

$$\overline{c_{misp}} := \{c \in class(D) \mid \{(c, \texttt{rdf:type}, P) \mid P \in \mathcal{P}\} \cap V(ns(c)) \neq \emptyset\}$$

$$\overline{p_{misp}} := \{p \in prop(D) \mid \{(p, \texttt{rdf:type}, C) \mid C \in \mathcal{C}\} \cap V(ns(p)) \neq \emptyset\}$$

In other terms, this metric checks for triples using a class c that is in the set of properties of the vocabulary of c (as defined in Metric IN3), which would mean that c is wrongly placed as a resource type, and similarly for properties p. A high value of this metric is interpreted as conformance to usage of classes and properties in a dataset.

Discussion: The usage of classes as properties and vice-versa are not common in the assessed datasets. Overall, 83% of the datasets score a value of 100% whilst the rest score 99.99%. The σ_s value for this metric is 0.01% (median 100%), which shows a very low deviation, whilst the mean is 99.99%. Upon further inspection, we saw that no properties were used in the object position of an `rdf:type` triple, although classes such as `http://creativecommons.org/ns#License` were used infrequently (two instances in this case) as properties. Figure 7 shows the box plot for metric CS2.

We noticed that in the iteration of this metric we did not exclude triples with properties that are also expecting a property in the object position of the triple, such as `rdfs:subPropertyOf` and `owl:onProperty`. These might change the quality value of some datasets, however, it will not change the computation of the metric.

(CS3) Misused OWL Datatype or Object Properties

OWL differentiates between properties referring to individuals (`owl:ObjectProperty`) and properties referring to data values (`owl:DatatypeProperty`). Incorrect usage of properties in this regard might lead to inapt functioning of an agent, for example, if a Linked

Data viewer is using `owl:ObjectProperty` and `owl:DatatypeProperty` characteristics in order to hyperlink objects or not. Zaveri et al. [55] classify this metric under consistency.

Metric Computation: This quality indicator assesses a dataset's statements for the correct usage of the predicate in terms the `owl:DatatypeProperty` and `owl:ObjectProperty` axioms. Therefore, this metric detects "erroneous" triples where a data value (literal) object is attached to an `owl:ObjectProperty`, and an entity (individual) to an `owl:DatatypeProperty`. Following this description, the metric can be formalised as follows:

$$CS3(D) := 1.0 - \frac{size(\{t \in D \mid misusedOWL(t)\})}{size(quads(D))}$$

$$misusedOWL(t) :=$$
$$(isLiteral(t.o) \wedge isOP(t.p)) \vee$$
$$(isIndividual(t.o) \wedge isDP(t.p))$$

where *isLiteral* is a function that returns *true* if the assessed triple's object is a literal (i.e. data value), *isIndividual* is a function that returns *true* if the assessed triple's object is a URI or a blank node, *isOP* and *isDP* are functions that check if the assessed triple's predicate has explicitly been defined by its vocabulary as an `owl:ObjectProperty` or `owl:DatatypeProperty` respectively (in analogy to checking for classes or properties as introduced in the context of Metric IN3). A high value of this metric indicates a low amount of misused properties.

Discussion: Figure 7 shows the box plot for metric CS3. Similar to the previously discussed metrics for this dimension, the datasets adhere to a high quality score (mean 98.88%) and a considerably low standard deviation value of 5.17% (median 100%). Overall, around 87% of the datasets scored 100% whilst in total 95% of the datasets scored 90% or higher. Nonetheless, the box plot shows that around 12% of the assessed datasets are outliers. The dataset with the lowest quality value scored 52.60%. Comparing this against the study on Microdata formats on the Web in [38], it is confirmed that on *average* datasets on the LOD cloud misuse OWL datatype and object properties less.

From our assessment the following datatype properties (top five) were used with resources:

- `http://swrc.ontoware.org/ontology#series` (28,269 times)
- `http://swrc.ontoware.org/ontology#journal` (21,731 times)
- `http://reegle.info/schema#sector` (1,876 times)
- `http://rdf.myexperiment.org/ontologies/components/link-datatype` (502 times)
- `http://eunis.eea.europa.eu/rdf/species-schema.rdf#sameSpeciesRedlist` (4 times)

whilst the following are object properties (top five) with literals:

- `http://www.europeana.eu/schemas/edm/collectionName` (50,000 times)
- `http://lexvo.org/ontology#represents` (49,966 times)
- `http://xmlns.com/foaf/0.1/based_near` (45,233 times)
- `http://vivoweb.org/ontology/core#dateTime` (25,538 times)
- `http://purl.org/NET/c4dm/event.owl#place` (7,952 times)

(CS4) Usage of Deprecated Classes or Properties

Removing classes and properties from schemas renders data using them incoherent. OWL introduces the two classes `owl:DeprecatedClass` and `owl:DeprecatedProperty` for such situations. Any class or property that is declared an instance of these is no longer recommended to be used in published data. This metric is classified under the consistency dimension in [55].

Metric Computation: This metric checks whether deprecated terms are used in a dataset. More specifically, all used classes and properties are checked if they are members of `owl:DeprecatedClass` or `owl:DeprecatedProperty` respectively:

$$CS4(D) := 1.0 - \frac{size(\overline{c_{dep}} \cup \overline{p_{dep}})}{size(class(D) \cup prop(D))}$$

$$\overline{c_{dep}} := \{c \mid (c, \texttt{rdf:type}, \texttt{owl:DeprecatedClass}) \in V(ns(c))\}$$

$$\overline{p_{dep}} := \{p \mid (p, \texttt{rdf:type}, \texttt{owl:DeprecatedProperty}) \in V(ns(p))\}$$

Discussion: With around 97% of the datasets scoring a quality value of 100%, data publishers tend to avoid using deprecated classes and properties. The LOD Cloud sample that was assessed used the minimal deprecated terms in most cases, with the lowest quality score of 97.41% marked as an outlier in the box plot (CS4) in Figure 7. The standard deviation, as in the other consistency metrics, is very low (0.23%), and the mean value is 99.97%.

(CS5) Valid Usage of the Inverse Functional Property

In the real world, a public key used for encryption is unique to every individual. If we want to represent this public key in a Linked Data document, then there should be one exactly one resource (possibly an individual of the type `foaf:Agent`) describing this public key, in order to represent this uniqueness between the key and the individual. Such properties are termed as *inverse functional*, meaning that if two different resources share the same value for that property, during reasoning or smushing[39] these two resources are treated as the same in the sense of `owl:sameAs`. The OWL vocabulary provides a class `owl:InverseFunctionalProperty`, which a property with the semantics described above should be a member of. Common examples of such properties include `foaf:mbox` and `foaf:homepage`. This metric is classified under the consistency dimension in [55].

Metric Computation: This quality indicator checks for incoherent values within the assessed dataset's values. More specifically, this metric checks if a value attached to a property member of `owl:InverseFunctionalProperty` (IFP) is shared by two or more *different* resources. With regard to the definition of different resources, this metric follows the *local unique name assumption*. In this metric, we only consider those statements that have an inverse functional property. We quantify this metric as follows:

$$CS5(D) := 1.0 - \frac{size(\overline{v_{IFP}})}{size(\overline{p_{IFP}})}$$

$$\overline{v_{IFP}} := \{t \in quads(D) \mid ifp(t.p) \wedge \exists \bar{t} \in quads(D).\varphi(t,\bar{t})\}$$

$$\overline{p_{IFP}} := \{t \in quads(D) \mid ifp(t.p)\}$$

$$\varphi(t,\bar{t}) := t.s \neq \bar{t}.s \wedge t.p = \bar{t}.p \wedge t.o = \bar{t}.o$$

where $ifp(t.p)$ checks whether the predicate of a triple is a member of `owl:InverseFunctionalProperty`, $\varphi(t,\bar{t})$ returns true if the two triples t and $\bar{t}$ violate the IFP characteristic. This definition does not work in the absence of inverse functional properties, as we would then divide by zero. In this case, we define $CS5(D) := 1$.

Discussion: The box plot for this metric (CS5) in Figure 7 shows the trend in this metric where a large part of the assessed datasets have no varying quality, bar a few number of datasets that are considered as outliers. These outliers, around 18% of the assessed datasets, increased the standard deviation to 12.29%, whilst the calculated median is 100%.

One should keep in mind that not all assessed datasets made use of inverse functional properties and were thus given a 100% score (since there was no triple breaking the IFP constraint); nevertheless, these where included in the assessment. From the assessment, around 3% of the datasets got a quality score of less than 50%.

Triples with the following IFP properties (top 5) where identified with violations during our assessment:

- `http://xmlns.com/foaf/0.1/homepage` (violated in 2861 triples)
- `http://rdf.myexperiment.org/ontologies/base/has-friendship` (violated in 635 triples)
- `http://eunis.eea.europa.eu/rdf/species-schema.rdf#sameSynonymGBIF` (violated in 380 triples)
- `http://eunis.eea.europa.eu/rdf/species-schema.rdf#sameSynonymITIS` (violated in 328 triples)
- `http://eunis.eea.europa.eu/rdf/species-schema.rdf#sameSynonymFaEu` (violated in 215 triples)

[39] This term is often used to name the process of aggregating resources based on inverse functional properties (`https://www.w3.org/wiki/RdfSmushing`).

Since each dataset is assessed individually, our assessment did not point out possible IFP violations across the assessed datasets. In order to ensure that the IFP constraint is not violated, data publishers should ensure that data values (such a email address, homepage) are validated for uniqueness before publishing, possibly across the Web of Data and not just locally in the dataset.

(CS6) Ontology Hijacking

Hogan et al. defined *ontology hijacking* as the *"redefinition or extension of a definition of a legacy concept [...] in a non-authoritative source"* [29]. s being an authoritative source for concept c means that the namespace of c coincides with that of s. For example, `http://xmlns.com/foaf/0.1/` is the authoritative source for the concept `foaf:Person`. Ontology hijacking may lead to incorrect inferencing throughout the data [29]. Nevertheless, ontology hijacking can be seen as restricting the Linked Data principle of an open world assumption, in a sense that discouraging ontology hijacking restricts what one can say about some concept. Zaveri et al. [55] classify this metric under consistency.

Metric Computation: This metric assesses a dataset for its redefinition of third party external classes and properties. More specifically, this metric identifies if a dataset is the authoritative document for all classes and properties it defines, following the axioms identified in [29]

Along these lines, we quantify the metric as follows:

$$CS6(D) := 1.0 - \frac{size(\{t \in tdef(D) \mid \mathcal{H}(t)\})}{size(tdef(D))}$$

where $tdef(D)$ is the set of triples in dataset D having one of predicates or objects for which hijacking rules have been defined, and the $\mathcal{H}(t)$ function checks whether triple t violates its corresponding hijacking rule.

Discussion: Similar to the Metric CS5, the variation in quality within most of the assessed datasets ($\approx$ 86% of the datasets) is very low, though a number of outliers (shown in Figure 7 Metric CS6) causes a standard deviation of around 19.99% (median is 100%). Furthermore, the mean value is 93.64%. Overall, publishers tend to avoid redefining terms that they are not authoritative to do so, with around 85% scoring a quality value of 100%. In general, publishers should try to avoid redefining terms, but instead they should extend existing terms (if needed), thus avoiding the confusion that can be caused by term cross-definition.

(CS9) Usage of Incorrect Domain or Range Types

In a schema, a property can optionally have domain and range defined. The domain is the expected type (class) of the subject of a triple using the given property. The range is the expected type (class of a resource, or datatype of a literal) of the object of such a triple. Using the incorrect domain or range types makes the data incoherent, as consumers who know the underlying schemas could query the data without looking at it, making it harder to retrieve the right or all results. Zaveri et al. [55] classify this metric under consistency.

Metric Computation: This metric assesses a dataset for the compliance of the types of the subjects and objects of its statements with the domains and ranges of its predicates according to the schema of the respective predicate. In particular, the predicate of each triple is dereferenced to identify the expected domain and range types. No type of a subject (or object, respectively) of a triple must be disjoint with any of the types in the domain (or range, respectively) of the predicate of the triple. At the typical scale of Linked Open Datasets it is, however, prohibitively expensive to infer all types of a resource and all disjointnesses between classes. For these reasons, we define this metric to efficiently compute the practically most relevant sub-case: checking whether the type of the subject/object of a triple or any of its superclasses matches the domain/range. For each triple being assessed, hierarchical inferencing is done on the subject and object in order to determine the parents of their[40]. This is required as we might have predicates with an abstract domain or range; for example, the `foaf:mbox` predicate has `foaf:Agent` as its domain. Therefore, if we assume a triple `:j foaf:mbox <...>` and `:j` is typed as `foaf:Person`, then we need to know that `foaf:Person` is a subclass of `foaf:Agent`, the domain type required for the `foaf:mbox` predicate. We define this metric as follows:

$$CS9(D) := 1.0 - \frac{size(\overline{dom}(D)) + size(\overline{ran}(D))}{2 \times size(D)}$$

$$\overline{dom}(D) := \{t \in D \mid dom(t.p) \in \mathcal{T}(t.s)\}$$

$$\overline{ran}(D) := \{t \in D \mid ran(t.p) \in \mathcal{T}(t.o)\}$$

[40]The hierarchical inference stops one level before `owl:Thing` and `rdfs:Resource`, the universal super-concepts.

where $\mathcal{T}(r)$ is a function that returns the type of the local resource[41] r together with its inferred parents, the functions $ran(p)$ and $dom(p)$ return a set of range and domain types respectively for the predicate p.

Discussion: This metric is implemented as a probabilistic metric using reservoir sampling as explained in [19]. Our assessment shows that data publishers tend to use incorrect domain and range types in the triples. Around 4% of the assessed datasets had a quality score of 90% or more, with the highest score being 99.51%. On the other hand, around 13% of the datasets scored less than 50%. The mean score for this metric is 60.11% whilst the standard deviation is around 13.43%. The box plot for Metric CS9 in Figure 7 is symmetrical with the median standing at 57.14%. It also depicts a set of outliers over the top whisker and one dataset marked as outlier under the bottom whisker. It is also lower than the rest of the consistency metrics (Metrics CS1 to CS6), suggesting that Linked Data publishers might be more complacent with using the right datatypes when creating resource triples. Linked Data publishers should be aware of the domain and ranges of the properties used in their datasets by consulting with the relevant vocabularies. Furthermore, simple on-the-fly type checking scripts can be created and used throughout the publishing activities, inspecting for such schema-to-data inconsistencies.

Since this metric is an estimate metric, the bias of these results lie within the reservoir sampler data objects being assessed, which can be under-represented. On the other hand, in [19] we have shown that with the right parameters probabilistic approximation techniques can provide a good estimate quality value.

(SV3) Compatible Datatype

Ranges with a data value (i.e. literal) are usually constrained to be of a certain datatype, for example, a property `ex:age` would have an `xsd:integer`. Being an important component in the RDF data model, literals are used to represent values for properties, whilst the datatype attached to the value can be used to interpret the value concisely. Beek et al. describe four benefits of having good quality literals including *efficient computation* [9]. This means that having a canonical representation of the datatype ensures a unique representation of a literal across the Web of Data, and thus actions such as comparing two literals of the same type would be easy [9]. It is recommended that publishers add explicit datatype information to literals. This metric is classified under the syntactic validity dimension [55].

Metric Computation: This quality indicator assesses the lexical form of the data values against the data type attached with the literal itself. Consider `"10"^^xsd:integer`, the value `10` is what is known as the lexical form, whilst `xsd:integer` (i.e. `http://www.w3.org/2001/XMLSchema#integer`) is its datatype. Along these lines we quantify this metric as follows:

$$SV3(D) := \frac{size(\{v \in lit_t(D) \mid \vartheta(v_{lf}, v_{dt})\})}{size(lit_t(D))}$$

where $lit_t(D)$ is the set of all *typed* literals, $\vartheta(v_{lf}, v_{dt})$ is a function that checks the validity of the value's lexical form v_{lf} against the value's datatype v_{dt}[42]. Untyped literals are ignored in this metric as they cannot be validated against an unknown datatype.

Discussion: The box plot for metric SV3 in Figure 7 shows that most of the datasets assessed adhere to a 100% quality value, though there were also a number of datasets that scored less and thus are marked as outliers. On average, the quality score of the assessed dataset is around 96.80% whilst the standard deviation is a high 14.16% (median 100%). Datasets that had no literal values were omitted from this assessment.

Similar to Meusel's and Paulheim's findings [38], our assessment identified `xsd:date` as the most problematic datatype, present in around 37.24% of the problematic triples. A literal typed with the `xsd:date` datatype should have the ISO 8601 lexical form of **CCYY-MM-DD**, where CC represents the century, YY year, MM month, and DD day [12]. Nonetheless, we found errors that can be grouped as follows:

- literals with a lexical form of `xsd:dateTime`;
- literals with a lexical form of `xsd:gYearMonth`;
- literals with a CCYY-M-D lexical form;
- literals with a CCYY-MM-D lexical form;
- literals with a lexical form of `xsd:gYear`;

[41] External resources are ignored as we assume a closed world during the assessment. Thus, only resources with locally defined types are included.

[42] Apache Jena (`https://jena.apache.org`) offers a function that validates a value's lexical form to its defined datatype. This metric validates only Jena pre-registered datatypes as defined in `https://jena.apache.org/documentation/notes/typed-literals.html#xsd-data-types`

– literals with an incorrect date, e.g. 9999-99-99 or 0000-00-00.

The other violated datatypes are:

– `xsd:double` ≈ 19.18%;
– `xsd:dateTime` ≈ 16.25%;
– `xsd:gYear` ≈ 16.01%;
– `xsd:gMonth` ≈ 6.13%;
– `xsd:integer` ≈ 3.04%;
– `xsd:int` ≈ 1.26%;
– `xsd:anyURI` ≈ 0.69%;
– `xsd:positiveInteger` ≈ 0.16%;

In order to reduce incompatible datatypes vis-à-vis the lexical form of a data value, publishers could publish and serialise their data using the latest Turtle 1.1[43] parser, as it relaxes and simplifies the serialisation of such literals.

Aggregated Results

Figure 7 shows a box plot illustration of the aggregated quality value compared with the category's metrics. The overall aggregated box plot shows a population with little dispersion (most of which results from Metric CS9) having a standard deviation of 12.89%, a median of 80.94% and mean of 77.36%. The majority of the metrics shows that a relatively high quality is adhered to by Linked Data publishers.

5.5. Accessibility Category

The dimensions in the accessibility category address the ease with which machines as well as humans can (re)use Linked Data resources. Zaveri et al. classify metrics under the following dimensions [55]: (i) *availability* – dealing with the access methods of the data; (ii) *licensing* – what are the permissions (if defined) to re-use a dataset; (iii) *interlinking* – the degree of internal and external interlinks between data sources; (iv) *security* – deals with the security and authenticity of datasets; (v) *performance* – how does the server hosting a dataset affect the efficiency of consuming data. In this section we assess metrics related to the *availability* dimension (2 metrics), *licensing* dimension (2 metrics), *interlinking* dimension (1 metric), and *performance* (2 metrics). Metrics in this category were measured at a single point in time, meaning that datasets might report different quality metric results over time. This is because data servers might have planned or unplanned down time. For this analysis, we only report the quality value of a dataset at one point in time, similar to what was done for other metrics. However, in the future we plan to have a service similar to SPARQLES[44] which tracks and assesses datasets over the mentioned accessibility metrics over time, visualising results online. In Table 5 we list the metrics that are assessed in this category, together with a summary of assessment results showing the mean value (μ), median value (Q_2), and standard deviation (σ_s).

	Metric Name	μ	Q_2	σ_s
A3	Dereferenceability of the URI	36.86%	34.11%	36.54%
L1	Machine-Readable License	14.4%	0%	35.11%
L2	Human-Readable License	8.8%	0%	28.33%
I1	Links to External Linked Data Providers	27.01 links	1 link	183.3 links
PE2	High Throughput	47.78%	29.67%	45.60%
PE3	Low Latency	57.55%	99.23%	47.12%

Table 5

List of metrics assessed in the Accessibility Category together with the assessed mean value (μ), median value (Q_2), and standard deviation (σ_s).

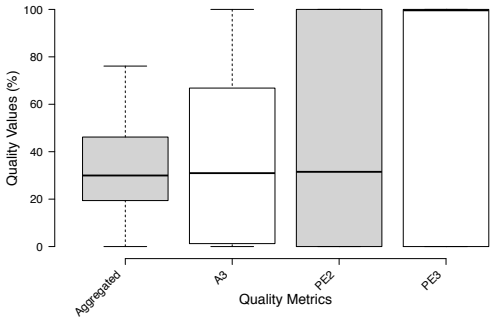

Fig. 8. Accessibility category box plot. Outliers are represented by dots.. Machine-Readable License, Human-Readble License, and Links to External Data Providers metrics are excluded, but included in the aggregated result box plot.

(A3) Dereferenceability of the URI

Dereferenceability is one of the main principles of Linked Data. HTTP URIs should be dereferenceable, i.e., HTTP clients should be able to retrieve the resources identified by the URI. According to the LOD principles, a typical web URI resource would return

[43] `http://www.w3.org/TR/turtle/`

[44] SPARQL Endpoint Status; `http://sparqles.ai.wu.ac.at`

a 200 OK code indicating that a request is successful and a 4xx or 5xx code if the request is unsuccessful. In Linked Data, a successful request should return an RDF document containing triples that describe the requested resource. Resources should either be hash URIs or respond with a `303 See Other` (redirect) code [47].

Metric Computation: The aim of this metric is to check the number of valid dereferenceable URIs used (according to these LOD principles) in a data source. More specifically, an `HTTP GET` request is performed on a URI defining a concept, together with a header accepting a variety of MIME types valid for Linked Data (including application/rdf+xml, text/n3, text/turtle, etc.). Unless a resource's URI is a hash URI, a correct server-side dereferencing mechanism should identify that the requested resource is a *real-world object* or *abstract concept* (a "non-information resource") and thus reply with a `303 See Other` and a redirect location where the *data object* in the desired format is (an "information resource"). Heath and Bizer explain that *"where URIs identify real-world objects, it is essential to not confuse the objects themselves with the Web documents that describe them"* [26, §2.3.1].

This `303` redirection is handled automatically by the client, with the server responding with a `200 OK` together with the semantically described object in the requested format. This metric checks all local and non-local URIs for dereferencability. Along these lines we adapt the metric from Hogan et al. [30, §5.1, Issue III]:

$$A3 := \frac{size(\{u \in dlc(D) \cap \mathcal{U} \mid deref(u)\})}{size(dlc(D) \cap \mathcal{U})}$$

where $\mathcal{U}$ is the set of URIs in the dataset D (of which we consider those that are data-level constants), and $deref(u)$ is a function that returns *true* if the URI u being examined follows the dereferencability rules.

Discussion: In [19, §5.1] we describe a probabilistic technique for this metric using reservoir sampling. In this technique, each resource URI is split into two parts: (1) the Pay-Level domain (PLD), and (2) the path to the resource. This is analogous to a dictionary data structure. For this metric we employ a "global" reservoir sampler for the PLDs. Furthermore, for each PLD we employ another sampler holding an evenly distributed sample list of resources to be dereferenced. However, such sampling might lead to an unbalanced representative sample. We therefore adopt a hybrid of the *reservoir* technique used in [19, §5.1] and the *stratified sampling* idea as described in [25]. Stratified sampling is a technique that can be used when the data can be partitioned into a number of disjoint subgroups [25]. The idea is that the sample is chosen per proportion of these subgroups, therefore improving the representative sample. Therefore, for each domain in the higher-level sampler, we keep track of the total number of items encountered during assessment, to simulate the size of the strata groups. A final sampler is then drawn from all lower-level sampled resources using the proportionate allocation method. The parameters used were 5000 as the global reservoir size (i.e. the number of possible different pay-level domains (PLD) in a dataset), and a PLD size of 10000. Nevertheless, one must keep in mind that these parameters introduce a bias in our results in a way that the sample might be under-represented.

The box plot for metric A3 in Figure 8 shows a large variance with values ranging 100% to 0%. The mean quality value of this metric is 36.86%, which is 33.44% lower than the mean recorded in [30, §5.1 – Issue III]. There are two reasons for this difference. First, in our study we do not just study local dereferenceable URIs, but we also take into consideration the dereferenceability of external resources the publishers use. This might mean that publishers are punished if the external URIs used are not dereferenceable. Secondly, we noticed that certain hosts blacklisted our IP address during this assessment following numerous HTTP requests. The box plot for metric A3 in Figure 8 is positively skewed, meaning that the assessment shows a high concentration of low quality values. Similar to [30, §5.1 – Issue III], our assessment shows a high variability between data producers on the dereferencability of resources. We report a standard deviation of 36.54%, with a median of 31.11%. In total our assessment attempted to dereference a total of 709,356 resources, out of which only 233,127 where valid dereferenceable resources. The rest of the resources resulted in the following problems:

- Hash URIs without parsable content – 5 resources;
- Status Code 200 – 61,922 resources;
- Status Code 301 – 7,281 resources;
- Status Code 302 – 13,878 resources;
- Status Code 303 without parsable content – 1,293 resources;
- Status Code 307 – 1 resource
- Status Code 4XX – 104,379 resources;
- Status Code 5XX – 5,444 resources;

– Failed Connection (either due to blacklisting or resource not online anymore) – 289,289 resources.

Surprisingly, not a lot of publishers abide by the dereferenceability guideline. Our assessment shows that only 33% of the assessed datasets have a dereferenceability value of 50% or more. Whilst this guideline is one of the Linked Data principles, one should understand the extra costs this mechanism requires, including the maintenance of content-negotiation and redirection schemes. However, one must investigate if the need of the dereferenceability mechanism is a must in Linked Data, or if agents can be adapted to understand Linked Data URIs automatically, for example by introducing catalog mechanisms similar to those used in XML to resolve URI references. In the meantime, a possible solution is that data publishers make use of Linked Data-based content management systems (such as OntoWiki[45]) that handle such mechanisms automatically.

Licensing

> *"It is a common assumption that content and data made publicly available on the Web can be re-used at will. However, the absence of a licensing statement does not grant consumers the automatic right to use that content/data."* – [26, §4.3.3]

Open licences, as defined by the Open Definition [41], are the heart of open data. They specify whether third parties can re-use or otherwise, and to what extent. In Linked Open Data, one would expect that such licences are either machine-readable using predicates such as `dct:license`, `dct:rights` and `cc:licence`, or at most human-readable (e.g. within `dc:description`). Such a license specification should be included in a dataset's metadata.

(L1) Machine-Readable License

Having machine-readable license definitions (such as those in the `http://purl.org/NET/rdflicense` dataset [46]), agents would be able to consume (for example to visualise) different parts of the license, such as the jurisdiction and duties (e.g. share-alike or attribution). Furthermore, agents would be able to understand the limitations of a license, and make informed decisions (e.g., if resources can be used within paid services) with less human interaction.

Metric Computation: The aim of this metric is to check if a dataset has a valid machine-readable license. By valid we mean that a license can be retrieved from a semantic resource (e.g. `http://purl.org/NET/rdflicense/.*`) or with an `owl:sameAs` link to one of the following URLs:

- `http://(www.)?opendatacommons.org/licenses/odbl.*`
- `http://(www.)?opendatacommons.org/licenses/pddl/.*`
- `http://(www.)?opendatacommons.org/licenses/by/.*`
- `http://creativecommons.org/publicdomain/zero/.*`
- `http://creativecommons.org/licenses/by/.*`
- `http://(www.)?gnu.org/licenses/.*`
- `http://creativecommons.org/licenses/by-sa/.*`
- `http://(www.)?.gnu.org/copyleft/.*`
- `http://creativecommons.org/licenses/by-nc/.*`
- `http://purl.org/NET/rdflicense/.*`

These should be attached to one of the following "license" predicates:

- `dct:license`;
- `dct:rights`;
- `dc:rights`;
- `xhtml:license`;
- `cc:license`;
- `dc:licence`;
- `doap:license`;
- `schema:license`.

We define this metric as follows:

$$L1(D) := \exists t \in D.lpr(t.p) \wedge lvld(t.o)$$

where, $lpr(t.p)$ is a function that checks the triple's predicate against the set of defined license predicates, and $lvld(t.o)$ is a function that checks if the triple's object is a valid machine-readable license. This metric returns true if the assessed dataset has a valid machine-readable license.

Discussion: In Section 3.2 we discuss the licences and rights in the LOD Cloud datasets' metadata. We show that around 41% of the whole LOD

[45] `https://ontowiki.net`

Cloud datasets have license or rights metadata, using the predicates `dct:license`, and `dct:rights`. In this metric we assessed the acquired data dumps and SPARQL endpoints for machine-readable licenses. The quality assessment resulted in just 17 datasets ($\approx$ 13%) that contained at least one machine-readable license. 14 out of the 17 datasets identified during this assessment had licence metadata in the LOD Cloud metadata. The other three datasets had licence information directly in the data itself. Machine-readable license statements can be easily included in a dataset by using other linked open datasets such as [46].

(L2) Human-Readable License

In contrast to Metric L1, a human-readable license enables human agents to read and understand a license in textual format, rather than in terms of triple statements.

Metric Computation: The aim of this metric is to verify whether a human-readable license text, stating the licensing model attributed to the dataset, has been provided as part of the dataset itself. The difference from Metric L1 is that this metric looks for objects containing literal values and analyses the text searching licensing related terms. More specifically, we check for the following:

1. A license **description** triple, identified by a predicate `dct:description`, `rdfs:comment`, `rdfs:label`, or `schema:description` and a literal object matching the following regular expression: `.*(licensed?|copyrighte?d?).*(under|grante?d?|rights?).*`;
2. A license triple, identified by a triple with a license predicate described in Metric L1, and a URI pointing to a human-readable documents (also defined in Metric L1).

We define this metric as follows:

$$L2(D) := \exists t \in D.t.p \in p_{hrdesc} \wedge lregex(t.o)$$

where p_{hrdesc} is the set of predicates representing human-readable descriptions, and *lregex* is a function that checks a literal against the defined license regular expression. This metric returns true if the assessed dataset has a valid human-readable license.

Discussion: Similar to Metric L1, the assessment shows a low overall level of conformance to this metric. We detected human-readable licenses in eleven ($\approx$ 8.46%) datasets, four of which also had a machine-readable license. Whilst it is understandable that publishers are less inclined to having statements with large textual literals containing licensing data, we suggest that publishers should at least define the license name in the datasets' metadata. Licenses are of utmost importance to open data [41, §1], therefore, publishers should define the license or rights either as machine-readable (preferable) or at least human-readable.

(I1) Links to External Linked Data Providers

One of the main Linked Data principles is to *"include links to other URIs, so that [agents] can discover more things."* [11]. Furthermore, Berners-Lee states that linking your data to external sources would earn the dataset the fifth star, given that the rest of the four guidelines are satisfied. Having external links in a dataset would enable data consumers to explore and understand better the data in question. Additionally, Heath and Bizer [26] emphasize the importance of external RDF links in the Web of Data since:

> *"they are the glue that connects data islands into a global, interconnected data space and as they enable applications to discover additional data sources in a follow-your-nose fashion."* – [26, §2.5]

These external outlinks is what makes the Linked Data paradigm stand out from other best practices about data management. Well-interlinked data enables better analysis and understanding of the data. The interlinking property is often used to identify the importance or authority of a data source in the Web of Data. For example in [49], the interlinking degree is used to visualise the importance of datasets within the LOD Cloud.

Metric Computation: The aim of this metric is to identify the total number of external RDF links used within the assessed dataset. An external link is identified if the object's resource URI in a triple has a PLD different from the assessed dataset's PLD. Furthermore, the external link should be a semantic resource that can be dereferenced and parsed by an RDF parser. For this metric we use a reservoir sampling approach [19] to estimate the number of external Linked Data providers in a dataset[46].

[46]The sampling method is the same as in the cited literature, however, we implemented the metric in a different manner in order to calculate the number of external providers instead of a ratio, as described in this article (Metric I1).

Along these lines, we quantify the metric as follows:

$$I1(D) := size(\{pld(u) \mid (u \in (dlc(D) \setminus ldlc(D)) \cap \mathcal{U}) \wedge isParseable(u)\})$$

where $pld(u)$ is a function that returns the pay-level domain of the resource's URI (u), $ldlc(D)$ is the set of local DLCs, and $\mathcal{U}$ is the set of URIs in dataset D. The value returned by this metric is the number of valid (dereferenceable) external RDF links the assessed dataset has.

Discussion: Similar to Metric A3, this metric was assessed using a sampling technique. In this case, each external PLD has a sampler of maximum 25 items. Estimation techniques create a bias since the parameters might create an under-represented sample. In this case, we might miss out possible Linked Data documents that identify a PLD as external. Table 6 shows the top five assessed datasets, the number of unique dereferenceable external PLDs linked in the dataset, and the total number of unique PLDs. In the sample acquired from the LOD Cloud, only 9 datasets had no external PLDs, whilst around 88% of the datasets had less than 50 unique external PLDs linked. In total, the number of external PLDs extracted from the assessed 3.7 billion triples amounted to 977,609. Three datasets, namely `dbpedia.org`, `kent.zpr.fer.hr`, and `www.pokepedia.fr` accounted for around 97% of these PLDs. However, the actual number of PLDs with dereferenceable resources is 3086, which is around 0.31% of the linked external PLDs.

Considering the Linked Data principles, one would have expected a higher ratio of external RDF links. However, there is no set number of external Linked Data PLDs each dataset should have. The assessed datasets feature a large standard deviation of 183.3 Linked Data PLDs, and a mean of 27.01 Linked Data PLDs. Nevertheless, one should consider that these two statistical measures are highly influenced by the top two datasets.

(PE2) High Throughput

Ideally, a Linked Data host can accommodate a large number of requests without affecting the consumers' productivity. That is, a consumer is not left waiting "in a queue" until other agents have been served. Therefore, in an ideal situation, a host has the capacity to handle a large number of parallel requests.

Metric Computation: Adapting a definition by Flemming [21], the *high throughput* metric measures the efficiency with which a system can bind to the data source by measuring the number of HTTP requests answered by the source of the dataset per second. From the dataset we use reservoir sampling to "randomly" choose a maximum of ten local resources (i.e. whose namespace is the same as the data source namespace) that will be used for this metric. The metric estimates the number of requests served per second, computed as the ratio to the total number of requests sent to the dataset's host. We define this metric adopting [21] as follows:

$$PE2(D) := \begin{cases} 1.0 & \geq 5 \text{ requests answered in} \leq 1 \text{ s} \\ \frac{servedRequestsPerSec}{200\ ms} & otherwise \end{cases}$$

where *servedRequestsPerSec* is number of requests that the host served per second. If five or more requests can be answered in a second or less, then the metric's value is defined as 100%, otherwise a percentage is calculated as the ratio of the number of served requests against the ideal time (200ms) taken to serve one request.

Discussion: The box plot for this metric (PE2) in Figure 8 shows a large varying quality with values ranging from 100% to 0%. The standard deviation stands around 45.60% (median value is 29.67%) whilst the mean value is 47.78%. The box plot is positively skewed, suggesting that observations at the low end are concentrated. Around 38% of the assessed datasets gave a result of 100%, which means that more than 5 requests were answered in 1 second or less. Around 8.52% of the datasets scored a quality value between 50% (inclusive) and 100% (not inclusive). All quality results are dependent on the data host during the time of the assessment, therefore, such a quality assessment should be performed more frequently.

(PE3) Low Latency

Latency is the amount of time an agent has to wait until the host responds with the particular request. The time taken largely depends also on how big the HTTP request is, and the number of HTTP round-trips the server has to make before serving the request. Therefore, the choice of Hash URIs and 303 redirects (i.e. Slash URIs) is also an important factor for latency [21,26]. Hash URIs would reduce the number of HTTP round-trips, as the document with the requested fragment resource description would contain descriptions of other resources in the same document. Therefore, the client would end up receiving unnecessary resources that would eventually increase the latency (since the document size will be larger). On the

Dataset	I1(D)	# Unique PLDs
http://energy.psi.enakting.org	1402	1623
http://lobid.org/organisation	1395	1604
http://dbpedia.org/	32	346,708
http://vocabulary.semantic-web.at/PoolParty/wiki/semweb	13	291
http://lod.geospecies.org	11	42

Table 6

Top 5 ranked datasets for the links to external RDF data providers metric.

other hand, Slash URIs require the client to go through the whole dereferencing process for every resource, but then the client will receive exactly the required resource. Ideally, the data source should serve resource requests with the lowest possible latency, which in turn means that data publishers should choose the right strategy for publishing data (Hash vs. Slash).

Metric Computation: The *low latency* metric measures the efficiency with which a system can bind to the data source by measuring the delay between submitting a request for that very data source and receiving the respective response. Similar to Metric PE2, a reservoir sampler is used to sample a maximum of ten local resources from the dataset under assessment. This metric is defined as the mean time taken for ten requests to respond, normalised to a percentage value between 0 and 100 by dividing by an ideal response time defined as one second [21]. Along these lines, Metric PE3 is defined as follows:

$$PE3(D) := \begin{cases} 1.0 & \geq 1 \text{ requests answered in} \leq 1 \text{ s} \\ \frac{1000\ ms}{meanResponseTime} & \text{otherwise} \end{cases}$$

where *meanResponseTime* is the mean response time of the 10 sampled resources. A 100% low latency means that the data source can respond to a resource within up to a second, otherwise, the percentage value is calculated as a ratio of the number of possible requests served in one second.

Discussion: Similar to Metric PE2, results of these metrics rely on the data host at time of assessment. The box plot for this metric (PE3) in Figure 8 confirms the large range varying quality, as in Metric PE2. The standard deviation is 47.12% with a mean value of 57.55%. However, unlike PE2, the metric's values are negatively skewed, with a median value of 99.23%. This shows that there is a large concentration of very high quality values. Around 49.61% of the datasets have a quality value of 100%, meaning that at least one request is answered in one second or less.

Aggregated Results

Figure 8 shows a box plot illustration of the aggregated quality value compared with the category's metrics. The overall aggregated box plot shows a population that is moderately varied having a standard deviation of 19.00% and a median of 29.96%. The mean aggregated quality score is 33.12%, with only 19% of the assessed datasets scoring 50% or more. The aggregated value is affected by the low values of the licenses metrics (L1 and L2), which is a concerning matter considering that the assessed datasets are part of the Linked **Open** Data cloud. Not having a defined license might make the adoption of linked dataset more difficult.

5.6. Ranking and Aggregation Remarks

All categories had an aggregated value $v(C, 1.0)$ calculated using the user-driven ranking function defined in [17], with a default weight of 1.0. In order to calculate a ranking for integer-based metrics (Metrics V1, V2 and I1), we followed a positional-based ranking, similar to a definition by Hogan et al. [30]:

$$pb_m(D) := \frac{((size(\overline{D_m}) + 1) - pos_m(D)) \times 100}{size(\overline{D_m})}$$

where, m indicates the metric (e.g., V1), $\overline{D_x}$ is the set of datasets that where assessed for metric x, and pos_x is a function that returns the assigned position of dataset D following the assessment of metric x. All datasets were given a score based on a scale of 0 to 100%. In all cases 100% translates to the highest level of conformance to the quality metric being assessed, whilst 0% translates to the lowest level of conformance. The aggregated score for a dataset ($as(D)$) was calculated as follows:

$$as(D) := \frac{\sum_{m \in \{RC1,\ldots,PE3\}} scr_m(D)}{size(\{RC1, \ldots, PE3\})}$$

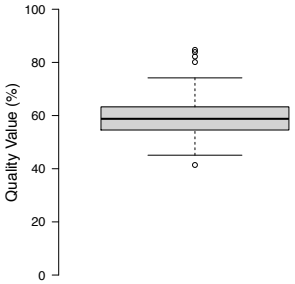

Fig. 9. Aggregated Conformance Score box plot. Outliers are represented by dots.

where scr_m is the score of a dataset for a metric m, and $\{RC1, \ldots, PE3\}$ is the set of metrics described in this article. The aggregated scores for each dataset took into consideration just the computed metrics, i.e. metrics that yielded a result. All quality results and the overall ranking are available at `http://jerdeb.github.io/lodqa/ranking` in a tabular format. Furthermore, readers of this article can filter and rank datasets according to their quality criteria on `https://w3id.org/lodquator`. Nonetheless, we do not claim that lower ranked datasets are hosting data of poor general quality, but just that these datasets are less conformant to the quality metrics assessed in this study, weighting each of them equally.

A total of 130 datasets were assessed in Luzzu, aggregating around 3.7 billion quads. The mean aggregated conformance score is 59.33% with a slight standard deviation of 7.63% (median value is 58.78%). Figure 9 depicts a symmetric box plot showing the spread of aggregated quality conformance scores. The box plot shows 5 outliers, four of which are "positive outliers", since their quality value is superior to the rest of the population.

Failing SPARQL endpoints

Most of the "failing" datasets are SPARQL endpoints, whilst others contained syntactic errors. In Luzzu, quality metrics are not written and executed on SPARQL endpoints, but instead triples are streamed from the endpoint[47] directly to the metric processors. In order to ensure that all triples are retrieved, the SPARQL processor makes use of the `ORDER BY` and `OFFSET` keyword, which takes much time to process especially on large knowledge bases. If the `ORDER BY` is removed, the endpoint responds faster, but since order is not guaranteed, multiple executions of the same query might result in different results. On the other hand, various endpoints have different settings, for example (i) (lack of) support of scrollable cursors – required for the query to stream triples; or (ii) different timeout settings (500 Server Error) – which might interrupt the assessment at a random point, thus failing the complete assessment.

[47] This is the only query the Luzzu framework does on the endpoint, until all results are retrieved.

6. Is *this* Quality Metric Informative?

In this section we present a statistical analysis of the quality assessment, primarily understanding which of the quality metrics assessed can potentially give the stakeholders more information on the quality of linked datasets than other metrics.

6.1. The Principal Component Analysis

The Principal Component Analysis (PCA) [43] is a statistical variable reduction technique that transforms a set of possibly correlated variables into a new set of uncorrelated components. Given some data, the PCA helps in finding the best possible characteristics to summarise the given data as well as possible. This is done by looking at the characteristics that provide the most variation across the data itself, ensuring that the data can be differentiated. On the other hand, the new set of uncorrelated components can be used to singularly describe correlated characteristics of the data. We will use the PCA in order to identify which of the assessed metrics are informative for Linked Data quality (cf. Section 6.2). This technique was favoured over ANOVA, which in simple terms is a technique usually used to determine whether there is significant difference between means. However, ANOVA was used in [8,39] to identify the quality metrics that are sensitive in images, for example what are the best metric(s) that should be used for images with watermarks. Nevertheless, these statistical tests gives an indication, that ideally is sustained with a subjective test.

6.2. Identifying the Informative Quality Metrics for a Generic Linked Data Quality Assessment

The aim of this analysis is to study how informative are the quality metrics assessed on the Linked Open Data Cloud. Therefore, our main research question for this analysis is:

What are the key quality indicators that are defined in Zaveri et al. [55] and assessed during this empirical study that can give us sufficient information about a linked dataset's quality?

Therefore we are using PCA to look at 27 different metrics in order to (1) reduce a number of quality metrics into a set of components that explain the variance of all quality values for all observations (linked datasets), and (2) possibly identify those metrics that are non-informative. The PCA helps us to find the best possible quality metrics that summarise the quality of linked datasets as well as possible, in terms of new characteristics (components). The process groups the quality metrics into a number of components, where all components are orthogonal and describe a substantial variability in the quality of Linked Data.

For this analysis we identify the following two hypotheses:

H_0**:** *No correlation exists among different metrics, thus each separate metric gives an informative value on the overall quality of a linked dataset.*

H_a**:** *Correlation exists among different metrics; therefore there are metrics that are non-informative to the overall quality value of a linked dataset.*

The null hypothesis (H_0) describes the scenario where all assessed metrics cannot be correlated and thus cannot be reduced to factors. We use the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) to check whether Principal Component Analysis (PCA) is appropriate for our data, and Bartlett's Test of Sphericity to check whether the null hypothesis (H_0) can be rejected.

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		0.96
Bartlett's Test of Sphericity	Approx. Chi-Square	991.81
	df	351
	Sig.	0.000

Table 7
KMO and Bartlett's Tests.

In Table 7 we display the results for the KMO and Bartlett's test. The KMO results shows that our data has an adequacy of 0.96, which makes the factor analysis appropriate for our data. Kaiser recommends that values greater than 0.5 are acceptable [32]. The 0.96 value means that correlation patterns between the input values (quality assessment output values) are compact and factor analysis will produce distinct and good factors. The Bartlett's Test of Sphericity confirms this and shows extremely strong evidence that correlation exists amongst different metrics.

Following the rejection of the null hypothesis, we use the Principal Component Analysis (PCA) in order to test the alternative hypothesis (H_a). Table 8 shows the total variance explained. In the initial eigenvalues column, the table displays the eigenvalues associated with each component (corresponding to our 27 quality metrics), and the total variance of the observed values for each factor. In simple terms, component 1 explains 12.75% of the total variance. Only components whose eigenvalues are greater than 1 are retained.

Therefore, the total number of components extracted is eleven. In order not to give too much importance to one component over another, a rotated component matrix (Table 9) is taken into consideration, in order to determine the informative quality metrics. The rotated component matrix is the main output following a Principal Component analysis. In total, these 11 factors can explain around 72.59% of the total variance.

In Table 9 we can see the 11 extracted components and the metrics each component represents. For the factor loading we use a cut-off point if the magnitude of the factor is at least 0.5 as suggested by Hair et al. [23] as the number of datasets is 130. This table also suggests which of the quality metrics, possibly combined (as in the case for components 1-9), are informative metrics.

By rejecting H_0, we are statistically confirming that most metrics on their own are not enough to provide an informative value on the quality of a dataset. Therefore, the PCA is used to create a descriptive summary of these metrics, which provides us with a number of components, thus proving our alternative hypothesis (H_a). Each component groups a number of quality metrics that defines an informative quality description. Recalling the research question for this study, the aim of this study is to highlight the key quality indicators that were classified in [55] and implemented in this empirical study. Therefore, for simplicity, we identify those metrics that are not in any of the 11 components as being metrics that describe the quality of a generic linked dataset in a non-informative manner. The PCA suggests that three metrics, namely Links to External Data Providers (Metric I1), Usage of Incorrect Domain or Range Datatypes (Metric CS9), and Dereferenceability (Metric A3), have values below the cut-off value for all of the 11 components.

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**	**Total**	**% of Variance**	**Cumulative %**
1	3.44	12.75	12.75	3.44	12.75	12.75	2.87	10.63	10.63
2	2.81	10.39	23.14	2.81	10.39	23.14	2.55	9.46	20.09
3	2.04	7.55	30.7	2.04	7.55	30.7	2.19	8.12	28.21
4	1.98	7.32	38.01	1.98	7.32	38.01	1.36	5.02	33.24
5	1.77	6.54	44.55	1.77	6.54	44.55	1.98	7.34	40.58
6	1.61	5.97	50.52	1.61	5.97	50.52	1.71	6.34	46.92
7	1.35	4.99	55.52	1.35	4.99	55.52	1.62	6	52.92
8	1.31	4.85	60.36	1.31	4.85	60.36	1.35	4.99	57.9
9	1.18	4.36	64.73	1.18	4.36	64.73	1.35	5.01	62.91
10	1.1	4.09	68.81	1.1	4.09	68.81	1.41	5.21	68.12
11	1.02	3.78	72.59	1.02	3.78	72.59	1.21	4.47	72.59
12	0.94	3.47	76.07						
13	0.88	3.27	79.34						
14	0.78	2.9	82.24						
15	0.72	2.68	84.92						
16	0.62	2.3	87.21						
17	0.58	2.14	89.35						
18	0.51	1.88	91.23						
19	0.48	1.77	92.99						
20	0.37	1.38	94.37						
21	0.34	1.26	95.63						
22	0.3	1.12	96.75						
23	0.26	0.97	97.72						
24	0.22	0.8	98.52						
25	0.16	0.61	99.13						
26	0.14	0.5	99.64						
27	0.1	0.36	100						

Table 8
Total variance explained.

Our initial quality assessment was generic, therefore all 130 datasets had the same 27 metrics assessed against them, irrelevantly of whether the metric is important to a particular dataset for a particular domain or not. This analysis helped us in identifying 11 orthogonal components that describe 72.59% of the variability. Hence, the results obtained after performing the PCA are just an indication of which metrics might not be informative in a generic Linked Data quality assessment. Furthermore, an initial observation of these components shows that some of these factors group metrics that follow the category and dimension categorisation of Zaveri et al. [55], however, this might not always mean that the grouped metrics are the same or similar. For example, if we take a look at the first component, we see that four metrics from the representational category are grouped together. This means that rather than having four metrics, we can create one metric gathering these four aspects in Linked Data quality. Moreover, one can notice that in the same component, Metric IO1 (re-use of existing terms) and Metric IN3 (usage of undefined classes and properties) are grouped together. From a conceptual point of view, these two metrics are highly related to each other since the first metric considers the usage of known existing vocabulary terms relevant for a particular domain, whilst the latter looks into the usage of undefined vocabulary terms. Similar observations can be identified in other orthogonal components. In the future, this analysis can act as a starting point for assessing the quality of the Web of Data. New metrics, representing the KQIs, can be developed to represent each component identified by our PCA findings. Furthermore, future quality assessment would not require the evaluation of three computationally time-consuming metrics (Metric CS9, Metric I1, Metric A3), as the described 11 components would be enough to describe the quality of a dataset.

	Components										
	1	2	3	4	5	6	7	8	9	10	11
IO1	0.85										
IN3	0.76										
V1	0.72										
V2	0.69										
CS9											
P2		0.86									
P1		0.78									
L1		0.74									
U1		0.58									
I1											
PE3			0.92								
PE2			0.91								
A3											
CS4				0.83							
CS6				0.63							
U3					0.93						
U5					0.89						
RC2						0.85					
IN4						0.81					
L2							0.8				
CS1							-0.75				
RC1								0.68			
CS2								0.61			
CN2									0.77		
CS3									0.68		
SV3										0.79	
CS5											0.84

Table 9
Rotated component matrix.

7. Concluding Remarks

Quality issues in datasets have severe implications on consumers who rely on information from the Web of Data. Currently, it is difficult for a consumer to find datasets that fit their needs based on quality aspects. The semantic quality metadata produced by this empirical study fills this gap. Prospective users can now search, filter and rank datasets according to a number of quality criteria, and more easily discover the relevant, fit for use dataset according to their requirements. Nonetheless, such an assessment should not be done once, but it should be a continuous (or periodical) process to reflect the dynamic Web of Data.

Large-scale empirical studies on data quality can raise awareness on the current problems in data publishing. Such empirical analyses are important to the community as (1) they help to understand what are the current (or recurring) problems, and (2) define future research and development directions – in this case of Linked Data. In this article we quantified and analysed a number of linked datasets vis-à-vis a number of quality metrics as classified in [55]. Furthermore, in Section 6.2 we statistically analysed the quality scores and performed the Principal Component Analysis (PCA) test in order to identify the non-informative Linked Data quality metrics in a generic assessment. This statistical method shows that following our assessment three out of 27 metrics were identified as non-informative to a datasets' quality. This empirical survey is one of the largest (in terms of triples) evaluation of LOD data quality to date. All quality metadata produced in this empirical study is published using Linked Data principles at `https://w3id.org/lodquator`.

In Section 3, we explained the *Open Data* principles and, using the LOD Cloud datasets metadata, performed a primary investigation to identify how well these abide by these principles. More specifically, we looked at the datasets' metadata in order to identify their accessibility points and licenses. We show that only around 42% had a valid Linked Data access point, whilst only 40% had a license.

In [27], Hitzler and Janowicz state that the general perception of Linked Data is that datasets are of poor quality. In line with research question described in Section 1 we looked at a number of datasets in order to understand better whether the perception label is deserved. In Section 5 we looked at the datasets themselves in order to assess their quality against a number of metrics. We have seen that data publishers are compliant to various degrees with the different Linked Data best practices and guidelines with regard to the quality metrics. Overall, if we consider the bigger picture, that is the aggregated conformance score, we see that on average the Linked Data quality is slightly below 60% (highest value of all metrics aggregated with a weight of 1 is 84.72%, lowest value is 41.41%) with a low standard deviation value of 7.63%. Whilst the general perception might be derived from various different factors, the aggregated results from the generic assessment shows that it might not be the case that Linked Data are necessarily of poor quality. However, there is no known literature that scales quality scores, therefore we cannot say that the assessed linked datasets are of high or medium quality. When we talk about the aggregate conformance scores, a high performing metric compensates for a lower one. Therefore, when we look at individual metrics, we see that there are certain aspects, more specifically quality metrics related to provenance and licenses, in which data publishers, collectively, should improve, as these are factors that can encourage Linked Data re-use. Nevertheless, this empirical study shows that there are still a number of problems related the Linked Data publishing and its conformance with a number of best practices and guidelines.

Our next step is to create a Linked Data Quality as a Service, that crawls the Web of Data and assesses its quality, providing quality metadata as Linked Data resources for consumption. For this we also plan to incorporate other objective and subjective metrics, however, we need to find methods that can enable the replicability of such metrics. Furthermore, we plan to assess the quality of the the LOD cloud periodically using the HDT[5] dataset LOD-a-lot[48], where the dataset contains 28 billion triples stored on 524 GB of disk space.

References

[1] H. Abelson, B. Adida, M. Linksvayer, and N. Yergler. ccREL: The Creative Commons Rights Expression Language, Mar. 2008.

[2] R. Albertoni, A. Isaac, C. Guéret, J. Debattista, D. Lee, N. Mihindukulasooriya, and A. Zaveri. Data quality vocabulary (DQV). W3C interest group note, World Wide Web Consortium (W3C), June 2015.

[3] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets with the VoID vocabulary. W3C interest group note, World Wide Web Consortium, Mar. 2011.

[4] C. B. Aranda, A. Hogan, J. Umbrich, and P. Vandenbussche. SPARQL web-querying infrastructure: Ready for action? In H. Alani, L. Kagal, A. Fokoue, P. T. Groth, C. Biemann, J. X. Parreira, L. Aroyo, N. F. Noy, C. Welty, and K. Janowicz, editors, *The Semantic Web - ISWC 2013 - 12th International Semantic Web Conference, Sydney, NSW, Australia, October 21-25, 2013, Proceedings, Part II*, volume 8219 of *Lecture Notes in Computer Science*, pages 277–293. Springer, 2013.

[5] M. Arias, J. D. Fernández, M. A. Martínez-Prieto, and C. Gutiérrez. HDT-it: Storing, Sharing and Visualizing Huge RDF Datasets. In *ISWC*, pages 23–27, 2011.

[6] A. Assaf, A. Senart, and R. Troncy. What's up LOD Cloud? observing the state of linked open data cloud metadata. In *2nd Workshop on Linked Data Quality*, 2015.

[7] J. Attard, F. Orlandi, S. Scerri, and S. Auer. A systematic review of open government data initiatives. *Government Information Quarterly*, 32(4):399–418, 2015.

[8] I. Avcibas, B. Sankur, and K. Sayood. Statistical evaluation of image quality measures. *J. Electronic Imaging*, 11(2):206–223, 2002.

[9] W. Beek, F. Ilievski, J. Debattista, S. Schlobach, and J. Wielemaker. Literally better: Analyzing and improving the quality of literals. *Semantic Web Journal*, 2017. forthcoming.

[10] S. K. Bera, S. Dutta, A. Narang, and S. Bhattacherjee. Advanced Bloom filter based algorithms for efficient approximate data de-duplication in streams. *CoRR*, 2012.

[11] T. Berners-Lee. Linked Data – Design Issues, 2006.

[12] P. V. Biron and A. Malhotra. XML Schema Part 2: Datatypes Second Edition. W3C recommendation, World Wide Web Consortium (W3C), Oct. 2004.

[13] C. Bizer, T. Heath, and T. Berners-Lee. Linked data – the story so far. *Int. J. Semantic Web Inf. Syst.*, 5(3):1–22, 2009.

[14] J. Bleiholder and F. Naumann. Data fusion. *ACM Comput. Surv.*, 41(1), Jan. 2009.

[15] C. Böhm, J. Lorey, and F. Naumann. Creating VoID descriptions for web-scale data. *J. Web Sem.*, 9(3):339–345, 2011.

[16] D. Brickley, R. Guha, and B. McBride. RDF schema 1.1. W3C recommendation, World Wide Web Consortium (W3C), February 2014.

[48] `https://datahub.io/dataset/lod-a-lot`

[17] J. Debattista, S. Auer, and C. Lange. Luzzu – a methodology and framework for linked data quality assessment. *Data and Information Quality*, 8(1), Oct. 2016.

[18] J. Debattista, C. Lange, and S. Auer. Representing dataset quality metadata using multi-dimensional views. In *Proceedings of the 10th International Conference on Semantic Systems - SEM '14*, pages 92–99, New York, New York, USA, Sept. 2014. ACM Press.

[19] J. Debattista, S. Londoño, C. Lange, and S. Auer. Quality assessment of linked datasets using probabilistic approximation. In F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux, and A. Zimmermann, editors, *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 – June 4, 2015. Proceedings*, pages 221–236, Cham, 2015. Springer International Publishing.

[20] B. Ell, D. Vrandecic, and E. P. B. Simperl. Labels in the web of data. In L. Aroyo, C. Welty, H. Alani, J. Taylor, A. Bernstein, L. Kagal, N. F. Noy, and E. Blomqvist, editors, *International Semantic Web Conference (1)*, volume 7031 of *Lecture Notes in Computer Science*, pages 162–176. Springer, 2011.

[21] A. Flemming. Quality characteristics of linked data publishing datasources. Master's thesis, Humboldt-Universität zu Berlin, Institut für Informatik, 2011.

[22] J. M. Giménez-García, H. Thakkar, and A. Zimmermann. Assessing trust with pagerank in the web of data. In E. Demidova, S. Dietze, J. Szymanski, and J. G. Breslin, editors, *Proceedings of the 3rd International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES '16) co-located with the 13th ESWC 2016 Conference, Anissaras, Greece, May 30, 2016.*, volume 1597 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[23] J. F. Hair, R. L. Tatham, R. E. Anderson, and W. Black. *Multivariate Data Analysis (5th Edition)*. Prentice Hall, 5th edition, March 1998.

[24] A. Hasnain, M. Al-Bakri, L. Costabello, Z. Cong, I. Davis, and T. Heath. Spamming in linked data. In *Proceedings of the Third International Conference on Consuming Linked Data - Volume 905*, COLD'12, pages 39–50, Aachen, Germany, Germany, 2012. CEUR-WS.org.

[25] J. A. Hausman and D. A. Wise. Stratification on Endogenous Variables and Estimation: The Gary Income Maintenance Experiment. In C. F. Manski and D. L. McFadden, editors, *Structural Analysis of Discrete Data with Econometric Applications*, chapter 10. Cambridge: MIT Press, 1981.

[26] T. Heath and C. Bizer. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, 1st edition, 2011.

[27] P. Hitzler and K. Janowicz. Linked data, big data, and the 4th paradigm. *Semantic Web*, 4(3):233–235, 2013.

[28] A. Hogan, A. Harth, A. Passant, S. Decker, and A. Polleres. Weaving the pedantic web. In *Linked Data on the Web Workshop (LDOW2010) at WWW'2010*, 2010.

[29] A. Hogan, A. Harth, and A. Polleres. SAOR: authoritative reasoning for the web. In *ASWC*, volume 5367 of *Lecture Notes in Computer Science*, pages 76–90. Springer, 2008.

[30] A. Hogan, J. Umbrich, A. Harth, R. Cyganiak, A. Polleres, and S. Decker. An empirical survey of linked data conformance. *J. Web Sem.*, 14:14–44, 2012.

[31] T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing linked data dynamics. In P. Cimiano, Óscar. Corcho, V. Presutti, L. Hollink, and S. Rudolph, editors, *ESWC*, volume 7882 of *Lecture Notes in Computer Science*, pages 213–227. Springer, 2013.

[32] H. F. Kaiser. An index of factorial simplicity. *Psychometrika*, 39(1):31–36, 1974.

[33] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, and J. Zhao. Prov-o: The prov ontology. W3c recommendation, World Wide Web Consortium (W3C), 2013.

[34] B. F. Lóscio, C. Burle, and N. Calegari. Data on the web best practices. W3C recommendation, World Wide Web Consortium, January 2017.

[35] B. F. Lóscio, E. G. Stephan, and S. Purohit. Data usage vocabulary (DUV). Technical report, World Wide Web Consortium, Dec. 2016.

[36] F. Maali, J. Erickson, and P. Archer. Data catalog vocabulary (DCAT). W3C recommendation, World Wide Web Consortium, 2014.

[37] P. N. Mendes, H. Mühleisen, and C. Bizer. Sieve: linked data quality assessment and fusion. In D. Srivastava and I. Ari, editors, *Proceedings of the 2012 Joint EDBT/ICDT Workshops*, pages 116–123. ACM, 2012.

[38] R. Meusel and H. Paulheim. Heuristics for fixing common errors in deployed schema.org microdata. In F. Gandon, M. Sabou, H. Sack, C. d'Amato, P. Cudré-Mauroux, and A. Zimmermann, editors, *The Semantic Web. Latest Advances and New Domains: 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 – June 4, 2015. Proceedings*, pages 152–168, Cham, 2015. Springer International Publishing.

[39] P. B. Nguyen, M. Luong, and A. Beghdadi. Statistical analysis of image quality metrics for watermark transparency assessment. In G. Qiu, K. M. Lam, H. Kiya, X.-Y. Xue, C.-C. J. Kuo, and M. S. Lew, editors, *Advances in Multimedia Information Processing - PCM 2010: 11th Pacific Rim Conference on Multimedia, Shanghai, China, September 21-24, 2010, Proceedings, Part I*, pages 685–696, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[40] M. Nottingham and E. Hammer-Lahav. Defining well-known uniform resource identifiers (URIs). RFC 5785 (Proposed Standard), Apr. 2010.

[41] Open Knowledge Foundation. The Open Definition.

[42] H. Paulheim and S. Hertling. Discoverability of SPARQL endpoints in linked open data. In *Proceedings of the 2013th International Conference on Posters & Demonstrations Track - Volume 1035*, ISWC-PD'13, pages 245–248, Aachen, Germany, Germany, 2013. CEUR-WS.org.

[43] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[44] S. Peroni. Media type as linked open data, 2016.

[45] K. J. Reiche and E. Höfig. Implementation of metadata quality metrics and application on public government data. In *COMPSAC Workshops*, pages 236–241. IEEE Computer Society, 2013.

[46] V. Rodriguez-Doncel, S. Villata, and A. Gomez-Perez. A dataset of RDF licenses. In R. Hoekstra, editor, *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014*, volume 271 of *Frontiers in Artificial Intelligence and Applications*, pages 187–188. IOS Press, 2014.

[47] L. Sauermann and R. Cyganiak. Cool URIs for the semantic web. W3C interest group note, World Wide Web Consortium, 2008.

[48] M. Schmachtenberg, C. Bizer, A. Jentzsch, and R. Cyganiak. Linking open data cloud diagram 2014, 2014.

[49] M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In P. Mika, T. Tudorache, A. Bernstein, C. Welty, C. A. Knoblock, D. Vrandecic, P. T. Groth, N. F. Noy, K. Janowicz, and C. A. Goble, editors, *13th Int. Semantic Web Conf.*, volume 8796 of *Lecture Notes in Computer Science*, pages 245–260. Springer, 2014.

[50] M. K. Smith, C. Welty, and D. L. McGuinness. OWL web ontology language guide. W3C recommendation, World Wide Web Consortium (W3C), February 2004.

[51] O. Suominen and C. Mader. Assessing and improving the quality of SKOS vocabularies. *J. Data Semantics*, 3(1):47–73, 2014.

[52] B. Szász, R. Fleiner, and A. Micsik. Linked data enrichment with self-unfolding URIs. In *2016 IEEE 14th International Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 305–309, Jan 2016.

[53] O. Théreaux. Common HTTP implementation problems. W3c note, World Wide Web Consortium, Jan. 2003.

[54] H. Wu, B. Villazon-Terrazas, J. Z. Pan, and J. M. Gomez-Perez. How redundant is it? - an empirical analysis on linked datasets. In *Proceedings of the 5th International Conference on Consuming Linked Data - Volume 1264*, COLD'14, pages 97–108, Aachen, Germany, 2014. CEUR-WS.org.

[55] A. J. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *Semantic Web Journal*, 2015.