Semantic Web 0 (2016) 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,*], Christoph Lange [a] and Sören Auer [a]
[a] *Enterprise Information Systems, Universität Bonn, Institut für Informatik, Römerstr. 164, 53117 Bonn Germany*
E-mail: debattis@iai.uni-bonn.de, lange@iai.uni-bonn.de, auer@iai.uni-bonn.de

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 reflects. The diversity of the Web of Data is reflected in the quality of the Web of Data. 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 ($\approx$ 5 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 published the quality metadata for each assessed dataset as Linked Data, using the Dataset Quality Vocabulary. This metadata could then be 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 will help us to 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* [35] 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 *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 [10]. The Linked Data principles, closely related to the five star scheme for publishing open data, can be summarised into *the publishing of open, linked, structured data, in non-proprietary formats, using URIs*.

This widespread and rapid adoption of the Linked Data principles has brought an unprecedented dimension 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 [25]. Document quality is often only subjectively assessable, and indirect measures such page rank and HITS (hubs and authorities), which calculate the importance of a docu-

*Corresponding author. E-mail: debattis@iai.uni-bonn.de.
[1] `https://datahub.io/group/lodcloud`

ment vis-à-vis the Web (via links), give a good indication whether a document is of good quality or a good authoritative source. In a parallel situation, resources (the data) 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 (e.g. since they are even more prone to link spamming than on the Web) but at the same time a number of other more direct quality indicators exist.

Linked Data resources are usually a 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."* – [31, §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 a number of datasets in the Linked Open Data Cloud over a number of quality indicators, as classified in [52]. Furthermore, such an investigation may lead 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 [15], a quality assessment framework for Linked Data, and a number of quality metrics (including a number of probabilistic approximation metrics), this study produces a quality metadata graph for each assessed dataset (publicly available for consumption as Linked Data resources), represented in Dataset Quality Vocabulary [16]. The benefits of these metadata graphs are twofold: (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 twenty-seven metrics related to four different quality categories as described in [52]. We then use the assessment results in order to identify the non-informative quality metrics in Section 6.2, 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 on how to improve the state, in this case of linked datasets, in order to improve, for example the overall quality. In this section, we review literature that analyses the quality of various aspects of Linked Data, as a prequel for the large-scale analysis described in this article.

In [26], 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 [28], Hogan et al. conduct 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. Our work overlaps with seven quality metrics defined in [28]: (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 [28], with some modifications as explained in Section 5. Nevertheless, the conclusions from [28] 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 metrics (as classified in [52]) on around 480 SPARQL endpoints. The authors report that only one third of the endpoints have descriptive metadata such as VoID and service descriptions[2], 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[3], 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. [5] shed light on the metadata availability in the Linked Open Data Cloud. This metadata was used in our dataset acquisition process. In [5], the metadata is checked for general, access, ownership and provenance information. The authors conclude, that metadata quality is in a bad condition. More specifically, licensing and accessibility metadata contains noisy data, thus resulting in incorrect information. We discuss the quality of LOD Cloud metadata in more detail in Section 3.

Suominen and Mader, in [48], 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: (1) labeling and documentation; (ii) structural issues (e.g. class disjoint 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.

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

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

> *... be freely used, modified, and shared by anyone for any purpose* [20].

More specifically, open data should [20, §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 i.e. using for example 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 defined in the Open Definition, whilst the last two are more related to the Linked Data principles, i.e. (4th star) the use of URIs to identify things, and (5th star) linking between the published data and external data [10].

Having metadata as part of a published dataset is the first step in putting a dataset on the open data map (thus encouraging discoverability [43]), 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 [24, §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. [6] 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.

Heath and Bizer 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] and DCAT [33]. Schemas like DCAT and VoID enable metadata description in a semantically interoperable format and can be exchanged between various agents. Currently, there are other schema initiatives such as the Dataset Quality Vocabulary (daQ) [16] and the W3C Data Quality Vocabulary (DQV) [2] to represent quality metadata for datasets, and the DUV [32] to describe various factors of a dataset such as citation and feedback from a human consumer perspective.

[2] http://www.w3.org/TR/sparql11-service-description/
[3] http://sparqles.ai.wu.ac.at/

The current LOD Cloud snapshot was taken in 2014, containing about 188 million crawled triples[4]. Metadata description of these datasets can be easily retrievable from the Linked Data catalog published together with the latest snapshot. In a recent study, Assaf et al. [5] gave an insight towards the metadata available in the Linked Open Data Cloud. The authors concluded that the quality regarding the available metadata information is in a bad condition. More specifically, licensing and accessibility metadata contained noisy data, thus resulting in incorrect information.

Whilst the CKAN API includes a metadata export functionality in terms of DCAT [33], metadata of new datasets imported to the catalog is generally manually added as textual description, thus it is prone to errors such as inconsistency and duplication. For example, in the *formats* tags, we find a variety 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* [41] should be considered in order to standardise the metadata effort between datasets within a catalog.

[4]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

3.1. LOD Cloud Datasets' Accessibility

In order to identify which datasets had some kind of access points, an initial experiment was performed on the latest LOD Cloud snapshot[5,6]. The LOD Cloud snapshot has a total of 569 datasets. Based on the metadata provided in the datahub, only around 42% (239 datasets) had a possible[7] Linked Data access point, i.e a data dump URI, SPARQL endpoint, or a VoID dataset description. From the 239 datasets, 50% of the datasets 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.

3.2. LOD Cloud Datasets' Licenses and Rights

Licences are the heart of Open Data. It is the mechanism that defines whether third parties can re-use 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 provide some kind of machine readable license on the datahub provided metadata. 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. In Table 3.2, 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[8] license since it does not sup-

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

[6]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

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

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

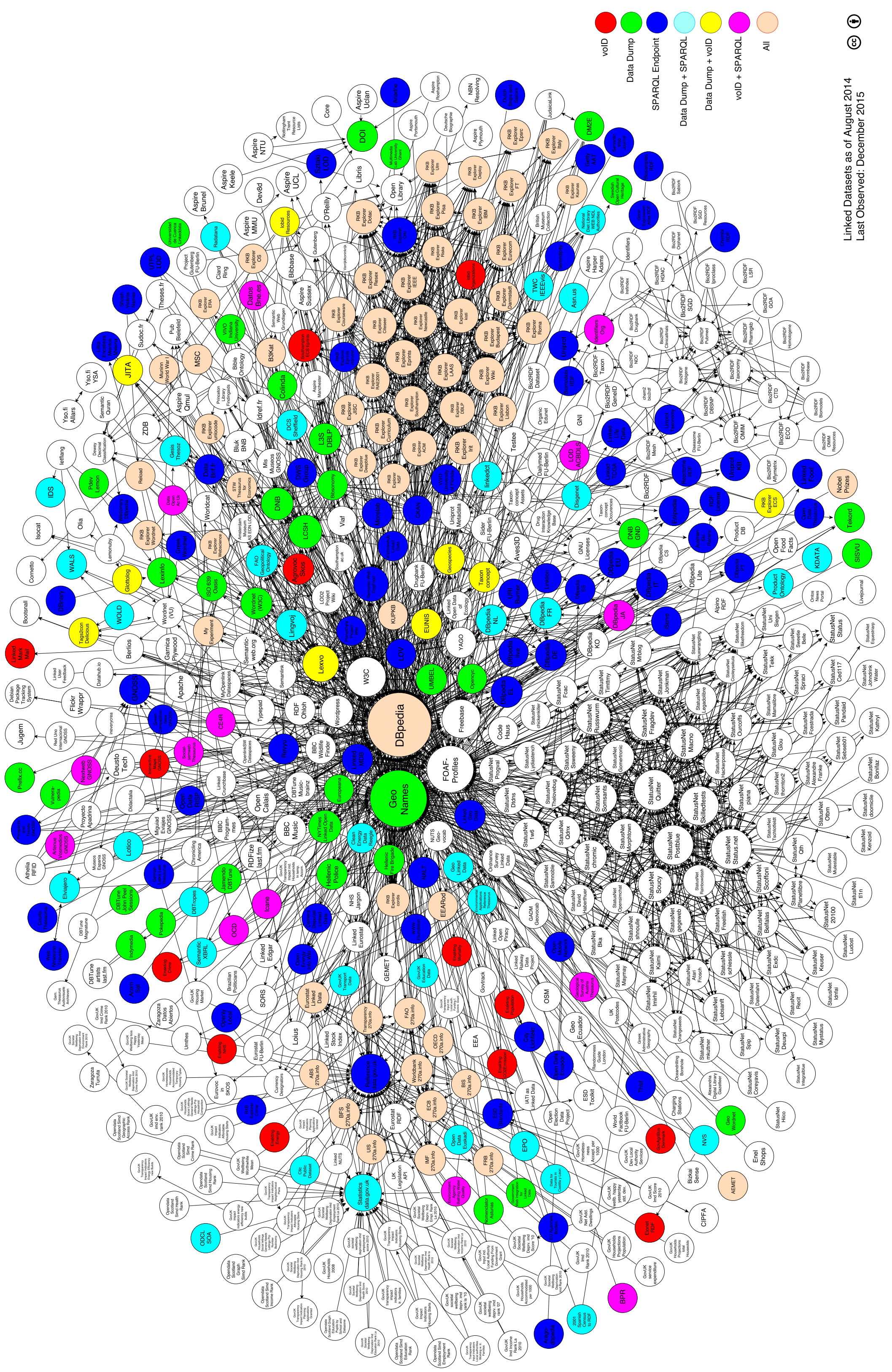

Fig. 1.: Coloring the LOD Cloud Datasets with various Access Methods (Data Dump, voID, SPARQL Endpoint, or a combination)

port some of the definition's principles, more specifically the principle that Open Data could be re-used for any purpose, including commercial purposes [20, §2.1.8]. It was noted that 7 out of 9 licenses used in the dataset's metadata were non-semantic resources (i.e. cannot be dereferenced to an RDF description). In Linked Data, publishers of such metadata should reuse RDF resources, such as Creative Commons[9] [1] and RDF License[10] [44]

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[11] 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 3.2). 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 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 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 documents themselves could also change the shape of the LOD Cloud as we know it. Such dynamics of the Web of Data are described further in [29]. Käfer et al. [29] presented the Dynamic Linked Data Observatory[12], 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 document itself. SPARQLES[13], 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[14]. Whilst this percentage is low, we noticed a small (insignificant) average change in the uptime of 0.002% between 24th February 2016 and 2nd March 2016. 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 242, therefore over a period of 6 months, 12% 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 defining 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)[15], has a corresponding generated DCAT metadata entry in the datahub.io portal. Metadata descriptions of these datasets can be easily retrieved from the catalogs Linked Data interface.

4.1. Identifying Datasets' Access Points

For this initial experiment we retrieved the distribution resources (from the property `dcat:distribution`) defined in the dataset metadata (`dcat:Dataset`), in order to identify the media types and corresponding URLs where the

[9] `https://creativecommons.org/ns`
[10] `http://purl.org/NET/rdflicense/`
[11] `.*(licensed?|copyrighte?d?).*(under|grante?d?|rights?).*`
[12] `http://swse.deri.org/dyldo/`
[13] `http://sparqles.ai.wu.ac.at/`
[14] As of 2nd March 2016
[15] 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.

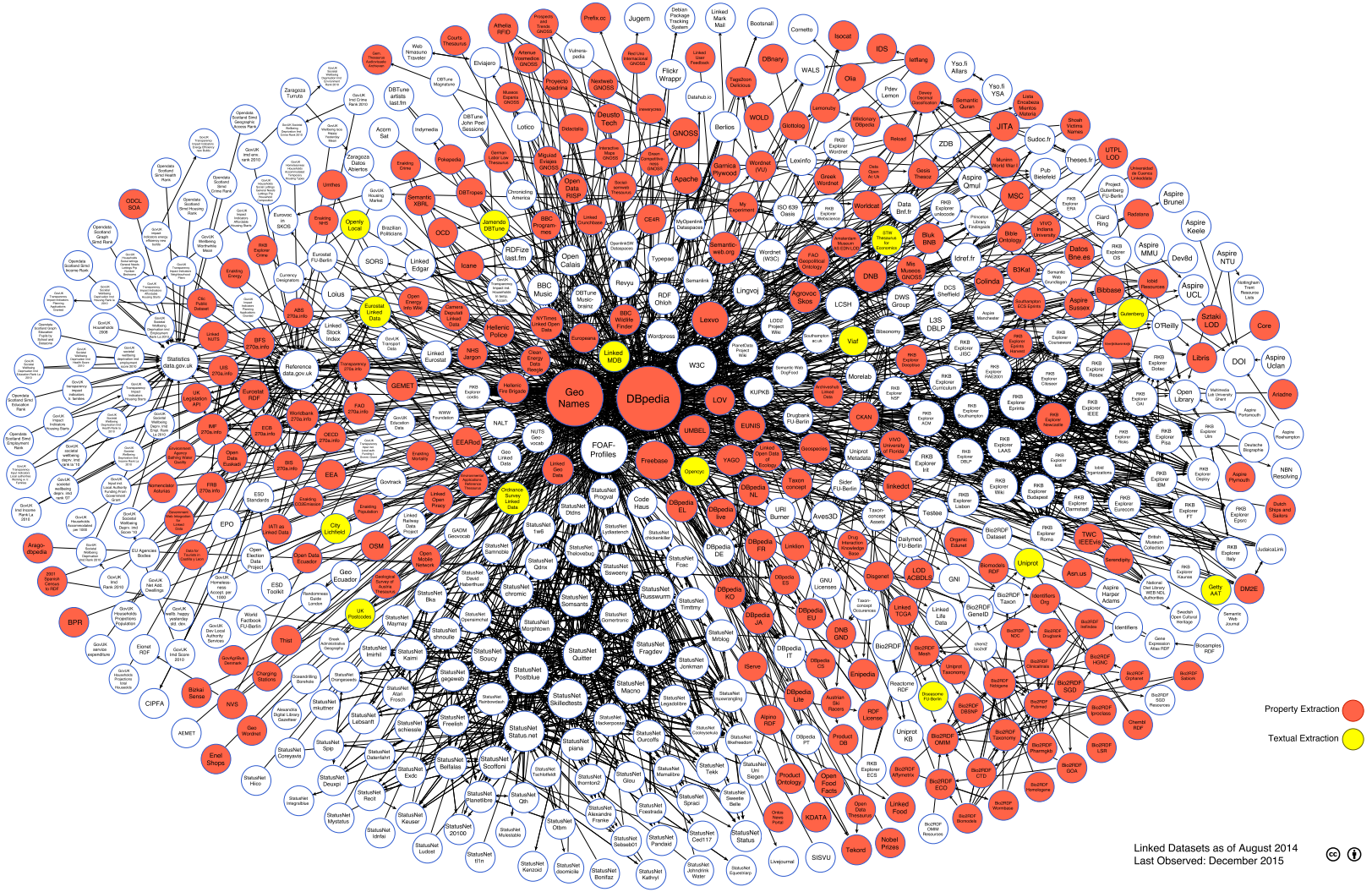

Fig. 2.: Coloring the LOD Cloud Datasets with Licence Availability extracted either via machine readable properties or using regular expressions from textual descriptions.

License Used	Type of License	URL Used	Semantic Resource	Frequency
Creative Commons Attribution License	Requires Attribution	http://www.opendefinition.org/licenses/cc-by	✗	93 (+2)
Creative Commons Attribution Share-Alike License	Requires Attribution and Share Alike	http://www.opendefinition.org/licenses/cc-by-sa	✗	47 (+1)
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/	✓	31 (+1)
Creative Commons CC Zero License	Public domain waiving all rights on the data	http://www.opendefinition.org/licenses/cc-zero	✗	30 (+1)
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	✗	5 (+1)
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)
OS Open Data License	Requires Attribution and Share Alike	-	-	(+2)
Eurostat Policy	Requires Attribution	-	-	(+1)
Project Gutenberg License	Restricts Commercial Use	-	-	(+1)
Creative Commons Attribution-NoDerivs License	Does not allow work to be re-used in derivative works	-	-	(+1)

Table 1

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

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. 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.

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 [24, §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[16].

Similarly, for SPARQL endpoints we looked at those distribution resources with a `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.

VoID descriptions were retrieved from media types containing `void` in their value. Typical media types

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

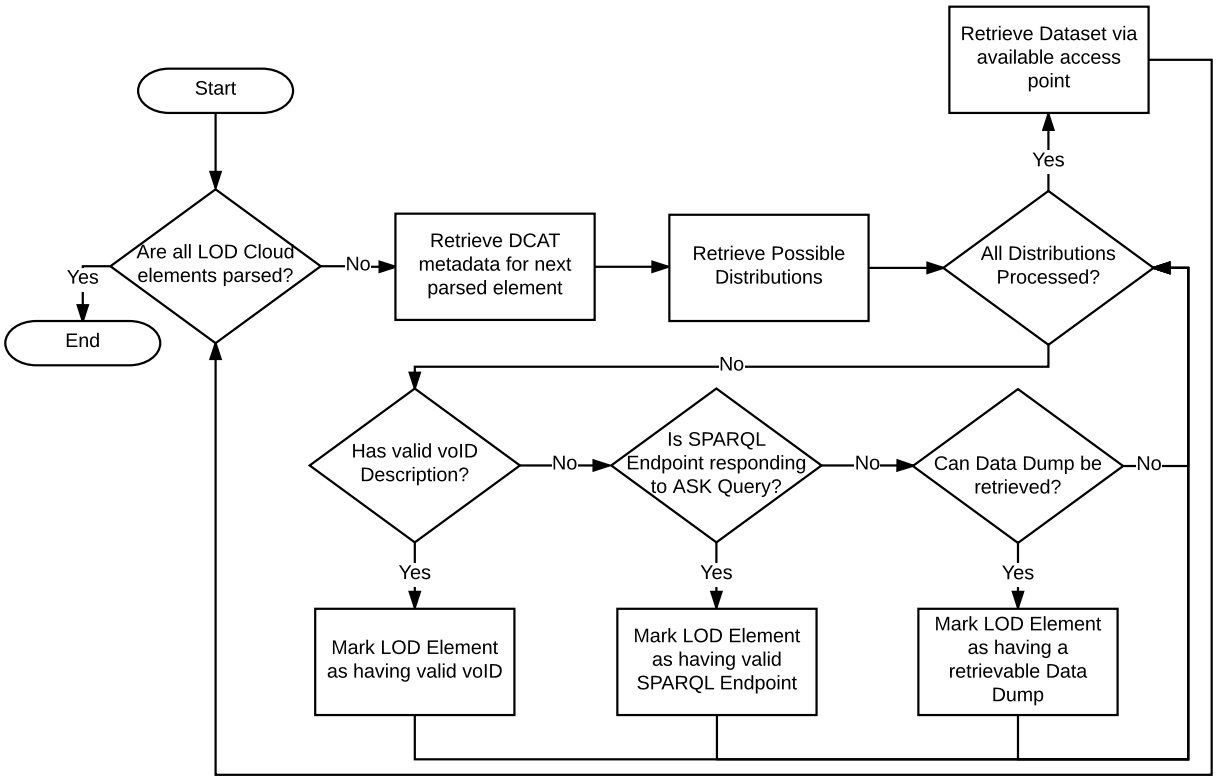

Fig. 3. A high-level flowchart depicting the marking and retrieval process of datasets from the LOD Cloud.

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 [39] 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:

- 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.

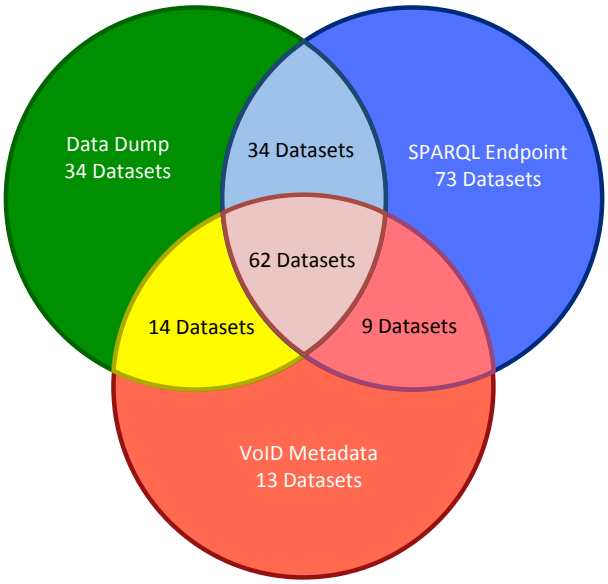

Fig. 4. A Venn Diagram illustrating a summery of the datasets' access points.

4.2. Datasets' Summary

During the acquisition process, we identified 239 accessible datasets – i.e. data dump, SPARQL endpoint, or VoID dataset:

- 13 datasets having only accessible VoID metadata;
- 34 datasets with only a data dump corresponding to the pre-defined media-types;
- 73 datasets with only a SPARQL endpoint;
- 34 datasets having both a data dump and a SPARQL endpoint;
- 9 datasets having both VoID metadata and a SPARQL endpoint;
- 14 datasets having both VoID metadata and a data dump;
- 62 datasets having all three possible access points.

Figure 4 illustrates this summery in a Venn diagram. The data dumps and SPARQL endpoints overall comprised approximately 5 billion quads.

5. Quality Assessment

In this study we are mainly interested in understanding the persistent quality issues within the LOD datasets, rather than the performance of the quality metrics. Furthermore, this study complements the work undertaken in the survey by Zaveri et al. [52] and the work that 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 inferencing or dereferenceability of external resources is done, unless required by the quality metric. For each data quality metric 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.

5.1. Choice of Data Quality Metrics

In this empirical study we assess the datasets against 27 quality metrics described in [52], and two additional quality metrics describing provenance information. 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. Furthermore, with this subjective metric we show that the Luzzu framework can handle both subjective and objective metrics, as described in Pipino et al. in [42].

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, users can then use the generated quality metadata to rank and filter datasets based on their choice of metrics.

The choice of generic quality metrics was based solely on the classification in [52]. 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. 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. [52] categorised a number of metrics in this category within the four dimensions *Representational Conciseness*, *Interoperability*, *Interpretability* and *Versatility*.

(RC1) Keeping URIs short

Classified in the representational-conciseness dimension, this metric observes the length of URIs. In the Cool URIs document [45], the editors remarked 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. [28] 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 [49, §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 having 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 on dataset D (i.e. the dataset being assessed). A data-level constant is defined by [28] 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.

Discussion: A box plot with the quality values is illustrated in Figure 5. The box plot for this metric (RC1) is comparatively tall, suggesting that publishers tend to have quite different inclinations on how long the URI identifiers should be. The sample over the LOD Cloud is centred on 97.92% with a sample standard deviation (σ_s) value of 24.90%. A number of outliers (around 13% of the datasets), were detected. These are datasets that scored lower than 54.55% (i.e the lower whisker). The range of quality values, including outliers, is 99.75%, whilst the range between quartile group 4 and quartile group 1 is 45.45%. We also notice that the population is skewed to the left (i.e the median is closer to the third quartile), whilst the bottom whisker is longer than the top whisker (since we cannot have a value greater than 100%), suggesting that most quality values are large with some smaller values. The average 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 where 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 fine, while longer ones are deemed to be "bad". There might be various reasons for publishers to use longer URIs. For example, URIs can comprise some structure, such as a directory scheme. In order to avoid the discriminatory power problem, Hogan et al. [28, §5.1 – Issue IV] calculate the metric value based on the average length of the URIs in a dataset, promoting those datasets that have short URIs. In our case the metric is more flexible with regard to the typical length of URIs used in datasets. Furthermore, we deem that publishers who use domain names with various levels (e.g. typical university URIs) and still adhere to the recommendation should not be given a lower quality value.

(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[17] in SPARQL 1.1 allows the retrieval of all mem-

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

bers 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 [24, §2.4.1.2], the author discourage the use of RDF reification since they "are 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 [28, §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 [28, §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. Therefore, the metric value is a ratio of quads in a dataset with and without discouraged RDF data structures.

Discussion: Similar to the findings of Hogan et al. [28], 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 average quality value of this metric is 99.44% and the calculated σ_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 1 σ_s (i.e. having a quality value of at least 96.369%). Similar to the Short URIs metric (RC1), a relatively small number of outliers (around 12% of the datasets) were detected. Nonetheless, the dataset with the lowest quality value for this metric (`http://bibsonomy.org`) is 70.25%. Upon further inspection of this dataset, we found that the publisher used `rdf:Seq` and `rdf:Bag` in order to list information such as editors and authors of some publication. 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 [11] argues 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) portal[18]. Together with W3C recommendation vocabularies such as *SKOS*, schemas such as *FOAF*, *Dublin Core*, and *SIOC*, amongst others, have become defacto standards with more than 15% of the LOD datasets using at least one of these vocabularies [46]. Furthermore, the W3C is striving to create standardised cross-domain vocabularies, such as *DCAT* and *PROV-O* amongst others. Zaveri et al. [52] classify this metric under the interoperability dimension, and focus on the overlap between the dataset in question and its overlap with recommended vocabularies [28, §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 the 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 (e.g. DBTropes is tagged with the label *media*). During the initialisation of the metric, the LOV API[19] is invoked to obtain the vocabularies available with the respective tag. Further-

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

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

more, based on the usage study conducted in [46], we included the following vocabularies by default for all datasets: RDF, RDFS, FOAF, DCTerms, OWL, GEO, SIOC, SKOS, VOID, DCAT.

We identify overlapping classes and properties in the same manner as defined in [28, §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:

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

$$\overline{cl_{exs}} := \{x \mid x \in \overline{v_c} \wedge x \in class(D)\}$$

$$\overline{pr_{exs}} := \{y \mid y \in \overline{v_p} \wedge y \in prop(D)\}$$

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. Therefore, the metric value is a ratio of the number of external terms (classes and properties) vs. the number of terms used in the dataset.

Discussion: The box plot for this metric (IO1 - Figure 5) is comparatively long and skewed to the left, 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 on a value of 24.00% with a sample standard deviation (σ_s) value of 29.10%. More concerning is the mean value of 34.01%, indicating the 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. [28, §5.3 – Issue IX], where the authors also suggest that the amount of re-used terms and properties in their sample is widely distributed (the σ value in their experiment is 29.05).

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. This metric does not consider user-defined terms with links to *existing* terms using predicates such as `owl:sameAs`, `owl:equivalentClass`, or `owl:equivalentProperty`, as being a valid re-used existing term as described in this metric.

In order to improve schema re-use, services such as LOV and Swoogle[20] should be used to find suitable schemas. On the other side, vocabulary curators should maintain and promote their schemas, for example, by making sure that vocabularies are properly dereferenceable.

(IN3) Usage of Undefined Classes and Properties

The invalid usage of undefined classes and properties metric is classified under the interpretability dimension [52], which targets the technical representation of the data itself. Using classes and properties without a formal definition (i.e. not defined in a schema) is undesirable, as agents would not be able to understand how the data should be interpreted, for example, during reasoning. Errors that lead to such invalid usage includes: capitalization errors (e.g. `foaf:person` vs. `foaf:Person`), syntactic errors (e.g. `foaf:img` vs. `foaf:image`), and dereferencability issues [17, §4] with external schemas (e.g. schema not available anymore, or not in machine-readable format).

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

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

$$\overline{cl_{undef}} := \{x \in V_c \mid \exists V \cdot ns(x) \mapsto V \wedge x \in class(D)\}$$

$$\overline{pr_{undef}} := \{y \in V_p \mid \exists V \cdot ns(y) \mapsto V \wedge y \in prop(D)\}$$

where V_c is the set of classes (where a class is defined as being of type `rdfs:Class` or `owl:Class`) in a vocabulary V which is resolved by an agent using the namespace of the term[21] x ($ns(x)$). Similarly, V_p is the set of properties (where a property is defined as being of type `rdf:Property`, `owl:ObjectProperty`,

[20] http://swoogle.umbc.edu

[21] In cases where slash URIs are used, the namespace does not necessarily resolve the schema, therefore the term is used to resolve the term's description.

`owl:DatatypeProperty`, `owl:AnnotationProperty`, or `owl:OntologyProperty`) in vocabulary V. Therefore, the metric value shows how much of a dataset uses classes and properties that are formally defined. In order to check if a class or property (term) is defined, the term is dereferenced for its semantic description and queried for properties and classes. If a term is non-dereferenceble, it is considered undefined.

Discussion: The box plot for this quality metric (IN3 - Figure 5) is relatively tall, covering 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 (median) at 53.33% with a σ_s value of 32.18%, whilst the average quality value is 54.48%. Although the box plot might seem symmetrical, the values are skewed to the right by a small margin ($\approx$ 5%).

A higher value means that less undefined terms were used in the 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. We noticed that apart from undefined terms, publishers use terms that were wrongly defined, 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"* [24].

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. In SPARQL, blank nodes behaviour is unpredictable in *RDF equivalent graphs*, whilst they cannot be referenced during querying [34].

Metric Computation: This metric assesses the usage of blank nodes within the subjects and objects. The metric value is assessed as suggested in [28, §5.1 – Issue I]:

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

Dataset	V1(D)
zbw.eu/stw	4
linkedmarkmail.wikier.org	3
nhs.psi.enakting.org	3
population.psi.enakting.org	3
crime.psi.enakting.org	3
...	
vocab.nerc.ac.uk	0
wals.info	0
www.productontology.org	0
bfs.270a.info	0
cordis.rkbexplorer.com	0

Table 2

Top and Bottom 5 Ranked Datasets for the Different Serialisation Formats Metric.

Dataset	V2(D)
nhs.psi.enakting.org	15
population.psi.enakting.org	15
crime.psi.enakting.org	15
co2emission.psi.enakting.org	15
rdfdata.eionet.europa.eu	13
...	
education.data.gov.uk	1
vocab.nerc.ac.uk	1
wals.info	1
www.productontology.org	1
cordis.rkbexplorer.com	1

Table 3

Top and Bottom 5 Ranked Datasets for the Multiple Language Usage Metric.

where $dlc(D)$ is the set of data-level constants in dataset D and $bn(D)$ is the set of blank nodes in D. The value represents the degree of **avoiding** the usage of blank nodes.

Discussion: The box plot (IN4) illustrated in Figure 5, is relatively short, suggesting that most data publishers agree to avoid blank nodes. Furthermore, the box plot range is 5.07%, with the upper whisker and third quartile at the 100% mark. The sample centrality is 100% and the σ_s is 12.15%. The higher the value, the less blank nodes are used in a dataset. The average quality metric value 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 are a number of datasets marked as outliers consequently stretching the σ_s value. In particu-

lar, the `prefix.cc` dataset uses blank nodes in almost every triple. This dataset affected the σ_s value significantly, which otherwise would be considerably lower than in [28, §5.1 – Issue I]. One should note that the corpus in [28] contained FOAF profiles, which traditionally contain many blank nodes. In certain situations, the usage of blank nodes is complementary 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, Quads, and JSON-LD. For example, Web applications prefer the JSON-LD format, rather 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 [24, §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 [52].

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.

According to the VoID W3C recommendation, the `void:feature` *"can be used for expressing certain technical features of a dataset, such as its supported RDF serialisation formats"*. [3, §2.6] Data publishers can serialise their data in up to 23 different formats[22]. The metric can be quantified as follows:

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

where $features(D)$ is the set of dataset features identified by the object in a triple **subject** $\times$ **void:feature** $\times$ **object**. Therefore, this metric returns a value indicating the number of supported serialisation formats.

Discussion: Table 2 shows the five top and bottom ranked datasets[23], according to the number of serialisation formats defined. In most cases, the publishers did not define any serialisation format in the metadata of their datasets. Only nine datasets had a serialisation format following our guideline. The σ_s value is 0.71 whilst the mean value is 0.18.

A dataset, serialised in different formats, widens possible uses in different scenarios. In order to encourage multiple format serialisation, tools such as *Raptor*[24] or *Serd*[25] provide command line functions that transform (bulk) data into various serialisations. One drawback is that using different serialisations takes up more storage resources. Regarding the generation of VoID metadata, generators such as [13], 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 used more often. 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."* – [31]

The usage of multiple languages metric is also classified by Zaveri et al. under the versatility dimension [52].

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:

$$V2(D) := round\left(\frac{size(lt = \{o \in lit(D) \mid hasLangTag(o)\})}{size(\overline{lt})}\right)$$

where lt is a set of literals with a language tag in dataset D, and $\overline{lt}$ is the set of unique literals with a language tag. This metric value will return a natural rounded number of languages that characterise the assessed dataset.

Discussion: Table 3 shows the datasets that have a high number of multi-lingual textual labels. In most

[22] `http://www.w3.org/ns/formats/`
[23] No ties were resolved.
[24] `http://librdf.org/raptor/`
[25] `https://drobilla.net/software/serd/`

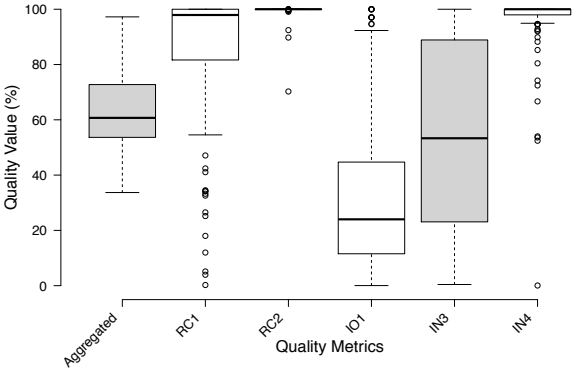

Fig. 5. Representational category box plot excluding Different Serialisation Formats and Usage of Multiple Languages metric are excluded, but included in the aggregated result. Outliers are represented by dots.

cases, publishers describe their textual literals using only one language ($\approx$ 83%). One possible reason is that publishers target a particular audience, or do not have the resources to create multilingual datasets. The mean value for this metric is 1.72 languages, whilst the standard deviation (σ_s) is 2.71 (median value: 1). The σ_s value shows that publishers are inclined towards supporting a lower number of languages.

Language tags allow agents to express linguistic or text-based information better, for example, providing better localisation. There are publishers who refrain from adding a language tag to the textual literals. Tools such as *Apache Tika*[26] detect the language of literals and can help publishers to add the correct tags.

Aggregated Results

Table 4 shows the aggregated ranking (top and bottom five datasets) of datasets in the representational dimension, as described in Section 5.6. 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, whilst the rest are float values). The overall aggregated box plot shows a population which is slightly skewed to the left, close to symmetrical (since the mean and median values are close), with a centrality median of 60.70%. This suggest that there is more variety amongst higher quality values (i.e. more than the median) amongst the sample. Nevertheless, the deviation value (σ_s) is 14.50%, which suggests a moderate distribution, whilst the average score is 63.60%.

[26] `http://tika.apache.org/`

5.3. Contextual Category

According to Zaveri et al. [52], 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-Ãă-vis the task at hand, (ii) degree of the 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 the resources within the dataset itself*. The provenance metrics can be seen similar to those classified under the *trustworthiness* dimension. Furthermore, in this category, we only tackle three metrics related to *understandability*, whilst no metrics classified under the *relevancy* and *timeliness* metrics are assessed.

(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."* – [31, §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 could 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:

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

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

Dataset	$v(C, 1.0)$	RC1	RC2	IO1	IN3	IN4	V1	V2
http://co2emission.psi.enakting.org/	97.24%	91.43%	100.00%	97.06%	97.06%	94.59%	3	15
http://crime.psi.enakting.org/	97.06%	91.43%	100.00%	97.06%	97.06%	94.59%	3	15
http://nhs.psi.enakting.org/	96.88%	91.43%	100.00%	97.06%	97.06%	94.59%	3	15
http://population.psi.enakting.org/	96.60%	91.43%	100.00%	97.06%	97.06%	94.59%	3	15
http://thesaurus.iia.cnr.it/	95.53%	100.00%	100.00%	100.00%	100.00%	100.00%	0	2
		...						
http://lod.taxonconcept.org/	43.22%	67.46%	100.00%	3.42%	9.99%	99.44%	0	1
http://wals.info/	42.66%	90.83%	100.00%	5.14%	5.14%	100.00%	0	1
http://vocabulary.semantic-web.at/PoolParty/wiki/semweb	42.49%	93.96%	100.00%	10.00%	15.30%	98.97%	0	1
http://sw.opencyc.org/	37.14%	85.10%	99.39%	0.37%	0.42%	98.36%	0	1
http://minsky.gsi.dit.upm.es/	33.69%	11.98%	100.00%	2.88%	4.33%	100.00%	0	1

Table 4
Overall ranking of datasets for the representational category.

'1' if $d \in \overline{ds}$ has a triple corresponding to **subject** $\times$ (**dc:creator** $\|$ **dc:publisher**) $\times$ **object**.

Results Overview: 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, qualitative speaking, very negatively short, suggesting that most of the sampled datasets contain no basic provenance information in their VoID or DCAT metadata (when available). The median value is 0%. Nonetheless, this metric has a number of outliers, amounting to around 16.27% of the sample population. From this 16.27%, 71% of the datasets have a quality value of 100%. The σ_s value stands around 32.89%, whilst the mean is 12.78%.

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"* [31], 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 rather to some metadata. Hence, we encourage publishers to use dataset profiles for VoID and DCAT, such as DCAT-AP [27] and VoID Editor[28].

(P2) Traceability of the Data

In the Data on the Web Best Practices document, 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"* – [31, §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 [23]. 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 represents some aspect which is being modelled in form of linked data, for example weather information from Malta. An activity describes the process of creating Linked Data resources.

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 [52], this metric can be related to the "trustworthiness of statements (T1)". 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

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

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

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. The metric's value represents the ratio of the dataset's resources conformity to this metric.

Results Overview: Similar to Metric P1, this metric (P2 - Figure 6) is also very negatively short. Unlike Metric P1, the granularity level of the metadata in this case 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) who creates such metadata to enable users to identify the origin of data. The overall median value is 0%. The σ_s value 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 the ambiguity of 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 [18]. 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 'Malta' from `http://dbpedia.org/resource/Malta`.

Heath and Bizer suggest that predicates such as `rdfs:label`, `foaf:name`, `skos:prefLabel`, `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 [24]. Nevertheless, there are a number of vocabularies having terms to describe human readable labels and comments[29]. The authors in [18] study the usage of labels in the Web of Data[30], and reported the occurrence of the various predicates used for resource labelling. In terms of classification, according to [52] this metric is classified under the *understandability* dimension.

Metric Computation: The aim of this metric is to calculate a dataset 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 [46]:

- `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` (for FOAF profiles).

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:

$$U1(D) := \frac{size(\{t \mid (\forall t \in \overline{ent} \cdot t.predicate \in desc)\})}{size(\overline{ent})}$$

where $\overline{ent}$ is the set of resources (i.e. triples with the same subject URI) in the assessed dataset D, t is a

[29] 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).

[30] The corpus used was the BTC2010 (`http://challenge.semanticweb.org/`)

triple in $\overline{ent}$, and *desc* is the set of predefined predicates that define a label or description. The metric's value represents the level of completeness of a dataset with regard to human-readable labels and descriptions.

Results Overview: The box plot for this quality metric (U1 - Figure 6) is relatively tall, covering the whole range of values, i.e. 100%. This suggests that data publishers follow varying practices with regard to human-readable labels and comments. The quality value is centred on 33.33% with a σ_s value of 40.93%, whilst the average quality value is 43.76%. The quality values of this metric provides the biggest variance against the rest of the contextual metrics. 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. [28, §5.3 – Issue XI]. Our assessment shows larger variation (σ_s value in [28, §5.3 – Issue XI] was 14.99%) in the quality result, the average value in our study increased by 28.76% when compared with the previous study conducted in 2012.

Whilst most of the publishers tend to attach labels and descriptions, other publishers might use other non de-facto schemas to describe resources in a human readable fashion. Overall, we can draw parallels between our assessment results and the results presented in [18], 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 from URI resource such as local name, or query a dataset retrieving local resources according to the specified URI structure. The presence of URI regular expression metric is classified under the *understandability* dimension [52].

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

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

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

Results Overview: This metric reports 100% if the assessed dataset has a URI regular expression pattern defined. Our assessment showed that only 10 of the datasets had such an expression, giving a mean value of 7.75%, and a σ_s value of 26.84%. The box plot for the metric U3 in Figure 6, illustrates this negatively short 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 [52].

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:

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

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 classes ($class(D)$) and properties ($prop(D)$).

Results Overview: Similar to most of the contextual metrics, the box plot for this metric (U5) is, also negatively very short, 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 σ_s value to 10.62%, whilst the mean is 2.71%.

From our assessment, around 2,800 different (not unique) vocabularies were used throughout the assessed dataset, whilst only 128 (around 4%) vocabularies where identified by the `void:vocabulary`

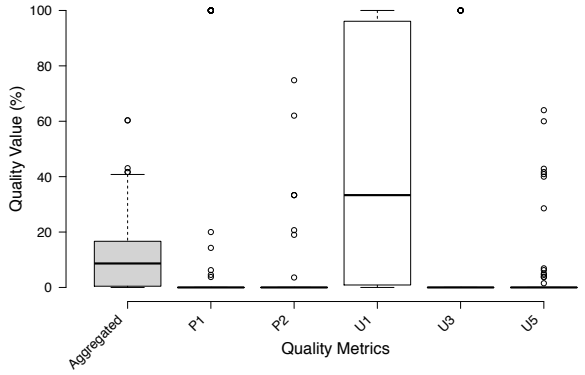

Fig. 6. Contextual category box plot. Outliers are represented by dots.

predicate. Moreover, only 63 of those 128 defined vocabularies (around 63%, and around 2% of the total number of vocabularies used) where 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

Table 5 shows the aggregated ranking of the five top and bottom datasets per category. 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 population that is symmetrical, with a centrality median of 8.66%. The deviation value (σ_s) is 13.84%, which suggests a moderate distribution, whilst the average 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 worryingly 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"* [52], the intrinsic category quality indicators are related to assess *correctness* and *coherence* of the data. Zaveri et al. [52] classified metrics according to the following dimensions:

1. *syntactic validity* – the conformance of an RDF document vis-à-vis the standard specification;
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 metrics related to the *conciseness* dimension (the extensional conciseness metric), the *consistency* dimension metrics (seven in total), and one metric from the *syntactic validity* dimension. No metrics were assessed for the other two dimensions mentioned in [52], as they would have required a different experiment set up. For example, for the *completeness* dimensions, we would require to assess the datasets according to their domain.

(CN2) Extensional Conciseness Metric

In [12], Bleiholder and Naumann define a 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. However, data redundancy can be challenging in terms of data curation. For example, a data curator has to ensuring that all "replicated" resources are updated accordingly. However, 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 [51].

At the Linked Data level, a linked dataset is concise if there are no redundant instances [36]. By redundancy, Mendes et al. [36] explains that there are no two instances (locally) 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 [52].

Metric Computation: The extensional conciseness metric checks for redundant resources in the assessed dataset, and thus measures the number of unique in-

Dataset	$v(C, 1.0)$	P1	P2	U1	U3	U5
http://bfs.270a.info/	60.35%	100%	74.80%	99.91%	0%	0%
http://lod.geospecies.org	60.29%	99.98%	0%	47.82%	100%	64%
http://statistics.data.gov.uk/	43.05%	0%	0%	98.33%	100%	60%
http://www.kupkb.org/	41.66%	100%	0%	100%	0%	0%
http://rod.eionet.europa.eu/	41.66%	100%	0%	100%	0%	0%
...						
http://curriculum.rkbexplorer.com	0%	0%	0%	0%	0%	0%
http://extbi.lab.aau.dk/resource/Dataset	0%	0%	0%	0%	0%	0%
http://data.dcs.shef.ac.uk/	0%	0%	0%	0%	0%	0%
http://prefix.cc/	0%	0%	0%	0%	0%	0%
http://id.ndl.go.jp/auth/ndla	0%	0%	0%	0%	0%	0%

Table 5

Overall ranking of datasets for the contextual category.

stances found in the dataset. In [17, §5.2], we showed that a naïve implementation of this metric results in large computational time, therefore we suggested the use of Bloom Filters [9] as an approximation technique. Using the bloom filter for identifying possible duplicate instances during the assessment process, we quantify this metric as:

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

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

where, *hash* is a function that hashes the resource and *isSet* is the function that checks if the produced hash is already contained in the filter. *r* is a resource whose hash bits might have been set before, thus indicating a possible duplicate resource. In simple terms, the value returned by this quality metric describes the dataset's level of non-redundant entities. Further discussion on Bloom Filters and how can this metric be approximated can be found in our previous publication [17, §5.2].

Results Overview: Our assessment estimated that the assessed datasets had an average of 7.6% redundancy (the mean value is 92.40%) in total. Nevertheless, this does not mean that there is low redundancy on the whole Web of Data, since the sample standard deviation (σ_s) stands at 13.22% (median 99.34%), which suggests a moderately varied quality value overall. Although the box plot (see Figure 7) for this metric (CN3) is comparatively short, the outliers stretch the σ_s value. 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, ensuring efficient runtime with a low loss in precision (cf. [17, §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 the traditional 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"* [47, §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 `John` (as an example) cannot be both a person and a document. This metric is classified under the consistency dimension in [52].

Metric Computation: Metric CS1 checks for disjointness between types in multi-typed resources. Moreover, each assessed explicit type is inferred in order to check disjointness also between parent classes.

Along these lines we quantify this metric as follows:

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

$$hasDisjointTypes(r) := \begin{cases} 1.0 & size(\overline{r_{dis}}) > 0 \\ 0.0 & otherwise \end{cases}$$

$$\overline{r_{dis}} := \{(pInf(t_i) \setminus pInf(t_j)) \mid \forall t \in types(r)\}$$

where $pInf(t)$ is the set containing the disjoint members of t ($\mathbf{t} \times$ **owl:disjointWith** $\times \bar{\mathbf{t}}$) and the disjoint members of the parent members of t ($\mathbf{t} \times$ **rdfs:subClassOf*** $\times \dot{\mathbf{t}}$), and $types(r)$ is the set of the types a resource is a member of ($\mathbf{r} \times$ **rdf:type** $\times \mathbf{t}$). The metric value indicates the degree of disjoint entities used within resources in the assessed dataset.

Results Overview: The assessment shows that 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 quality 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 average quality value for this metric is 100%, whilst the standard sample deviation (σ_s) is 0% (median is 100). The box plot CS1 in Figure 7 shows that there is no variation in the quality value of this metric, with the quartile ranges having the same value. Such low values in OWL inconsistencies were also reported in [26], were the authors attribute inconsistency problems caused by various incompatible exporters, such as FOAF exporters.

(CS2) Misplaced Classes or Properties Metric

RDF Schema provides property-centric mechanisms for defining classes (`rdfs:Class`) and properties (`rdf:Property`) in vocabularies [14]. 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."* - [14, §2]

OWL has its own class axiom (`owl:Class`), which implicitly is a subclass of its RDF Schema counterpart. The schema has specialised property axioms (`owl:DatatypeProperty` and `owl:ObjectProperty` amongst others) that extend the `rdf:Property`, in order to (1) distinguish between the supposed values of the property, (2) enforce property-value constraints, and (3) describe logical characteristics of a property (cf. Metric CS3).

The RDF data model is represented by a *triple form* (**subject** $\times$ **predicate** $\times$ **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 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 it 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. Nonetheless, there are two OWL axioms, `owl:equivalentProperty` and `owl:inverseOf` that require a property to be in the *object* position. Therefore, triples with these two properties as predicates should be excluded from the assessment. This metric is classified under the consistency dimension in [52].

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:

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

$$\overline{c_{misp}} := \{c \mid \forall c \in class(D) \cdot c \in V_p\}$$

$$\overline{p_{misp}} := \{p \mid \forall p \in prop(D) \cdot p \in V_c\}$$

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

Results Overview: 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 average is 99.99%. The

range value is 0.09%. 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 this metric CS2.

(CS3) Misused OWL Datatype or Object Properties Metric

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 properties or not. Zaveri et al. [52] 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:

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

$$misusedOWL(t) :=$$
$$(isLiteral(t.object) \wedge isOP(t.predicate)) \vee$$
$$(isIndividual(t.object) \wedge isDP(t.predicate))$$

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 is an `owl:ObjectProperty` or `owl:DatatypeProperty` respectively. A high value of this metric indicates a low amount of (or no) misused properties.

Results Overview: Figure 7 shows the box plot for this metric CS3. Similar to the previously discussed metrics for this dimension, the datasets adhere to a high quality score (average 98.88%) and a considerably low deviation (σ_s) 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, having a value lower than 100%, which is less than the box plot lower whisker.

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 Metric

Removing classes and properties from schemas renders data using these incoherent. OWL introduces the two classes `owl:DeprecatedClass` and `owl:DeprecatedProperty` for such situations. These two properties indicate that a class or property that belongs to these, are no longer recommended to be used in published data. This metric is classified under the consistency dimension in [52].

Metric Computation: This metric assesses a dataset to check if depreciated terms are used. More specifically, all used classes and properties are checked if

they are members of `owl:DeprecatedClass` or `owl:DeprecatedProperty` respectively:

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

$$\overline{c_{dep}} := \{c \mid \forall c \in class(D) \cdot c \in V_c \wedge dc(c)\}$$

$$\overline{p_{dep}} := \{p \mid \forall p \in prop(D) \cdot p \in V_p \wedge dp(p)\}$$

where $dc(c)$ is a function that returns true if the class c member of (**c** $\times$ **rdf:type** $\times$ **owl:DeprecatedClass**), whilst $dp(p)$ is a function that returns true if the property p member of (**c** $\times$ **rdf:type** $\times$ **owl:DeprecatedProperty**). The metric's value calculates the ratio of used deprecated classes and properties against all used classes and properties.

Results Overview: 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 deviation (σ_s), as in the other consistency metrics, is very low (0.23%) with the median being 100. The overall average is 99.97%.

(CS5) Valid Usage of the Inverse Functional Property Metric

In the real world, an encryption public key 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 and 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[31] these two resources are treated as the same. The OWL schema provides a term `owl:InverseFunctionalProperty`, in which a vocabulary property with the above described semantics should be member of. Common examples of such properties include `foaf:mbox` and `foaf:homepage`. This metric is classified under the consistency dimension in [52].

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. In this metric, we only consider those statements with an inverse functional property. We quantify this metric as follows:

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

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

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

$$\varphi(t, \bar{t}) := \begin{cases} true & if\ ((t.sb \neq \bar{t}.sb) \wedge \\ & (t.pr = \bar{t}.pr) \wedge \\ & (t.ob = \bar{t}.ob)) \\ false & otherwise \end{cases}$$

where $ifp(t.predicate)$ is a function that checks if a term is a member of (**t.predicate** $\times$ **rdf:type** $\times$ **owl:InverseFunctionalProperty**), $\varphi(t, \bar{t})$ is a function that returns true if a triple t and a previously seen triple $\bar{t}$ are violating the IFP functionality. Therefore, the metric value is a ratio between the number of violating IFP triples, against the number of statements having an IFP predicate.

Results Overview: The box plot for this metric (CS5) in Figure 7 shows the trend in this metric where a large part of the assessed datasets are 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 σ_s value to 12.29%, whilst the calculated median is 100%.

One should keep in mind that not all datasets assessed made use of inverse functional properties and were 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 singled out in the assessment:

- `http://xmlns.com/foaf/0.1/homepage` (violated in 2861 triples)
- `http://rdf.myexperiment.org/ontologies/base/has-friendship` (violated in 635 triples)

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

- `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)

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 Metric

In [27], the term ontology hijacking was described as the *"re-definition or extension of a definition of a legacy concept [...] in a non-authoritative source"*. An authoritative source s for concept c means that the namespace of c coincides with that of s. In simple terms, `http://xmlns.com/foaf/0.1/` is the authoritative source for the concept `foaf:Person`. Nevertheless, ontology hijacking can be seen as restricting the Linked Data idea of open world assumption, in a sense that such terms restricts what one can say about some concept. On the other hand, ontology hijacking may lead to incorrect inferencing throughout the data [27]. Zaveri et al. [52] classifies 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 defining a class or property, following the axioms identified in [27]. The hijacking rules (axioms - triple position for authoritative document) are:

- `rdfs:subClassOf` - subject;
- `owl:equivalentClass` - subject or object;
- `rdfs:subPropertyOf` - subject;
- `owl:equivalentProperty` - subject or object;
- `owl:inverseOf` - subject or object;
- `rdfs:domain` - subject;
- `rdfs:range` - subject;
- `owl:SymmetricProperty` - subject;
- `owl:onProperty` - object;
- `owl:hasValue` - subject;
- `owl:unionOf` - object;
- `owl:intersectionOf` - subject or object;
- `owl:FunctionalProperty` - subject;
- `owl:InverseFunctionalProperty` - subject;
- `owl:TransitiveProperty` - subject.

This metric analyse defined classes and properties in a dataset by checking if these definitions are violating the hijacking rules. Along these lines, we quantify the metric as follows:

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

where $tdef(D)$ is the set of triples in dataset D having one of the hijacking rules axioms in its predicate or object position, and $\mathcal{H}(t)$ is a function that checks if triple t is violating one of the hijacking rules. Therefore, the value of this metric illustrates the percentage of triples that have some form of ontology hijacking, against all possible ontology hijacking triples.

Results Overview: Similar to the Metric CS5, the variation in quality within most of the assessed datasets ($\approx$ 86% of the datasets) is very low, though due to a number of outliers (shown in Figure 7 Metric CS6), the standard deviation value (σ_s) stands 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 Datatypes Metric

In a schema, a property can optionally have a domain and range types defined. These definitions determine in what class type a resource should be used (the domain) and what is the expected type for its value. Similar to the most metrics defined in this section, using incorrect domain and range datatypes would not break the RDF data model. Nevertheless, it 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. [52] classifies this metric under consistency

Metric Computation: This metric assesses a dataset for the type validity of the domain and range of its statements, according to the schema of the predicate used. In particular, the predicate of each triple is dereferenced where the domain and range types were extracted, together with the types' inferred parent types. Following that, the subject and the object resource types are checked against the domain and range types for the particular property. We quantify this metric as follows:

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

$$\overline{dom}(D) := \{t \mid \forall t \in \mathcal{R} \cdot ((\mathcal{T}(t.s) \cap dom(t.p)) = 0)\}$$

$$\overline{ran}(D) := \{t \mid \forall t \in \mathcal{R} \cdot ((\mathcal{T}(t.o) \cap ran(t.p)) = 0)\}$$

where $\mathcal{R}$ is a set of sampled (which sample can be as big as the dataset under assessment) triples from the assessed dataset D (i.e. $\mathcal{R} \subseteq D$), $\mathcal{T}(r)$ is a function that returns the type of the local resource[32] r, the functions $ran(p)$ and $dom(p)$ return a set of range and domain types respectively for the predicate p together with their inferred parents. The metric value is a ratio between the total number of incorrect domain/range datatypes in statements and the total number of items in the reservoir $\mathcal{R}$ multiplied by 2 - since we are assessing the predicate of a triple twice (once for its domain, and another for its range).

Results Overview: This metric is implemented as a probabilistic metric using the reservoir sampling, in a similar manner as explained in [17]. Our assessment shows that data publishers tend to use the 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 average score for this metric is 60.11% whilst the standard deviation (σ_s) 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 (Metric CS1 to Metric CS6), suggesting that Linked Data publishers might be more laid-back 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 [17] we have shown that with the right parameters probabilistic approximation techniques can provide a good estimate quality value.

(SV3) Compatible Datatype Metric

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 can represent infinitely anything, whilst the datatype attached to the value can be used to interpret the data concisely. In [8], the authors describe four benefits of having good quality literals including *efficient computation*. 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 [8]. It is recommended that publishers add the datatype to the literals. This metric is classified under the syntactic validity dimension [52].

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` (translated to `http://www.w3.org/2001/XMLSchema#integer`) is its datatype. Along these lines we quantify this metric as follows:

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

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}. Untyped literals are ignored in this metric as they cannot be validated against an unknown datatype. Therefore, the value of the metric is a ratio between the number of correctly typed literals and the total number of typed literals in the assessed dataset D.

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

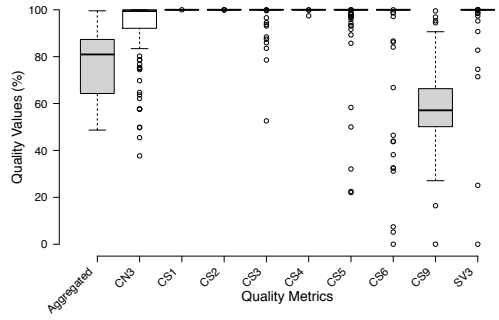

Fig. 7. Intrinsic category box plot. Outliers are represented by dots.

Results Overview: The box plot for metric SV4 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 σ_s value is a high 14.16% (median 100%). Datasets that had no literal values were omitted from this assessment. In order to reduce incompatible datatypes vis-a-vis the lexical form of a data value, publishers could publish and serialise their data using the latest Turtle 1.1 parser, as it relaxes and simplifies the serialisation of such literals.

Aggregated Results

Table 6 shows the aggregated ranking of the top and bottom 5 datasets from the intrinsic category point of view. 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 that is slightly varied having a σ_s value of 12.89% and a median of 80.94%. The majority of the metrics shows that a relative high quality (mean value of 77.36%) is adhered to by Linked Data publishers.

5.5. Accessibility Category

The accessibility category groups quality indicators related to the proper access functions of the Linked Data resources. The dimensions in this category deals with the ease of using Linked Data resources. In [52], Zaveri et al. classify metrics under the following dimensions: (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 hosting servers affect the efficiency of a data consumer. 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).

(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 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 Redirect code [45].

Metric Computation: The aim of this metric is to check the number of valid deferencable 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 Linked Data valid mime-types (e.g. application/rdf+xml, text/n3, text/turtle, etc...). A correct server-side dereferencing mechanism, should identify that the requested resource is an *abstract concept* and thus replies with a `303 See Other` and a redirect location where the *real-world object* (of the desired format) is. 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"* [24, §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. For this metric we use adapt the sampling approach, similar to the one described in [17, §5.1], in order to get a representative sample of the assessed dataset URIs. Along these lines we adapt the metric from Hogan et al. [28, §5.1, Issue III]:

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

where $\mathcal{R}$ is the set of sampled URIs in the dataset, $\mathcal{U}$ is the set of URIs in the dataset D, and $deref(u)$ is a function returns *true* if the URI u being examined follows the dereferencability rules as described above.

Dataset	$v(C, 1.0)$	CN3	CS1	CS2	CS3	CS4	CS5	CS6	CS9	SV3
http://extbi.lab.aau.dk/resource/	99.55%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	90.63%	100.00%
http://fao.270a.info/	98.74%	99.97%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	73.83%	100.00%
http://worldbank.270a.info/	98.40%	99.99%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	66.75%	100.00%
http://uis.270a.info/	98.39%	99.99%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	66.19%	100.00%
http://imf.270a.info/	98.31%	99.99%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	64.68%	100.00%
				...						
http://citeseer.rkbexplorer.com/	57.69%	80.20%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	50.00%	N/A
http://lingweb.eva.mpg.de/ids/	57.55%	78.60%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	58.35%	N/A
http://acm.rkbexplorer.com/	56.09%	75.40%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	50.00%	N/A
http://jisc.rkbexplorer.com	54.92%	78.55%	100.00%	100.00%	52.60%	100.00%	99.90%	100.00%	50.96%	N/A
http://www.productontology.org	48.68%	49.90%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	73.07%	N/A

Table 6
Overall ranking of datasets for the Intrinsic category.

The metric's value represents the percentage of valid dereferenceable URIs in a dataset.

Results Overview: In [17, §5.1] we describe a probabilistic technique for this metric using reservoir sampling, however, such sampling might lead to an unbalanced representative sample. We therefore adopt a hybrid of the technique used in [17, §5.1] and the *stratified sampling* idea as described in [22]. Stratified sampling is a technique that can be used when the data can be partitioned into a number of disjoint subgroups [22]. The idea is that the sample is chosen per-proportion of these subgroups, therefore improving the representative sample. 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. However, 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 this metric (A3) in Figure 8 shows a large varying quality with the box plot ranging all values from 100% to 0%. The average quality value of this metric is 36.86%, which is 33.44% lower than the average recorded in [28, §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. 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 right skewed, meaning that the assessment shows a high concentration of low quality values. Similar to [28, §5.1 – Issue III], our assessment shows a high variability between data producers on the dereferencability of resources. We report a σ_s value 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 an 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. In the meantime, a possible solution is that data publishers make use of Linked Data-based content management systems that handles 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 state-*

ment does not grant consumers the automatic right to use that content/data." - [24, §4.3.3]

Licences, as defined by the Open Definition [20], are the heart of open data. It is the mechanism that defines 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 license specification should also be included in a dataset's metadata.

(L1) Machine-Readable License

Having machine-readable license definitions (such as `http://purl.org/NET/rdflicense` [44]), 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, attribution, etc ...). 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/.*`) 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 a "license" predicate:

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

We quantify this metric as follows:

$$L1(D) := \begin{cases} true\ if\ (lpr(t_p) \wedge lvld(t_o)) \\ false\ otherwise \end{cases}$$

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.

Results Overview: In Section 3.2 we discuss the licences and rights in the LOD Cloud datasets' metadata. We show how around 41% of the whole LOD 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. However, our assessment resulted in just 17 datasets ($\approx$ 13%) that contained at least one machine-readable license. Whilst we have to acknowledge that our data acquisition process did not take into consideration sources other than the LOD Cloud metadata (CKAN metadata was not included in the assessment), such **open** datasets should make this information explicit, as not all agents will have access to the LOD Cloud metadata. For example, dataset metadata can easily add machine-readable license statements by using other linked open datasets such as [44].

(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 triple with a predicate `dct:description`, `rdfs:comment`, `rdfs:label`, or `schema:description` and a literal 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 quantify this metric as follows:

$$L2(D) := \begin{cases} true\ if\ (t_p \in p_{hrdesc} \wedge lregex(t_o)) \\ false\ otherwise \end{cases}$$

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.

Results Overview: Similar to Metric L1, the assessment shows a low overall level of conformance to this metric. We detected human-readable licenses in 11 ($\approx$ 8.46%) datasets, 4 of which also had a machine-readable license. Whilst it is understandable that publishers are less inclined to have 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 [20, §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 they (referring to agents) can discover more things."* [10]. Furthermore, Berners-Lee states that linking your data to external sources would earn the dataset the fifth star (`http://5stardata.info`), given that the rest of the 4 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 [24] describe 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."* – [24, §2.5]

These external outlinks is what makes the Linked Data ideology stands out from others. Well-interlinked data enables better analysis and understanding of the data. The interlinking property is often used in order to identify the importance or authority of a data source in the Web of Data. For example in [46], 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 than 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 similar to how we described it in [17][33]. Along these lines, we quantify the metric as follows:

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

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 external RDF links the assessed dataset has.

Results Overview: Similar to Metric A3, this metric was assessed using a sampling technique. We used the reservoir sampling technique, where 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 7 shows the top five assessed datasets, the number of unique dereferenceable external PLDs linked in the dataset, and the total number of unique PLDs. From the LOD Cloud dataset acquired sample, 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 amounted to 977,609. Three datasets,

[33] The sampling method is the same as in the cited literature, however, we implemented the metric in a different manner, as described in the article (Metric I1.

namely dbpedia.org, kent.zpr.fer.hr, and www.pokepedia.fr accounted for around 97% of these PLDs. However, the actual number of dereferenceable PLDs is 3086, which is around 0.31% of the linked external PLDs.

If one considers the ratio of actual Linked Data external PLDs and the total possible external PLDs in a dataset, we found that 7 datasets resulted in 100%, whilst 36 datasets scored 50% or more. On average, the ratio of total possible external PLDs and actual Linked Data external PLDs is 27.71%, whilst the deviation (σ_s) value is 30.94% For example, the top two datasets scored 86.38% and 86.97% respectively, whilst for dbpedia.org the value was less than 0.01%.

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 provide a large deviation (σ_s) of 183.3 Linked Data PLDs, and an average of 27.01 Linked Data PLDs. Nevertheless, one should consider that these two statistical descriptions are highly influenced by the top two datasets. Data publishers are encouraged to use interlinking tools such as Silk [50], LIMES [37], or DHR [21], therefore ensuring that they abide by the Linked Data principles. Silk[34] is a flexible link discovery framework allowing users to define linkage heuristics using a declarative language. Similarly, LIMES[35] is a large-scale link discovery framework based on metric spaces. Unlike Silk and LIMES, in DHR [21] the links are discovered by closed frequent graphs and similarity matching.

(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 are served. Therefore, in an ideal situation, a host has the capacity to handle a large number of parallel requests.

Metric Computation: Adapting the metric from [19], 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 10 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 served requests per second, computed as the ratio between the total number of requests sent to the dataset's host. We quantify this metric, adopting [19] as follows:

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

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.

Results Overview: The box plot for this metric (PE2) in Figure 8 shows a large varying quality with the box plot ranging all values from 100% to 0%. The σ_s value stands around 45.60% (median value is 29.67%) whilst the mean value is 47.78%. The box plot is right 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 metric 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 [19,24]. Hash URIs would reduce the number of HTTP round-trips, as the document with the requested fragment resource description would have other resource descriptions 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 other hand, Slash URIs should have to do the whole dereferencing process, though the client will only receive 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).

[34] http://silk-framework.com
[35] http://aksw.org/Projects/limes

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 7
Top 5 ranked datasets for the links to external RDF data providers metric.

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 10 local resources from the dataset under assessment. This metric is defined as the average 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 [19]. Along these lines, Metric PE3 is quantified as follows:

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

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

Results Overview: 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 value (σ_s) is 47.12% with a mean value of 57.55%. However, unlike PE2, the metric's values are left 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 1 request is answered in 1 second or less.

Aggregated Results

Table 8 shows the aggregated ranking of the top and bottom 5 datasets from the accessibility category point of view. 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 σ_s value of 19.00% and a median of 29.96%. The box plot is skewed right, showing a large concentration of low quality values, with the average aggregated quality score being 33.12%, with only 19% of the assessed datasets scoring 50% or more. The aggregated value is affected by the low 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.

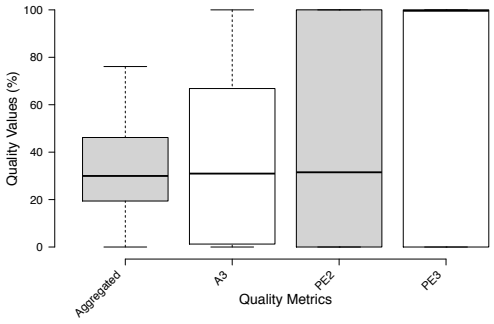

Fig. 8. Accessibility category box plot. Outliers are represented by dots.. Machine-Readable License, Human-Readble License, and Links to External Data Providers metrics are excluded, but included in the aggregated result box plot.

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 [15], 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 as defined in [28]:

$$pb_x(D) := \frac{((size(\overline{D_x}) + 1) - pos_x(D)) \times 100}{size(\overline{D_x})}$$

where, x 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 con-

Dataset	$v(C, 1.0)$	A3	L1	L2	I1	PE2	PE3
http://fao.270a.info/	76.10%	66.82%	100.00%	0.00%	73.28%	100.00%	100.00%
http://frb.270a.info/	75.87%	64.29%	100.00%	0.00%	73.28%	100.00%	100.00%
http://ecb.270a.info/	75.82%	68.36%	0.00%	100.00%	73.28%	100.00%	100.00%
http://oecd.270a.info/	74.41%	51.9%	100.00%	0.00%	73.28%	100.00%	100.00%
http://uis.270a.info/	72.21%	35.26%	100.00%	0.00%	73.28%	100.00%	100.00%
...							
http://vocabulary.wolterskluwer.de/court	0.08%	0.60%	100.00%	0.00%	0.00%	11.23%	55.34%
http://www.lingvoj.org/	0.00%	-	0.00%	0.00%	0.00%	0.00%	0.00%
http://prefix.cc/	0.00%	-	0.00%	0.00%	0.00%	0.00%	0.00%
http://transport.data.gov.uk/	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
http://msc2010.org/mscwork/	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%

Table 8
Overall ranking of datasets for the accessibility category.

formance 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:

$$as(D) := \frac{\sum_{m_{scr} \in \{RC1 \ldots PE3\}} m_{scr}(D)}{size(\{RC1 \ \ldots \ PE3\})}$$

where m_{scr} is the result of a dataset for a computed metric, and $\{RC1 \ \ldots \ PE3\}$ are the metrics described in this article. The aggregated scores only took into consideration the computed metrics. For example, in the case of the top placed dataset where all metrics were computed for the dataset, the average was taken over all 27 metrics. On the other hand, the second placed dataset was only available from a SPARQL endpoint which unfortunately did not manage to complete the evaluation (after a number of tries) due to various exception that we describe in the next subsection. Therefore, in that case, the aggregated score for those datasets was taken over 16 metrics.

All quality results and overall ranking is available at `http://jerdeb.github.io/lodqa/ranking`. Nonetheless, we do not claim that lower ranked datasets are hosting poor quality data, but instead our claim is that following this study these datasets are less conformant to the quality metrics assessed.

From a total of 239 datasets, only 130 datasets (totalling 3.7 billion quads were assessed. The average aggregated conformance score is 59.33% with a slight deviation (σ_s) 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.

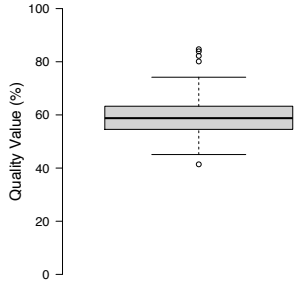

Fig. 9. Aggregated Conformance Score box plot. Outliers are represented by dots.

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[36] 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

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

– required for the query to stream triples; or (ii) different timeout settings (500 Server Error) – which might interrupt the assessment at a random point.

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.

6.1. The Principal Component Analysis

The Principal Component Analysis (PCA) [40] 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 [7,38] 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. [52] and assessed during this empirical study that can give us sufficient information about a linked dataset's quality?

Therefore, in this analysis, using PCA, we are looking 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 will help us to find the best possible quality metrics that summarises the quality of linked datasets as well as possible, in terms of new characteristics (components). In doing so we group the quality metrics into a series of components, where each group means that the metrics in that component would have significant variance on describing 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 BartlettâĂŹs 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 9
KMO and Bartlett's Tests.

In Table 9 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 [30]. The Bartlett‘s test gave a significance level of .000, that is, we can reject the null hypothesis (H_0) at $p < 0.05$, where p value is the significance level.

Following the rejection of the null hypothesis, we will use the Principal Component Analysis (PCA) in order to test the alternative hypothesis (H_a). Table 10 shows the total variance explained. In the Initial Eigenvalues column, the Table displays the eigenvalues associated with each component, 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 factors extracted is 11. In order not to give too much importance to one component over another, a rotated component matrix (Table 11) 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 10 factors can explain around 72.59% of the total variance. The other 16 components will only explain 27.14% of the variance.

In Table 11 we can see the 11 extracted components and the metrics each component represents. Each cell represents the correlation of a metric with a component. For the factor loading we use a cut-off point of 0.5[37] 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 main research question, the aim of this study is to highlight the key quality indicators that where classified in [52] 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 3 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.

[37] Based on: `http://imaging.mrc-cbu.cam.ac.uk/statswiki/FAQ/thresholds`. Accessed on 20th August 2016

Our initial quality assessment was generic, therefore all 130 datasets had the same 27 metrics assessed against them, irrelevantly if the metric is important to a particular dataset for a particular domain or not. 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.

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 the future directions – in this case of Linked Data. In this article we quantified and analysed a number of linked datasets vis-à-vie a number of quality metrics as classified in [52]. 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 3 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 we performed a primary investigation in order 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

	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 10
Total variance explained.

only around 42% had a valid Linked Data access point, whilst only 40% had a license.

In [25], 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 look at a number of datasets in order to understand better whether the perception label is deserved. In Section 5 we look at the datasets themselves in order to assess their quality against a number of metrics. We have seen that data publishers are compliant in 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 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 this might not be the case. 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.

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

	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 11
Rotated component matrix.

(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] 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.

[6] 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.

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

[8] W. Beek, F. Ilievski, J. Debattista, S. Schlobach, and J. Wielemaker. Literally better: Analyzing and improving the quality of literals (under review). *Semantic Web Journal*, 2016.

[9] 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.

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

[11] 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.

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

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

[14] D. Brickley, R. Guha, and B. McBride. Rdf schema 1.1. W3c recommendation, World Wide Web Consortium (W3C), February 2014.

[15] J. Debattista, S. Auer, and C. Lange. Luzzu - a methodology and framework for linked data quality assessment. (To Appear).

[16] 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.

[17] J. Debattista, S. Londoño, C. Lange, and S. Auer. *Quality Assessment of Linked Datasets Using Probabilistic Approximation*, pages 221–236. Springer International Publishing, Cham, 2015.

[18] 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.

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

[20] T. O. K. Foundation. The Open Definition.

[21] N. Hau, R. Ichise, and B. Le. Discovering missing links in large-scale linked data. In A. Selamat, N. T. Nguyen, and H. Haron, editors, *ACIIDS (2)*, volume 7803 of *Lecture Notes in Computer Science*, pages 468–477. Springer, 2013.

[22] 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.

[23] P. J. Hayes and P. F. Patel-Schneider. Rdf 1.1 semantics. W3c recommendation, World Wide Web Consortium (W3C), February 2014.

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

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

[26] 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.

[27] 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.

[28] 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.

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

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

[31] B. F. Lóscio, C. Burle, and N. Calegari. Data on the web best practices. W3c recommendation candidate, World Wide Web Consortium, February 2016.

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

[33] F. Maali, J. Erickson, and P. Archer. Data catalog vocabulary (dcat). W3c recommendation, World Wide Web Consortium, 2014.

[34] A. Mallea, M. Arenas, A. Hogan, and A. Polleres. On blank nodes. In *10th Int. Semantic Web Conf.*, ISWC'11, pages 421–437, Berlin, Heidelberg, 2011. Springer.

[35] A. J. Max Schmachtenberg, Christian Bizer and R. Cyganiak. Linking open data cloud diagram 2014, 2014.

[36] 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.

[37] A.-C. N. Ngomo and S. Auer. Limes: A time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Three*, IJCAI'11, pages 2312–2317. AAAI Press, 2011.

[38] P. B. Nguyen, M. Luong, and A. Beghdadi. *Statistical Analysis of Image Quality Metrics for Watermark Transparency Assessment*, pages 685–696. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

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

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

[41] S. Peroni. Media type as linked open data.

[42] L. Pipino, Y. Lee, and R. Wang. Data quality assessment. *Communications of the ACM*, 45(4), 2002.

[43] 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.

[44] V. Rodriguez-Doncel, S. Villata, and A. Gomez-Perez. A dataset of rdf licenses. In I. O. S. Press, 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.

[45] L. Sauermann and R. Cyganiak. Cool uris for the semantic web. W3c interest group note, World Wide Web Consortium, 2008.

[46] 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.

[47] M. K. Smith, C. Welty, and D. L. McGuinness. Owl web ontology language guide. W3c recommendation, World Wide Web Consortium (W3C), February 2004.

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

[49] O. Thréaux. Common http implementation problems. W3c note, World Wide Web Consortium, Jan. 2003.

[50] J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *Proceedings of the 8th International Semantic Web Conference*, ISWC '09, pages 650–665, Berlin, Heidelberg, 2009. Springer-Verlag.

[51] 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, Germany, 2014. CEUR-WS.org.

[52] A. Zaveri, A. Rula, A. Maurino, R. Pietrobon, J. Lehmann, and S. Auer. Quality assessment for linked data: A survey. *Semantic Web Journal*, 7, 2015.