Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud

Jeremy Debattista
debattij@tcd.ie
ADAPT Centre,
School of Computer Science and Statistics
Trinity College, Dublin
Ireland

Judie Attard
attardj@tcd.ie
ADAPT Centre,
School of Computer Science and Statistics
Trinity College, Dublin
Ireland

Rob Brennan
rob.brennan@dcu.ie
School of Computer Science
Dublin City University
Ireland

Declan O'Sullivan
declan.osullivan@scss.tcd.ie
ADAPT Centre,
School of Computer Science and Statistics
Trinity College, Dublin
Ireland

ABSTRACT

The Linked Open Data (LOD) cloud has been around since 2007. Throughout the years, this prominent depiction served as the epitome for Linked Data and acted as a starting point for many. In this article we perform a number of experiments on the dataset metadata provided by the LOD cloud, in order to understand better whether the current visualised datasets are accessible and with an open license. Furthermore, we perform quality assessment of 17 metrics over accessible datasets that are part of the LOD cloud. These experiments were compared with previous experiments performed on older versions of the LOD cloud. The results showed that there was no improvement on previously identified problems. Based on our findings, we therefore propose a strategy and architecture for a potential collaborative and sustainable LOD cloud.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability; • **Information systems → World Wide Web**; *Web searching and information discovery*; *Semantic web description languages.*

KEYWORDS

Linked Data, LOD cloud, metadata quality, data quality, sustainable services

ACM Reference Format:
Jeremy Debattista, Judie Attard, Rob Brennan, and Declan O'Sullivan. 2019. Is the LOD cloud at risk of becoming a museum for datasets? Looking ahead towards a fully collaborative and sustainable LOD cloud. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3308560.3317075

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3317075

1 INTRODUCTION

The Linked Open Data (LOD) cloud is regarded by many as a significant contribution towards the cause of adopting Linked Data(sets) and Semantic Web technologies, both in an academic context, but also to some extent in a commercial/industry one. Over time, the LOD cloud evolved into a *clustered catalog* of individual domain specific knowledge graphs. Despite this separation into various knowledge graphs, the LOD cloud demonstrates cohesion between these data sources with interlinks (links between datasets of a different domain), and intralinks (links between datasets of the same domain). The LOD cloud can be viewed by publishers as a catalog that can be crawled by data consumers for discovering datasets that can be re-used or linked to. For this to be possible, datasets on the LOD cloud has metadata attached to them, which its aim is to provide a level of understanding of how the data can be accessed and used.

The LOD cloud has been subject to a number of studies, especially on its metadata [2, 8, 15]. These studies highlight the shortcomings of the dataset metadata of the LOD cloud with regard to dataset accessibility, licensing, and metadata structure. These problems resulted in (1) dead links/data sources, (2) incorrect resolving of datasets, and (3) unclear usage of datasets in terms of licenses. These experiments were performed on previous versions of the LOD cloud, where the visualisation was less dynamic than lately; where monthly publication of the LOD cloud is currently performed. We therefore followed up on previous experiments to observe whether the problems with the metadata persist. Our results show that they do. We were therefore motivated to identify and discuss the current challenges going forward, and hence propose a sustainable architecture to help the Linked Data community to overcome these challenges. In a similar fashion, we also followed up on our previous quality assessment [8] survey in order to identify whether the

quality of accessible datasets on the LOD cloud has improved or otherwise.

The main contributions of this paper are therefore:

- An evaluation on the metadata of the LOD cloud (Section 2);
- A discussion of an on-going periodical data quality assessment over LOD cloud datasets on 17 different quality metrics (Section 3);
- Identification of a strategy and architecture for a potential collaborative and sustainable LOD cloud (Section 4).

We conclude our paper with related work in Section 5 and final remarks in Section 6.

2 THE DISCOVERABILITY AND OPENNESS OF DATASETS IN THE LOD CLOUD

Currently[1], the LOD cloud visualises 1,369 datasets[2]. The cloud diagram is updated every month with a tendency in increasing the number of dataset at each iteration. The inclusion criteria is based on the publishing of datasets following the Linked Data principles [3] and a set of five-rule inclusion guideline [13]. Furthermore, the "open" keyword suggests that datasets should also follow the open data definition, meaning that the dataset should "*be freely used, modified, and shared by anyone for any purpose*" [16]. Based on these premises, we conduct a number of experiments on the available metadata in order to understand better how these are enforced. Therefore, we will be looking at (1) what licenses are used, (2) if linked datasets are using the correct media types, and (3) the accessibility of the datasets. The size of the dataset and the number of external interlinks are disregarded in this experiment as these are related to the dataset itself. All experiments performed in this section are available online[3]. LOD cloud data can be retrieved from https://lod-cloud.net/lod-data.json and in this section we refer to this data as the *JSON data file* or the *data file*.

Experiments on dataset metadata have been conducted on previous versions of the LOD cloud [2, 8, 15]. In this paper we will compare the observed results to our previous work [8], where similar experiments were conducted. In [8], the observations were made in December 2015, based on the datasets that were linked and visualised in August 2014[4].

2.1 Can I freely use the dataset?

Licences are the heart of Open Data. They define whether third parties can re-use data or otherwise, and to what extent. In Linked Open Data, one would expect that such licenses are at least in a machine-readable format. Additionally, having the license mentioned in human-readable format, within the metadata's description, is a bonus. The open knowledge foundation recommend the usage of one of the following main licences to be in conformant with the open definition principles [16]:

- Creative Commons CCZero (CC0)
- Open Data Commons Public Domain Dedication and Licence (PDDL)
- Creative Commons Attribution 4.0 (CC-BY-4.0)
- Open Data Commons Attribution License (ODC-BY)
- Creative Commons Attribution Share-Alike 4.0 (CC-BY-SA-4.0)
- Open Data Commons Open Database License (ODbL)

For this experiment we parsed through each dataset in the data file and looked for the value attributed to `license` key. Our experiments show that around 45% (619 datasets) had a defined license. Figure 1 shows the frequency of the licenses used. The Creative Commons Attribution license is the most frequently used with 208 datasets or 15% using one of the recommended licenses. This is followed by the Creative Commons Attribution Share-Alike license (116 datasets) and Creative Commons Zero (89 datasets). 85 datasets are using the Creative Commons Non-Commerical license, which is not recommended by [16]. A more worrying aspect is that most of these licenses are links to human readable webpages which therefore cannot be understood by machine agents. In comparison to [8], our observations show an increase of 5% in the number of datasets that included a license in their metadata. We also noticed that whilst each dataset metadata can be exported to an RDF-based serialisation, the license is missing (e.g. https://lod-cloud.net/rdf/bio2rdf-taxonomy?format=ttl).

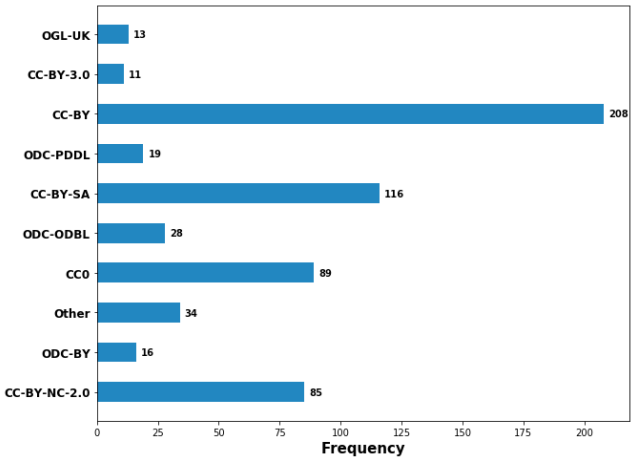

Figure 1: Overview of licenses used on the latest version of the LOD cloud.

We performed a second experiment in order to observe whether license data was included in the datasets' description field in a human-readable form. A regular expression that captures *license* or *copyright* and one of *under*, *grant* or *right* was performed on all 1,369 description fields. This resulted in 22 matches. We manually inspected the descriptions and observed that there were a total of 10 datasets that had a conformant license described, and 5 non-conformant licenses described. This gives an increase of 2 datasets from the observations in [8]. We also observed that there were 4 bad matches, meaning that these datasets were only listing the licenses of the data sources used to create the linked dataset. We also noticed that there were 3 datasets that had a particular license

[1]Based on the January 2019 crawl
[2]The website says 1,234 datasets, however, the official JSON file with the LOD cloud data indicates otherwise.
[3]https://github.com/jerdeb/lodexperiments
[4]For reference to the visualised LOD cloud diagram https://lod-cloud.net/versions/2014-08-30/lod-cloud.png. Last Accessed: 21st January 2019.

mentioned in the description but did not match the one used in the license field of the metadata.

2.2 Usage of the right media types for dataset distribution.

One of the main principles to add a dataset to the LOD cloud is that the datasets *"must resolve with or without content negotiation, to RDF data"* [13]. In this experiment we were interested in exploring the media types attached to the different distributions in the LOD cloud. Ideally, distributions in the LOD cloud use the respective media types, for example an RDF/XML data dump should use `application/rdf+xml`, as this would facilitate the uptake of the distributions by different agents by using the right content negotiation request. In Table 1 we list the different data types used for the distributions and their frequency. We observed a mixture of media types, however, the frequently used media type is `text/html`. Whilst this kind of media type is encouraged for RDFa serialisation, we observed that none of the distributions contained actual RDFa data. Similar to the findings in [8], we observed a number of unregistered media types. These included `RDF`, `n-quads`, `HTML`, and `application/x-ntriples` amongst others. 91 distributions had no media type assigned, whilst 109 distributions where assigned to media types that across the LOD cloud were less frequently used. We also noticed that there was a large number of distributions using `meta/void` (226 distributions) and `meta/rdf-schema` (370 distributions), however, these are also considered to be unknown media types.

Table 1: List of media types used in the distributions

Media Type	Frequency
mapping/owl	26
meta/owl	27
text/plain	31
application/x-gzip	32
n-quads	32
None	91
application/x-ntriples	91
meta/sitemap	102
application/x-nquads	103
Others	109
application/rdf+xml	114
application/octet-stream	118
HTML	119
application/zip	137
text/turtle	252
meta/void	266
meta/rdf-schema	370
RDF	401
text/html	1107

2.3 Accessibility of datasets

The final experiment on the datasets' metadata is to identify which datasets have a *potential* access point that allows for RDF crawling via an RDF dump, SPARQL endpoint, or a voID dataset description. We highlight the word *potential* because doing basic checks does not mean that the available distributions are well-formed RDF-based serialised datasets. For these three access points we set a number of criteria. The common criteria is that we first check if the access URL is online, with a 10 second timeout. For the data dumps the additional criteria is that the distributions have one of the following media types:

- `application/x-ntriples`
- `application/rdf+xml`
- `text/turtle`
- `application/x-nquads`
- `application/trig`
- `application/n-triples`
- `gzip:ntriples`
- `application/x-gzip`
- `application/octet-stream`
- `application/x-ntriples`
- `RDF`
- `plain/text`

For data dumps we added some invalid media types based on their frequency of use, however, this does not mean endorsement for wrong usage. With regard to SPARQL endpoints, the endpoint should answer to a simple `ASK` query, whilst the voID metadata should be able load in an in-memory graph structure and must answer return true to the query `ASK { ?s a <http://rdfs.org/ns/void#Dataset> . }`.

In contrast to [8] where it was observed that around 42% of the datasets had a potential direct access point from the LOD cloud, only 33% (454 datasets) of the datasets can potentially be accessed. Furthermore, from the 915 datasets that had no potential access points, 209 datasets had no data distributions, for example https://lod-cloud.net/dataset/slideshare2rdf, which violates the current LOD cloud inclusion criteria.

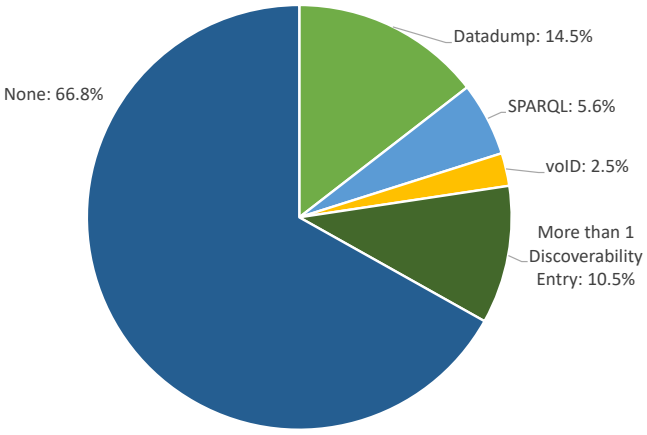

Figure 2: Visualisation of different access points for datasets.

Figure 2 illustrates the different access points. We can break this down as follows:

- Only Data dump - 199 datasets;
- Only SPARQL Endpoint - 77 datasets;

- Only voID - 34 datasets;
- Datadump and SPARQL Endpoint - 72 datasets;
- Datadump and voID - 62 datasets;
- SPARQL Endpoint and voID - 4 datasets;
- All three entry points - 6 datasets.

In Figure 3 we depict how the LOD cloud would look if only the datasets with a potential endpoint are visualised.

3 ASSESSING QUALITY ASPECTS OF REACHABLE DATASETS

Our next experiment focuses on the quality of datasets. Since August 2018, we performed monthly quality assessments on accessible datasets on the LOD cloud. The accessibility criteria used is the same as described in Section 2.3. In order not to re-download and pre-process data dumps[5], we monitor these for changes by querying the dumps URL header (`HTTP HEAD`) and check whether the `last-modified` or the `etag` values have changed. Unlike the accessibility metadata experiment, for quality assessment we also considered datasets that were only available via a LODLaundromat mirror. The aim of this experiment is to statistically observe datasets' quality, hence quality problems encountered are out of scope of this paper.

We used Luzzu [6], a Linked Data quality assessment framework, to assess the quality of the datasets on 17 different metrics from 3 different categories, mainly intrinsic, contextual, and representational. These metrics can be objectively assessed by any quality assessment framework, therefore limiting any bias that can arise from subjectively assessed metrics. The choice of metrics was left independent of any potential task, hence we understand that certain metrics might not be critical for different tasks, nonetheless, are important for the Linked Data community. On the other hand, the reason we have excluded accessibility category metrics from this experiment is due to the fact that we had a considerable number of datasets downloaded from LODLaundromat. Having a dataset crawled from LODLaundromat does not guarantee that the source is effectively online and hence would skew results. For a description and a formal definition of all metrics mentioned in this section, we refer the reader to [8]. The primary aim of this assessment was to create a service whereby data consumers can search for current LOD cloud datasets based on their quality. Furthermore, this assessment indicates whether the quality of linked datasets have improved or otherwise since the last assessment in [8]. The assessment in [8] was performed over 130 linked datasets. At time of writing we performed quality assessment on 451 datasets that could be directly accessed from the LOD cloud metadata (cf. Section 2.3). However, we also noticed that we had access to datasets that had very few or no triples. Therefore, prior to analysing the results we discarded the results of 71 quality assessments that were performed over datasets that had less than 100 triples. Overall, we assessed over 13B triples, and all downloaded and assessed datasets are available online as HDT dumps[6] and GZ dumps[7].

Figure 4 shows how the average quality changed between the previous assessment and over the 6 month period August 2018 and January 2019 for the representational category (6 metrics) and contextual category (2 metrics) metrics. The chart also demonstrates the number of datasets assessed at each observation date. The metrics assessed in these two categories are:

- (RC1) Keeping URIs Short
- (RC2) Minimal Usage of RDF Data Structures
- (IN3) Usage of Undefined Classes and Properties
- (IN4) Usage of Blank Nodes
- (V1) Different Serialisation Formats
- (V2) Usage of Multiple Languages
- (P1) Provision of Basic Provenance Information
- (U1) Human Readable Labelling and Comments

In most cases we observe that the average has increased between the 2016 assessment and the latest 6 month assessment. Given that the number of datasets has increased drastically, we also calculated the standard deviation for the different months and compared them with [8]. We observed that the spread of the values of all metrics were similar (± 5% standard deviation points), with the exception of IN4. The observation of August 2018 for IN4 had a standard deviation of 2.89% (median value of 99.95), when compared to 12.15% of the 2016 assessment. However, the standard deviation increased to around 6.66% for the following months. We observe that the datasets obtained in the period August 2018 and January 2019 have improved on metrics RC1, IN3, P1, and U1. With regard to P1 and U1, we observe an average increase of 7.7% and 7.2% points respectively, whilst for RC1 we saw an average of 5.9% increase, whilst for IN3 an average of 8.5% increase. There was no significant change with regard to the V1 and V2 metrics, which are not displayed in the chart as their value is an integer value in contrast to the rest which are percentages.

Figure 5 shows how the average quality values changed from the 2016 assessment and over the 6 month period for the intrinsic category metrics. The metrics assessed in this categories are:

- (CN2) Extensional Conciseness
- (CS1) Entities as Members of Disjoint Classes
- (CS2) Misplaced Classes or Properties
- (CS3) Misused OWL Datatype or Object Properties
- (CS4) Usage of Deprecated Classes or Properties
- (CS5) Valid Usage of the Inverse Functional Property
- (CS6) Ontology Hijacking
- (CS9) Usage of Incorrect Domain or Range Datatypes
- (SV3) Compatible Datatype

For the intrinsic metrics the situation seems more balanced. We observed an increase in metric CS9 or around 9.16% when comparing the first assessment against the 6 consecutive months, however the spread of the data points is similar. On the other hand, we saw a decrease in quality for the CN2 and CS6 metrics. The average decrease is of 4.24% and 6.47% respectively. Furthermore, we noticed that the standard deviation for CS6 increased by around 10% (median 100%), thus the average value would ultimately be affected by the less conformant datasets. Metric SV3 also saw a slight increase on average, however, we also noticed that the spread of quality among datasets in this regard is much less (on average 3.7% vs 14.6%, with both medians 100%) than in [8]. Therefore, from a statistical point of view datasets assessed between August 2018 and January 2019 tend to be more conformant towards this aspect than CS6.

[5] We follow the pre-processing step described in [8, §4.1]
[6] http://s001.adaptcentre.ie/lod/hdtdumps/
[7] http://s001.adaptcentre.ie/lod/gzdumps/

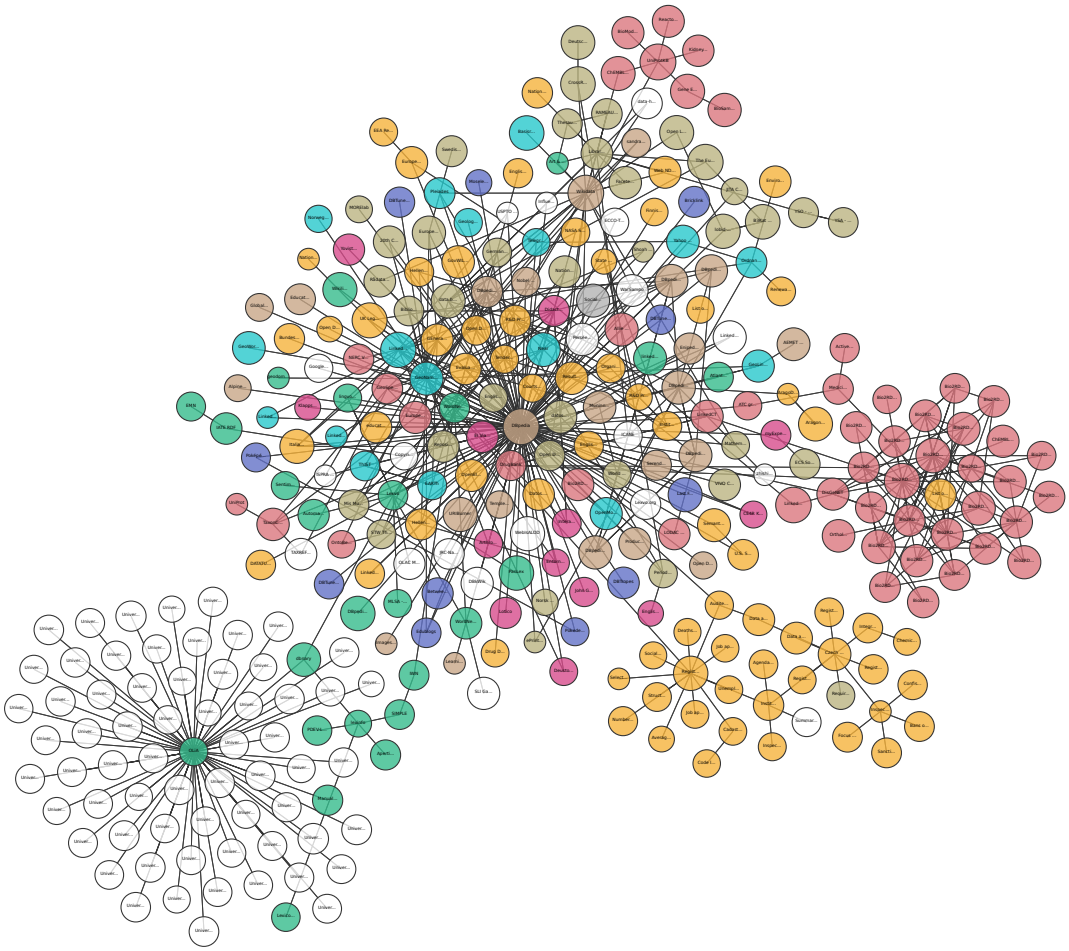

Figure 3: Datasets in LOD with a potential access point. The colours refer to domains as identified by the LOD cloud maintainers.

All quality metadata results and the crawled datasets used are available in an prototype online catalog: http://luzzu.adaptcentre.ie. The metadata schema is based on the Dataset Quality Vocabulary [7] (daQ), which is part of Luzzu's underlying semantic framework [6]. Unfortunately, we do not yet have enough data to potentially predict how the quality could look like in a number of years, however, we plan to keep our monthly crawls to gather more observations.

4 THE SUSTAINABLE LOD CLOUD

The current LOD cloud poses a number of barriers when data consumers are trying to identify a particular RDF dataset. Previous work [2, 8, 15] has already highlighted problems related to generated dataset metadata, however, as we discuss in Section 2, these have not been solved. Given the increase of RDF-based linked datasets on the web, the Linked Data community should ensure that the LOD cloud does not end up dormant and outdated, as happened between 2014 and 2017. Therefore, in order to make the LOD cloud a sustainable service and to provide a motivation for a new approach, we have to tackle the following challenges:

C1 - Publishers should own and maintain the datasets' metadata.
The LOD cloud is perceived as a monolithic structure with dataset metadata stored in a centralised catalogue. The metadata is initially created by the publishers themselves, however, changing this metadata is more difficult since this data is now "owned" by the LOD cloud maintainers. Therefore, whilst datasets might be updated, the metadata in the LOD cloud might still be old.

C2 - Lack of systematic and fine-granular metadata structures.
One of the major problems or challenges in the current state of the LOD cloud is that the metadata has no systematic structure in terms of properties, the property's values, and categorical values (e.g. media type). The current LOD cloud metadata attempts to leverage on both DCAT [12] and voID [1] standards, using predicates from both vocabularies to achieve a granular metadata description. This challenge

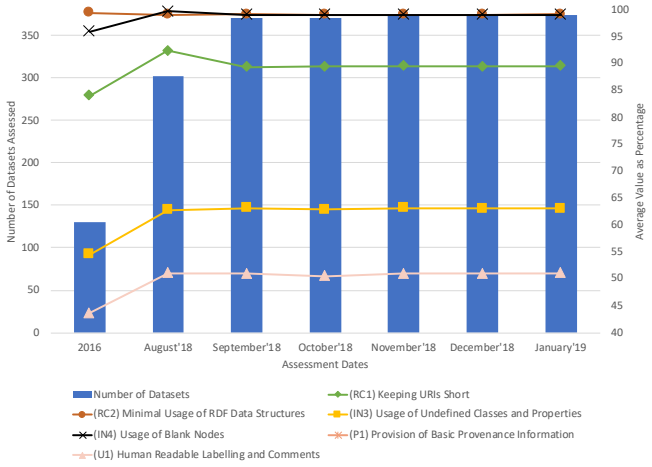

Figure 4: Aggregating quality metric results for representational and contextual metrics.

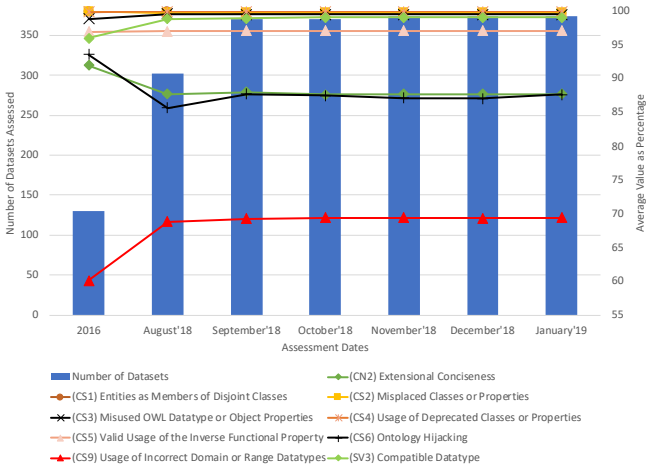

Figure 5: Aggregating quality metric results for intrinsic metrics.

is highly coupled with the next challenge, C3, which looks at the validation of metadata descriptions.

C3 - Invalid metadata descriptions.
Giving ownership of the metadata back to the publishers is no a guarantee that the metadata will be updated or factual. Therefore, validators should be implemented. The current LOD cloud suffers from metadata validation. Manually inspecting the DBpedia metadata [8] as an example, we found that the `void:dataDump` predicate links to the DBpedia downloads page. This is incorrect as the editors of the voID vocabulary documentation highlights that "*the* `void:dataDump` *property should not be used for linking to a download web page ... [it] should only be used for linking directly to dump files*" [1]. Furthermore, the license points to the a web page that describes the license, rather than the actual license itself (http://opendefinition.org/licenses/cc-by-sa/), whilst among the different media types attached to the dataset's distribution, we found `HTML`, `meta/void`, `linked data`, and `RDF`.

C4 - Many dead and outdated datesets listed.
Datasets that have perished from the web are still being visualised on the LOD cloud. The eleven 270a.info [4] datasets[9] are a notable example. In this case, we observe that an archived partial snapshot has been stored in LOD Laundromat[10], however, the metadata still shows the original description. Whilst archiving these datasets for preservation using tools such as LOD Laundromat is important, these datasets should not be visualised in the LOD cloud.

C5 - Lack of involvement of data consumers in the structure.
Up till now, data consumers could, painfully, crawl or search the LOD cloud diagram for a potential dataset. Prior to the 2018 update, most crawling or searching was done either by using datahub.io API or else by parsing the visualised SVG diagram. The latest versions provide a JSON file with all the datasets. Given the assumption that RDF/Linked Data consumers are potentially the largest set of users using the LOD cloud, the current versions still lack adequate filtering and searching tools, and lessons could be learned from the recent Google Dataset search[11] portal.

These five challenges guide us throughout this section to propose (1) a set of sustainability strategies for the LOD cloud, and (2) a potentially sustainable architecture based on Linked Data principles.

4.1 Sustainability strategies for the LOD cloud

The first step that is required in order to make the LOD cloud sustainable is to define a strategy. This strategy is based upon *people* or stakeholders, *processes* and *technology*. In a nutshell, the stakeholders of the LOD cloud service are the dataset publishers, the service maintainer/s, third party service providers, and the data consumers. The main process is the publishing of datasets on the LOD cloud using standards and an interoperable data model. Finally, the technology ensures the automation of adding and validating the status of datasets on the LOD cloud.

Based on Collibra's[12] experience with regard to building data governance solutions [14], we adapted their best practices as guidelines for the proposed LOD cloud service strategy as follows:

(1) **The service operating model.**
Challenge(s) to be tackled: C1, C5
Whilst acknowledging the previous and current maintainers, we propose that the LOD cloud service is operated on a federated model ensuring that the cloud remains open and the responsibility of the community as a whole. However, this requires instilling a culture that encourages interaction between the various stakeholders. *Publishers* should provide dataset metadata and be responsible to maintain it, ensure

[8] https://lod-cloud.net/dataset/dbpedia

[9] One such example: https://lod-cloud.net/dataset/fao-linked-data
[10] http://download.lodlaundromat.org/2a3bed796c47b67919645 9f3b5612b65
[11] https://toolbox.google.com/datasetsearch
[12] https://www.collibra.com

that their datasets are of high quality, and guarantee high availability uptime of their data services (e.g. SPARQL endpoints). The *LOD cloud service maintainer/s* must control services related to the generation, cataloging (i.e. keeping a list of all submitted and verified dataset metadata resource URIs), availability and maintenance of the cloud. *Consumers* should be able to comment (e.g. on usage) and vote on different data sources, helping potential future consumers to decide on whether a data source is right for their use case. Finally, *third party service providers*, such as quality assessors, should be able to identify changes in the LOD cloud and generate output that can be easily linked to the datasets' metadata in the LOD cloud.

(2) **Identification of critical data elements.**
Challenge(s) to be tackled: C2
In order to have a uniform view of the datasets in the LOD cloud, the critical element is the identification of a metadata standard, and a glossary or taxonomy for non-descriptive values (e.g. for licenses, or media types). Over the years, standards and taxonomies have been defined and based on Linked Data principles we advocate for reuse where possible. Datasets' metadata should clearly identify the ownership and usage of the dataset, providing: (a) *who* owns the data; allowing consumers to identify whether the publisher is a trusted source that can be reached to answer any questions; (b) *what* the dataset is about; including a basic description, purpose of the dataset, and the schemas underlying the data; (c) *where* could dataset distributions and/or other data services such as endpoints be found; and (d) *how* can the dataset be used in terms of licensing. Whilst most of this is already catered for in the current LOD cloud service, and improved over the years (for example using drop down lists in forms with predefined values) there is no agreed upon glossary amongst the publishers and the service maintainers themselves.

(3) **Defining the key activities and control structures for sustainability of service.**
Challenge(s) to be tackled: C3, C4
In order to define control measurements for the LOD cloud service, we first need to identify the key workflow activities. There are a number of key activities between the different stakeholders mentioned previously that would require different workflows. The first key workflow in the LOD cloud service is to check the validation and the correctness of a submitted dataset metadata resource URI by a publisher. The metadata resource URI is dereferenced as required by the Linked Data principles, and validated prior to inclusion in the LOD cloud. Another key workflow between the LOD cloud service and the publisher is the provision of regular heartbeat checks to ensure the availability of subscribed datasets for data consumers. Allowing for data consumers to vote and comment on specific data sources requires an authority or filtering from the LOD cloud service. A third workflow is a mechanism that prevents abuse or spam, ensuring sanity checks on subjective views of the datasets. This ensures that potential future consumers are not misled in using or disregarding a particular dataset or publisher. Similarly, the LOD cloud service should guarantee that external service providers are not biasing towards particular datasets or publishers, and any output that is generated (e.g. quality metadata), should contain extensive data lineage and provenance information, providing traceability to all stakeholders. Trustworthiness between the stakeholders, of the dataset, and other metadata generated by the different service providers, and the sustainability of the service, depends on the successful implementation of these control structures.

4.2 Capabilities and architecture of a LOD cloud service.

Based on the identified strategies, in this section we define the capabilities of the proposed LOD cloud service. The architecture captures the three pillars of the strategy, that is the *people*, the *processes* and the *technology*. The architecture is driven on the following capabilities:

Discovery - The LOD cloud service must enable data consumers, in this case both people and agents, to search, explore and identify the datasets required for a particular task in the most quick and efficient manner.

Understandability - The LOD cloud service is built on top of an interoperable semantic layer of standards and taxonomies that all publishers use uniformly. Furthermore, the external service providers must provide any output data using the same data model as in the LOD cloud service, i.e. RDF.

Social - The LOD cloud service allows all stakeholders to participate in the upkeep and uptake of the LOD cloud.

4.2.1 Architecture. Based on the idea of the LOD Research cloud[13], the underlying LOD cloud service should run using a Linked Data platform, which accepts Linked Data Notifications [5]. The idea is that once publishers publish their metadata resource on their servers, the publisher sends a Linked Data notification with the resource. Listing 1 illustrates a sample publishers' notification that is sent to the LOD cloud service. Upon receiving the notification, the LOD cloud service dereferences the received request and validates the dataset's metadata (Section 4.2.2). Once validated, the metadata's URI is stored in a registry and eventually visualised as part of the LOD cloud. Furthermore, the LOD cloud service will perform regular heartbeat checks on the publishers. Service providers and data consumers can then directly communicate with the publisher by dereferencing the metadata's URI that is stored in the LOD cloud service. Any output from the service providers and consumers (e.g. votes or comments) can be stored on the LOD cloud service triple store. Examples of service providers might include frameworks for quality, value, profiling, archiving amongst others, producing metadata that can be linked to datasets' metadata resources subscribed within the LOD cloud. Human consumers should be able to filter and search datasets based on various criteria, allowing them to vote or comment on these datasets, hence filtering mechanisms should be in place. Figure 6 depicts a high-level architecture diagram of the proposed solution.

```
@prefix schema: <http://schema.org/> .
[   a schema:CreateAction;
     schema:agent [
```

[13] https://linkedresearch.org/cloud. Last Accessed: 1st February 2019.

```
      a schema:Organisation;
      schema:name "The Org"
    ];
    schema:object <http://the.org/metadata/thedataset>
] .
```

Listing 1: Publisher's Linked Data Notification.

4.2.2 Dataset Metadata. To date, adding a dataset to the LOD cloud was either by adding dataset metadata directly to datahub.io with the LOD cloud tag, or more recently by filling out a form with all required fields. The former resulted into DCAT metadata, whilst for the latter form data is mapped into a combination of voID and DCAT metadata. Whilst ideally all publishers will publish their dataset metadata using one agreed upon standard, the LOD cloud ecosystem should allow any metadata published either by DCAT, voID or schema.org. The mandatory fields in the metadata are:

- **Title** - The dataset's title in textual form;
- **Description** - A concise textual description of the dataset which might also include information such as how it was used and a human-readable license;
- **Creator** - A resource which describes the creator or the publisher of the dataset, whom consumers should contact for questions related to the dataset;
- **Website** - A human-readable web page describing in more detail the dataset;
- **Full Download** - A resource which describes the distribution of the full data dump of the dataset. Each distribution must have the media type (mime type) described, for example for and RDF/XML data dump, one should use "application/rdf+xml". Furthermore, one should use the Media Types as Linked Data resources [17], which contains semantic descriptions for different RDF-based serialisation This is not mandatory if a SPARQL endpoint is provided;
- **SPARQL Endpoint** - Similar to full download, a resource that describes the access for the SPARQL endpoint, which might or might not include the different SPARQL protocols available. This is not mandatory if a full download is provided;
- **Domain** - A textual description of the domain of the dataset, for example financial or geospatial; and
- **License** - A machine-readable resource that describes the legality of reuse of the given dataset (dump or endpoint). Usage of correct machine-readable licenses such as [18] or creative commons semantic URIs are mandatory.

Other fields such as DOI, example resources, data catalogue, number of triples, and links could also be provided. Nonetheless, the LOD cloud services should identify the linked and number of triples for each subscribed dataset automatically.

This paper contributes towards an evaluation of the LOD cloud metadata, a periodical quality analysis of linked datasets that can be crawled from the LOD cloud, and propose strategies and an approach for making the LOD cloud service more sustainable.

5 RELATED WORK

Schmachtenberg et al. [19] crawled the Web of Data in order to present the 2014 version of the LOD cloud diagram. The authors analysed how different best practices were adopted by the crawled datasets, more specifically looking at provenance, licensing, and access methods amongst other experiments. In a similar crawling approach, Neto et al. [15] attempted to identify the actual Linked Open Data cloud by obtaining metadata from different sources. One of their goals was to assess the quality of the available RDF metadata. Similar to our findings, the authors report that the metadata suffered from a number of quality issues, mainly, lack of usage of standard vocabularies, incorrect usage of specific properties, and erroneous data. Assaf et al. [2] discussed the quality of the metadata of datasets available in the 2014 version of the LOD cloud. For this, the authors made use of the datahub.io API, thus using datahub's provided metadata as their evidence. Assaf et al. concluded that the metadata is in bad condition, mostly as a result of noisy data in licensing and accessibility metadata fields.

More recently, Debattista et al. [8] performed a study on the quality of the metadata and datasets available on the 2014 version of the LOD cloud. With regard to metadata quality, the authors conclude that based on the Open Defintion [16], approximately less than half of the datasets should not be part of the LOD cloud. The authors also suggested that the LOD cloud should reflect the Web of Data and be more dynamic. In terms of data quality, Debattista et al. assessed 27 metrics over 130 datasets. The authors concluded that the resulting overall aggregated average of slightly below 60% indicated that linked datasets might have a better quality than perceived [9]. In earlier studies, Hogan et al. [10] crawled and assessed the quality of around 12 million RDF statements. The aim of this study was to find and discuss common problems related to accessibility, reasoning, syntactical and non-authoritative contributions. In a follow up study [11], Hogan et al. assessed the quality of over 1 billion quads on a number of best practices and guidelines.

While a number of literature analysed the LOD cloud and metadata of linked datasets, to the best of our knowledge there is no work discussing how to concretely mitigate the identified problems. This motivate further our aim to propose a sustainable strategy for the LOD cloud service.

6 CONCLUSIONS

The first part of the article's title asks the question *Is the LOD cloud at risk of becoming a museum for datasets?* We have performed a number of experiments on the LOD cloud metadata on its structure, access and licensing parts. Values from these gave us a clear indication on the current status. We observed that there are datasets that have been offline but are still visualised, datasets that have incorrect or no license, and moreover datasets that do not adhere to the 5 LOD cloud inclusion principles. Furthermore, being Linked *Open* Data, one would expect that datasets follow the open definition [16] or the well known Linked Open Data 5 star. When we create the intersection of the dataset that have an access point and an open licence, we ended up with 35 datasets that can be called linked open datasets. On the other hand, the overall quality of datasets generally improved for the majority of datasets. Nonetheless, data publishers should invest in having quality checks within their publishing frameworks. Finally, when compared with similar experiments performed on earlier version of the LOD cloud, we can conclude that this significant contribution is in a great risk of becoming a museum for *linked* (partially *open*) *datasets*.

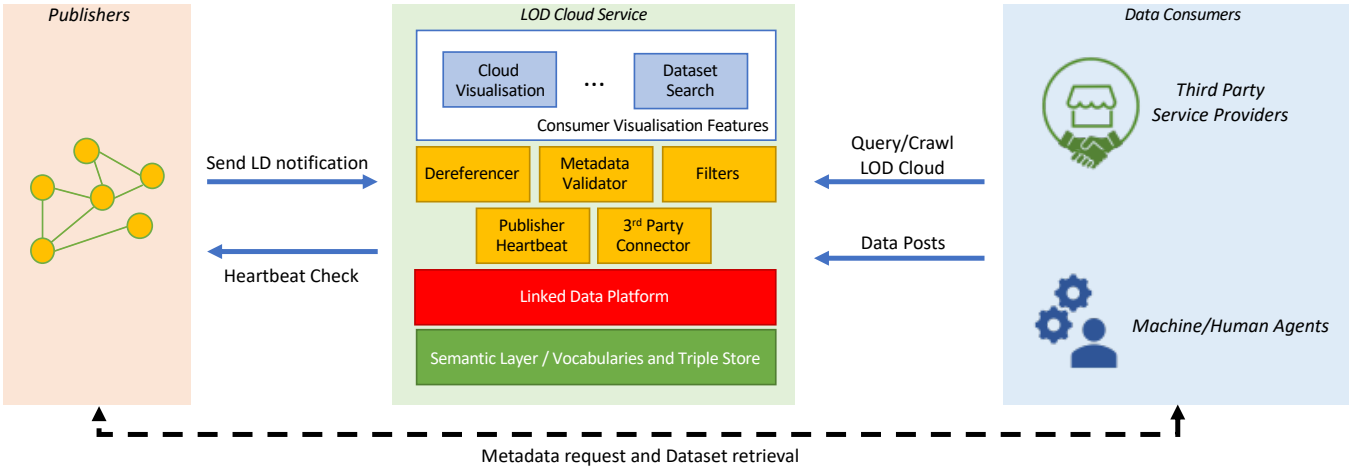

Figure 6: Proposed High-Level Architecture Diagram of the LOD cloud Service and Principle Data Flow.

Striving for the survival of the LOD cloud, in this article we propose a sustainable strategy for a new LOD cloud service, based on all stakeholders of the LOD cloud and standard Linked Data vocabularies and technologies. We discussed challenges that were identified in both previous literature and this study. Finally, we described the capabilities and an potential architecture for the LOD cloud service. As for future work, we encourage more discussion with the various stakeholders on how to save the LOD cloud from its untimely death, create a first prototype of the proposed architecture, and finally foster a culture of collaboration between all stakeholders of the LOD cloud.

ACKNOWLEDGMENTS

This research was partially supported by the Irish Research Council Government of Ireland Postdoctoral Fellowship [project ID GOIPD/2017/1204], Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie (EDGE) [grant agreement no. 713567], and Science Foundation Ireland and co-funded by the European Regional Development Fund through the ADAPT Centre for Digital Content Technology [grant number 13/RC/2106].

REFERENCES

[1] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. 2011. *Describing Linked Datasets with the VoID Vocabulary.* W3C Interest Group Note. World Wide Web Consortium.

[2] Ahmad Assaf, Raphaël Troncy, and Aline Senart. 2015. What's up LOD Cloud? - Observing the State of Linked Open Data Cloud Metadata. In *The Semantic Web: ESWC 2015 Satellite Events - ESWC 2015 Satellite Events Portorož, Slovenia, May 31 - June 4, 2015, Revised Selected Papers (Lecture Notes in Computer Science)*, Fabien Gandon, Christophe Guéret, Serena Villata, John G. Breslin, Catherine Faron-Zucker, and Antoine Zimmermann (Eds.), Vol. 9341. Springer, 247–254. https://doi.org/10.1007/978-3-319-25639-9_40

[3] Tim Berners-Lee. 2006. Linked Data – Design Issues. https://www.w3.org/DesignIssues/LinkedData.html. Accessed: 2017-12-15.

[4] Sarven Capadisli, Sören Auer, and Axel-Cyrille Ngonga Ngomo. 2015. Linked SDMX Data: Path to high fidelity Statistical Linked Data. *Semantic Web* 6, 2 (2015). https://doi.org/10.3233/SW-130123

[5] Sarven Capadisli and Amy Guy. 2017. *Linked Data Notifications.* W3C Recommendation. World Wide Web Consortium (W3C).

[6] J. Debattista, S. Auer, and C. Lange. 2016. Luzzu – A Methodology and Framework for Linked Data Quality Assessment. *Data and Information Quality* 8, 1 (Oct. 2016).

[7] Jeremy Debattista, Christoph Lange, and Sören Auer. 2014. Representing dataset quality metadata using multi-dimensional views. In *Proceedings of the 10th International Conference on Semantic Systems, SEMANTICS 2014, Leipzig, Germany, September 4-5, 2014*, Harald Sack, Agata Filipowska, Jens Lehmann, and Sebastian Hellmann (Eds.). ACM, 92–99. https://doi.org/10.1145/2660517.2660525

[8] Jeremy Debattista, Christoph Lange, Sören Auer, and Dominic Cortis. 2018. Evaluating the quality of the LOD cloud: An empirical investigation. *Semantic Web* 9, 6 (Sep 2018), 859–901. https://doi.org/10.3233/SW-180306

[9] Pascal Hitzler and Krzysztof Janowicz. 2013. Linked Data, Big Data, and the 4th Paradigm. *Semantic Web* 4, 3 (2013), 233–235.

[10] Aidan Hogan, Andreas Harth, Alexandre Passant, Stefan Decker, and Axel Polleres. 2010. Weaving the Pedantic Web. In *Proceedings of the WWW2010 Workshop on Linked Data on the Web, LDOW 2010, Raleigh, USA, April 27, 2010 (CEUR Workshop Proceedings)*, Christian Bizer, Tom Heath, Tim Berners-Lee, and Michael Hausenblas (Eds.), Vol. 628. CEUR-WS.org. http://ceur-ws.org/Vol-628/ldow2010_paper04.pdf

[11] Aidan Hogan, Jürgen Umbrich, Andreas Harth, Richard Cyganiak, Axel Polleres, and Stefan Decker. 2012. An empirical survey of Linked Data conformance. *Web Semantics: Science, Services and Agents on the World Wide Web* 14 (jul 2012), 14–44. https://doi.org/10.1016/j.websem.2012.02.001

[12] Fadi Maali, John Erickson, and Phil Archer. 2014. *Data Catalog Vocabulary (DCAT).* W3C Recommendation. World Wide Web Consortium. http://www.w3.org/TR/vocab-dcat/

[13] John P. McCrae, Andrejs Abele, Paul Buitelaar, Richard Cyganiak, Anja Jentzsch, and Vladimir Andryushechkin. 2019. Linked Open Data Cloud. http://lod-cloud.net

[14] Kash Mehdi. 2017. 4 Data Governance Best Practices to Kickstart your Data Governance Program. Blog Post. https://www.collibra.com/blog/4-data-governance-best-practices Last Accessed: 2019-01-24.

[15] Ciro Baron Neto, Dimitris Kontokostas, Amit Kirschenbaum, Gustavo Correa Publio, Diego Esteves, and Sebastian Hellmann. 2017. IDOL: Comprehensive & Complete LOD Insights. In *Proceedings of the 13th International Conference on Semantic Systems, SEMANTICS 2017, Amsterdam, The Netherlands, September 11-14, 2017*, Rinke Hoekstra, Catherine Faron-Zucker, Tassilo Pellegrini, and Victor de Boer (Eds.). ACM, 49–56. https://doi.org/10.1145/3132218.3132238

[16] Open Knowledge Foundation. [n. d.]. The Open Definition. http://opendefinition.org. Accessed: 2017-12-15.

[17] Silvio Peroni. 2016. Media type as Linked Open Data. http://www.sparontologies.net/mediatype/.

[18] Victor Rodriguez-Doncel, Serena Villata, and Asuncion Gomez-Perez. 2014. A dataset of RDF licenses. In *Legal Knowledge and Information Systems - JURIX 2014: The Twenty-Seventh Annual Conference, Jagiellonian University, Krakow, Poland, 10-12 December 2014 (Frontiers in Artificial Intelligence and Applications)*, Rinke Hoekstra (Ed.), Vol. 271. IOS Press, 187–188. https://doi.org/10.3233/978-1-61499-468-8-187

[19] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. 2014. Adoption of the Linked Data Best Practices in Different Topical Domains. In *13th Int. Semantic Web Conf. (Lecture Notes in Computer Science)*, Peter Mika, Tania Tudorache, Abraham Bernstein, Chris Welty, Craig A Knoblock, Denny Vrandecic, Paul T Groth, Natasha F Noy, Krzysztof Janowicz, and Carole A Goble (Eds.), Vol. 8796. Springer, 245–260. https://doi.org/10.1007/978-3-319-11964-9_16