Meta-Data for a lot of LOD

Laurens Rietveld [a], Wouter Beek [a], Rinke Hoekstra [a,b] and Stefan Schlobach [a]
[a] *Department of Computer Science, VU University Amsterdam, The Netherlands*
E-mail: {laurens.rietveld,w.g.j.beek,stefan.schlobach,rinke.hoekstra}@vu.nl
[b] *Leibniz Center for Law, Faculty of Law, University of Amsterdam, The Netherlands*
E-mail: hoekstra@uva.nl

Abstract.
This paper introduces the LOD Laundromat meta-dataset, a continuously updated RDF meta-dataset describing documents that are crawled, cleaned and (re)published by the LOD Laundromat [5]. This meta-dataset of over 110 million triples contains structural information for more than 650,000 documents (and growing). While traditionally dataset meta-data is often not provided, incomplete, or incomparable in the way they were generated, the LOD Laundromat meta-dataset provides a wide variety of structural dataset properties, including the number of triples in LOD Laundromat documents, the average degree in documents, and the distinct number of Blank Nodes, Literals and IRIs. This makes it a particularly useful dataset for data comparison and analytics, as well as for the global study of the Web of Data.

Keywords: Dataset Meta-data, Linked Data, Dataset Descriptions

1. Introduction

In this paper we present the LOD Laundromat meta-dataset, a uniform collection of dataset meta-data that describes the structural properties of very many (over 650,000) Linked Data documents containing over 38 billion triples. These Linked Data documents range from large compressed data dumps and RDFa embedded in web pages, to dereferenceable URIs and SPARQL `CONSTRUCT` query references. This meta-dataset is unique in its scale (both in terms of the 650,000 datasets it describes, and the number of meta-data properties), the consistent way in which meta-data properties are calculated, the explicit description of the computational processes used to calculate these properties, and the use cases it supports. The meta-dataset uniquely facilitates the analysis and comparison of many datasets, and supports research scenarios in which algorithms make innovative use of meta-data values to improve performance[18].

Analyzing, comparing, and using multiple Linked Open Datasets currently requires the hassle of finding a download location, hoping the downloaded data dumps are valid, and parsing the data in order to analyze or compare it based on some criterion. It is even more difficult to search for datasets based on characteristics that are relevant for machine-processing, such as syntactic conformance and structural properties such as the average outdegree of nodes. What is needed is a uniform representation of the *dataset* and a uniform representation of *dataset descriptions*.

The LOD Laundromat [5] realizes the first: it (re)publishes the largest (collection of) dataset(s) on the Web of Data (over 38 billion triples and counting). Every dataset is published in the same format that is fully conformant with Linked Open Data (LOD) publication standards for machine-processability. The purpose of the LOD Laundromat is to drastically simplify the task of data preprocessing for the data consumer.

However, the creation of meta-data describing the datasets is still left to the original data publisher. We see that many data publishers do not publish a dataset description that can be found by automated means, and that those dataset descriptions that *can* be found do not always contain all (formally or de-facto) standardized meta-data. More importantly, the meta-data values are generally not comparable between datasets since different data publishers may interpret and calculate the same meta-data property differently. For instance, it is not generally the case that a dataset with a higher value for the `void:triples` property contains more triples: this value might be outdated with respect to the original dataset, or it might have been incorrectly calculated. Because of such incompatibilities between existing dataset descriptions, it is difficult to reliably analyze and compare datasets on a large scale.

Therefore, next to the uniform *dataset* representations that are published by the LOD Laundromat, we need the same uniform representation for publishing *dataset meta-data*. In addition to uniformity, even straightforward meta-data should come with provenance annotations that describe how meta-data was generated. The here presented LOD Laundromat meta-dataset brings exactly this: a collection of dataset descriptions, linked to the same canonical dataset representation, all modeled, created, and published in the same manner, and with provenance annotations that explain how the meta-data was generated.

In section 2 we give an overview of comparable datasets. In section 3 we identify shortcomings in existing meta-data standards and collections, and formulate a set of requirements for a dataset that would allow large collections of datasets to be analyzed, compared, and used. Section 4 presents the meta-data we publish, the model that is used to publish in, the used external vocabularies, a discussion in the context of the five stars of Linked Data Vocabulary use, and clarification on how the LOD Laundromat meta-dataset is generated and maintained. Section 5 shows the applications and use cases that the LOD Laundromat meta-dataset offers. We conclude with section 6.

2. Comparable Datasets

SPARQL Endpoint Status[1] [6] presents an overview of dataset descriptions that can be found by automated means. These results show that even the uptake of the core meta-data properties (such as the ones from the VoID [1] specification) is still quite low: only 12.9% of the analyzed SPARQL endpoints are described using VoID. Because of this apparent lack of LOD meta-data, several initiatives tried to fill this gap by creating uniform meta-data descriptions for multiple datasets.

Firstly, LODStats [2] provides statistical information for all Linked Open Datasets that are published in the CKAN-powered[2] Datahub[3] catalog. It offers a wide range of statistics, e.g., including the number of blank nodes in a dataset and the average outdegree of subject terms. Unfortunately, only a small subset of those statistics are themselves being published as Linked Data. Secondly, Sindice [17] provides statistical information similar to LODStats, but mostly analyzes smaller datasets that are crawled from Web pages. The meta-data provided by Sindice are similar to those in the VoID specification but they are not published in a machine-readable format such as RDF.

Although Sindice and LODStats provide a step in the right direction by uniformly creating meta-data descriptions for many Linked Datasets, they only support a subset of existing meta-data properties, they do not publish exhaustive meta-data descriptions as Linked Data, and they do not publish structural information on the meta-data generation procedure. Also, they are constrained to Linked Datasets that are published in only certain locations.

3. Meta-Data Requirements

In this section we present a requirements analysis for a dataset that satisfies our goal of supporting the meaningful analysis, comparison, and use, of very many datasets.

We explain problems with respect to meta-data specifications (section 3.1), dataset descriptions (section 3.2) and collections of dataset descriptions (section 3.3). Based on these considerations, the requirements are presented in section 3.4.

3.1. Meta-data specifications

Existing dataset vocabularies include VoID [1], VoID-ext [15], DCAT[4], and Bio2RDF [7]. VoID is a vocabulary for expressing meta-data about Linked Datasets. It supports generic meta-data (e.g., the homepage of a dataset), access meta-data (e.g., which protocols are available), links to other datasets, exemplary resources, as well as dataset statistics (e.g., the number of triples). Only some of the VoID meta-data properties can be automatically generated. Others can only be given by human authors, –such as exemplary resources– since they depend on interpretation. Bio2RDF presents a collection of dataset meta-data properties that extends the set of VoID properties and provides more detail. For example, Bio2RDF includes

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

[2] `http://ckan.org/`

[3] `http://datahub.io/`

[4] See `http://www.w3.org/TR/vocab-dcat/`

properties that describe how often particular types are used in the subject position and in the object position for a given property; e.g. property `ex:livesIn` links 10 subjects of type *`ex:Person`* to 6 objects of type *`ex:City`*. The use of such descriptive properties can increase the size of a meta-dataset significantly when the described dataset has a large number of classes and properties.

VoID-ext extends the set of meta-data properties that are found in VoID as well. It includes the in- and out-degree of entities, the number of blank nodes, the average string length of literals, and a partitioning of the literals and URIs based on string length. The Data Catalog Vocabulary (DCAT) is a vocabulary for describing datasets on a higher level; i.e., it includes properties such as the dataset title, description and publishing/modification date. Such information is difficult to reliably extract from the dataset in an automated fashion.

We observe the following problems with these existing meta-data specifications:

First, some existing meta-data properties are subjective. For example, `void:entities` is intended to denote a subset of the IRIs of a dataset based on "arbitrary additional requirements" imposed by the authors of the dataset description. Since different authors may impose different requirements, the number of entities of a dataset may vary between zero and the number of resources.

Secondly, some existing meta-data properties are defined in terms of undefined concepts. For example, LODStats specifies the set of vocabularies that are reused by a given dataset. The notion of a 'reused vocabulary' is itself not formally defined but depends on heuristics about whether or not an IRI belongs to another dataset. LODStats calculates this set by using relatively simple string operations according to which IRIs of the form `http://<authority>/<string>/<value>` are assumed to belong to the vocabulary denoted by `http://<authority>/<string>`. Although this is a fair attempt at identifying reused vocabularies, there is not always a bijective map between datasets and URI substrings that occur in datasets. The number of links to other datasets suffers from the same lack of a formal definition.

3.2. Dataset descriptions

We observe the following problems with existing dataset descriptions: First, uptake of dataset descriptions that can be found by automated means is still quite low (section 2). Secondly, for reasons discussed above, the values of meta-data properties that do not have a well-founded definition cannot be meaningfully compared across datasets. E.g., if two dataset descriptions contain different values for the `void:entities` property it is not clear whether this denotes an interesting difference between the two datasets or whether this is due to the authors having different criteria for identifying the set of entities. Thirdly, even the values of well-defined meta-data may have been calculated in different ways by different computational procedures. We observe that there are significant discrepancies between meta-data which occurs *in* the original dataset description and those from the LOD Laundromat. For example, a dataset about a Greek fire brigade contains 3,302,302 triples according to its original VoID description[5], but 4,134,725 triples according to the LOD Laundromat meta-dataset[6].

Similar discrepancies exist between meta-data values that occur in different dataset description *collections*, e.g. between LODStats and the LOD Laundromat meta-dataset.[7]

Since it is difficult to assess whether a computational procedure that generates meta-data is correct, we believe it is necessary that all generated meta-data is annotated with provenance information that describes the used computational procedure. Although relatively verbose, this approach circumvents the arduous discussion of which version of what tool is correct/incorrect for calculating a given meta-data value. We assume that there will always be multiple values for the same meta-data property. The fact that there are different values, and that these have been derived by different means, is something that has to be made transparent to the consumer of this meta-data. The onus is on the data consumer to trust one computational procedure for calculating a specific meta-data value more than another. This requires provenance that details the mechanism behind the calculated meta-data.

[5] See `http://greek-lod.auth.gr/Fire/void.ttl`

[6] See `http://lodlaundromat.org/resource/0ca7054f382b29319c82796a7f9c3899`

[7] E.g., according to LODStats the dataset located at `http://www.open-biomed.org.uk/open-biomed-data/bdgp-images-all-20110211.tar.gz` contains 1,080,060 triples while the LOD Laundromat meta-dataset states 1,070,072.

3.3. Dataset description collections

We observe two problems with existing collections of dataset descriptions: Firstly, even though the meta-data may be calculated consistently within a collection, the computational procedure that is used is not described in a machine-processable format (if at all). This means that values can only be compared within the collection, but not with dataset descriptions external to the collection (e.g. occurring in other collections). Secondly, meta-data that is calculated within existing collections is not always published in a machine-interpretable format (e.g. LODStats).

3.4. Requirements

Based on the above considerations, we formulate the following requirements which allow multiple datasets to be meaningfully compared based on their meta-data:

1. The LOD Laundromat meta-dataset must cover very many datasets in order to improve data comparability.
2. The meta-dataset should reuse official and de-facto meta-data standards as much as possible, in order to be compatible with other dataset descriptions and to promote reuse.
3. The meta-dataset must be generated algorithmically in order to assure that values are calculated in the same way for every described dataset.
4. Only those meta-data properties must be used that can be calculated efficiently, because datasets can have peculiar properties that may not have been anticipated when the meta-data properties were first defined.
5. The LOD Laundromat meta-dataset must contain provenance annotations that explain how and when the meta-data was calculated.
6. The meta-data must be disseminated as LOD and must be accessible via a SPARQL endpoint.
7. The LOD Laundromat meta-dataset must be able to support a wide range of real-world use cases that involve analyzing and/or comparing datasets such as Big Data algorithms that process LOD.

4. The LOD Laundromat meta-dataset

In this section we present the meta-data we publish, the model we use, and how we generate this dataset.

4.1. Published Meta-Data

The LOD Laundromat meta-dataset is generated in adherence to the requirements formulated in section 3. Since there are multiple ways in which these requirements can be prioritized and made concrete, we will now discuss the considerations that have guided the generation of the meta-data.

Firstly, there is a trade-off between requirements 2 and 3: since the meta-dataset has to be constructed algorithmically, only well-defined meta-data properties can be included.

Secondly, there is a conflict between requirements 1 and 4 on the one hand, and requirement 2 on the other: since the LOD Laundromat meta-dataset must describe many datasets, some of which are relatively large, and we want calculations to be efficient, we chose to narrow down the set of meta-data properties to those that can be calculated by *streaming* the described datasets. This excludes properties that require loading (large parts of) a dataset into memory, e.g. in order to perform joins on triples.

Thirdly, because of the scale at which the LOD Laundromat meta-dataset describes datasets, it is inevitable that some datasets will have atypical properties. This includes datasets with extremely long literals, datasets where the number of unique predicate terms is close to the total number of predicate terms, or datasets where the number of unique literal datatype equals the total number of literals. It is only when meta-data is systematically generated on a large scale, that one finds such corner cases. These corner cases can make dataset descriptions impractically large. This is especially true for meta-data properties that consist of enumerations. E.g., for some datasets the partition of all properties, as defined by VoID-ext and Bio2RDF, is only (roughly) a factor 3 smaller than the described dataset itself (and this is only one meta-data property). Or, take as example the `void-ext:subjectPartition`, that refers to a partition that contains triples for a certain subject. Using such partitions for all the subjects in a dataset would generate a meta-dataset that equals the size of the original dataset. Therefore, in order to keep data descriptions relatively small w.r.t. the dataset described, the meta-dataset does not include properties whose values are dataset partitions.

Under these restrictions, the meta-dataset is able to include a large number of datasets while still being relatively efficient to construct. Implementation-wise, the generation of the meta-dataset takes into ac-count the many advantages that come from the way in which LOD Laundromat (re)publishes datasets. LOD Laundromat allows datasets to be opened as gzip-compressed streams of lexicographically sorted N-Triples and N-Quads. Since these streams are guaranteed to contain no syntax error nor any duplicate occurrences of triples, they can be processed on a line-by-line / triple-by-triple basis, making it convenient to generate meta-data for inclusion in the LOD Laundromat meta-dataset. Because of these advantages, the meta-data server (with 5TB SSD Disk space, 8-core CPU and 256GB memory) manages to stream and analyze 400.000 triples per second.

Table 1 gives an overview of the meta-data properties included in the LOD Laundromat meta-dataset, together with those that are included in existing dataset description standards. As can be seen from the table, the only meta-data properties that are excluded from our dataset (because of computational issues) are the distinct number of classes that occur in either the subject, predicate, or object position, as specified in VoID-ext. These three meta-data properties cannot be calculated by streaming the data a single time. In addition, all meta-data properties whose values must be represented as partitions are excluded in order to preserve brevity for all dataset descriptions, and to maintain scalability. Considering these limitations, the meta-data properties presented in Bio2RDF are similar to those in VoID and VoID-ext. Therefore, Bio2RDF is not referenced in our vocabulary. The generation of several statistics (e.g. the distinct number of URIs) requires in-memory lists. To reduce this memory consumption, we use an efficient in-memory dictionary (RDF Vault [3]).

Since we want the LOD Laundromat meta-dataset to be maximally useful for a wide range of use cases (requirement 7), we have added several meta-data properties that do not occur in existing specifications:

1. Next to the number of distinct IRIs, blank nodes and literals (i.e., *types*), we also include the number of (possibly non-distinct) occurrences (i.e., *tokens*).
2. Existing vocabularies specify the number of properties and classes (although they do so incorrectly, see section 3). The meta-dataset also includes the number of classes and properties that are *defined* in a dataset, such as `<prop> rdf:type rdf:Property`
3. Existing dataset description vocabularies such as VoID-ext use arithmetic means to describe number series such as the literal lengths in given document. The LOD Laundromat meta-dataset uses more detailed descriptive statistics, that include the median, minimum, maximum and standard deviation values as well.
4. Similar statistics are provided for network characteristics such as Degree, In Degree and Out Degree.

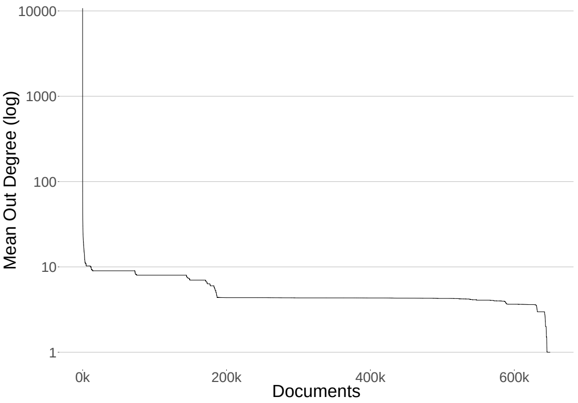

Fig. 1. Average out degree distribution of LOD Laundromat documents

Considering that only 0.5% of the datasets publish a corresponding dataset license via RDF, we exclude this information for now. We expect these dataset licenses to increase in use and popularity though, and will include this meta-data in a future crawl.

Figure 1 illustrates one of the published meta-data properties: the average out degree of datasets. The figure illustrates our previous remark that analyzing many datasets will inevitably include datasets with atypical properties or 'corner cases'. E.g., the dataset with the highest average out degree, contains 10.004 triples, and only one subject, thereby strongly skewing the dataset distribution. Such a-typical properties of datasets are potentially important as e.g. a means of explaining deviating evaluation results between datasets. Note, that generating the data behind this figure requires the following SPARQL query, illustrating the ease of use:

```
SELECT * {[] llm:outDegree/llm:mean ?mean}
```

Besides publishing the meta-data, and in line with requirement 5, the meta-dataset contains a provenance trail of how the meta-data was generated. The provenance trail includes a reference to the code that was used to generate the meta-data. For this we use a Git commit identifier in order to uniquely identify the exact version that was used. The provenance trail also in-

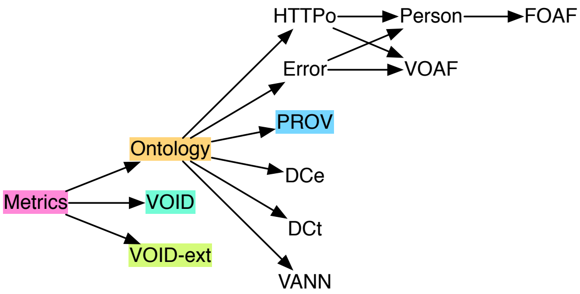

Fig. 2. Dependencies of LOD Laundromat meta-dataset vocabulary

cludes all the steps that preceded the calculation of the meta-data:

1. Where the file was downloaded (either the original URL or the archive that contained the file).
2. When the file was downloaded (date and time).
3. Meta-data on the download process, such as the status code and headers from the original HTTP reply. For archived data the applied compression techniques (possibly multiple ones) are enumerated as well.
4. Detailed meta-data on the data preparation tasks performed by the LOD Laundromat in order to clean the data. This includes the number of bytes that were read (not necessarily the same as the value for `Content-Length` HTTP header) and syntax errors that were encountered (e.g., malformed syntax, unrecognized encoding, undefined prefixes).
5. The number of duplicate triples in the original dataset.
6. A reference to the online location where the cleaned file is stored, and from which the meta-data is derived.

4.2. Model

The meta-data is specified in the LOD Laundromat meta-dataset[8]. Of the 26 meta-data properties that are included, 22 are linked to one or more other dataset description vocabularies. Figure 2 shows the dependencies between our meta-dataset vocabulary and other vocabularies. The referenced dataset description vocabularies are VoID and VoID-ext. Figure 3 shows an example dataset description that illustrates the structure of this meta-dataset[9]. The meta-dataset also includes information about the vocabulary *itself*, such as its license (Creative Commons[10]), last modification date, creators, and homepage. As such, it implements the first 4 of the 5 stars for vocabulary re-use [12]. The fifth star (re-use *by* other vocabularies) is not reached yet because the vocabulary is quite recent. However, the LOD Laundromat meta-dataset has been submitted to the Linked Open Vocabulary catalog [11], thereby hopefully supporting its re-use and findability.

The provenance information of datasets is described using the PROV-O vocabulary [14], a W3C recommendation. Figure 4 presents an overview on how PROV-O is used by the LOD Laundromat meta-dataset. Similar vocabularies exist, such as the VoiDp [16] vocabulary which matches the provenance of Linked Datasets with the VoID vocabulary. However, because VoiDp uses a predecessor of the PROV-O standard, we model our provenance in PROV-O directly. The Provenance Vocabulary [9] aims to describe the provenance of Linked Datasets as well, but is too specific for our use considering the wide range of provenance (see below) we describe.

As the LOD Laundromat cleaning process is part of the provenance trail, we model this part of the dataset using separate vocabularies: Firstly, the LOD Laundromat vocabulary[12] describes the crawling and cleaning process of LOD Laundromat. This description includes the download time and date of the original document, and therefore specifies which version of the original document is described by the meta-dataset. Secondly, the HTTP vocabulary[13] describes HTTP status codes. Thirdly, the error ontology[14] models all exceptions and warnings, and is used by the LOD Laundromat vocabulary to represent errors that occur during the crawling and cleaning process. Each of these vocabularies are linked to other vocabularies. E.g., the HTTP vocabulary is an extension of the W3C HTTP in RDF vocabulary[15].

[8] See `http://lodlaundromat.org/metrics/ontology/`

[9] For brevity, only a subset of the available meta-data properties are included in this figure

[10] See `http://creativecommons.org/licenses/by/3.0/`

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

[12] `http://lodlaundromat.org/ontology/`

[13] `http://lodlaundromat.org/http/ontology/`

[14] `http://lodlaundromat.org/errors/ontology/`

[15] `http://www.w3.org/2011/http`

Table 1

An overview of dataset meta-data properties, grouped by the vocabularies that define them and dataset description collections that include them. For brevity's sake, properties whose values are dataset partitions and properties that require manual intervention are excluded

Meta-data Property	**VoID**	**Bio2RDF**	**VoID-ext**	**LOD Laundromat**	*DataType / Range*
Triples	v	v	v	v	`xsd:integer`
Entities	v	v	v	v	`xsd:integer`
Distinct Classes	v	v	v	v	`xsd:integer`
Distinct Properties	v	v	v	v	`xsd:integer`
Distinct Subject	v	v	v	v	`xsd:integer`
Distinct Objects	v	v	v	v	`xsd:integer`
Distinct RDF Nodes			v	v	`xsd:integer`
Distinct IRIs			v	v	`xsd:integer`
IRIs				v	`xsd:integer`
Distinct Blank Nodes			v	v	`xsd:integer`
Blank Nodes				v	`xsd:integer`
Distinct Literals	v		v	v	`xsd:integer`
Literals				v	`xsd:integer`
Distinct URIs in subject position			v	v	`xsd:integer`
Distinct Blank Nodes in subject position			v	v	`xsd:integer`
Distinct URIs in object position			v	v	`xsd:integer`
Distinct Blank Nodes in object position			v	v	`xsd:integer`
Distinct Literal Data-Types			v	v	`xsd:integer`
Distinct Literal Languages			v	v	`xsd:integer`
Length statistics of IRIs			v	v	`xsd:integer`
Length statistics of IRIs in subject position			v	v	`llm:DescriptiveStatistics`
Length statistics of IRIs in predicate position			v	v	`llm:DescriptiveStatistics`
Length statistics of IRIs in object position			v	v	`llm:DescriptiveStatistics`
Length statistics of Literals			v	v	`llm:DescriptiveStatistics`
Degree Statistics				v	`llm:DescriptiveStatistics`
Indegree Statistics				v	`llm:DescriptiveStatistics`
Outdegree Statistics				v	`llm:DescriptiveStatistics`
Defined Classes				v	`xsd:integer`
Defined Properties				v	`xsd:integer`
Distinct Classes occurring in the subject position			v		
Distinct Classes occurring in the predicate position			v		
Distinct Classes occurring in the object position			v		

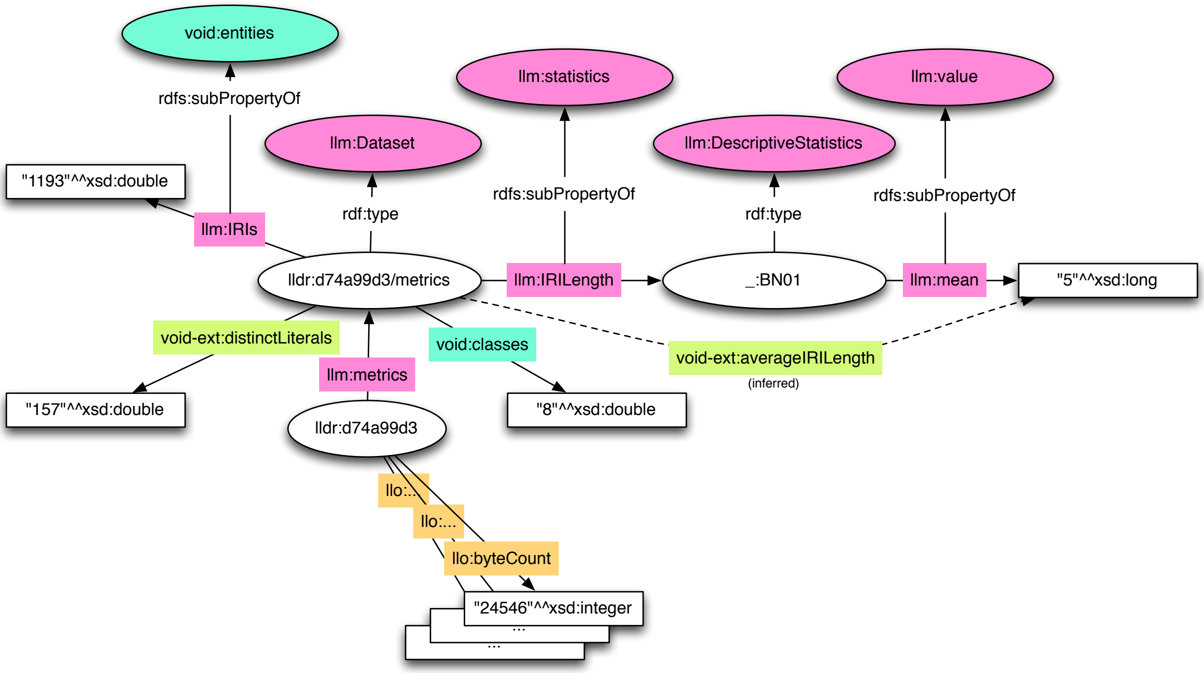

Fig. 3. Example (partial) dataset meta-data description, color-coded using vocabularies from Figure 2

4.3. Naming Scheme

The LOD Laundromat meta-dataset uses the following naming scheme. As a running example, we take a Semantic Web Dog Food file that is crawled by LOD Laundromat[16].

- The LOD Laundromat document identifier for this dataset is generated by appending an MD5 hash of the data source IRI to `http://lodlaundromat.org/resource/`[17].
- The calculated structural properties of this dataset are accessible by appending `/metrics` to the LOD Laundromat document identifier[18].
- Provenance that describes the procedure behind the metrics calculation is accessible by appending `metricCalculation` to the LOD Laundromat identifier[19].

4.4. Dissemination

The LOD Laundromat [5] continuously crawls and analyses Linked Data dumps. In order to get a maximum coverage of the LOD Cloud, it searches both linked data catalogs and the LOD Laundromat datasets themselves for references to datadumps. Because it does not claim to have a complete seed list that links to all LOD in the world, users have the option to manually or algorithmically add seed-points to the LOD Laundry Basket[20].

The code[21] used to generate the LOD Laundromat meta-dataset runs immediately after a document is crawled and cleaned by the LOD Laundromat, and is directly published via a public SPARQL endpoint[22]. SPARQL is preferred over HDT as publishing method, because HDT files are static and do not support updates. In line with requirement 6, a nightly version of the meta-dataset is extracted from the SPARQL endpoint and published as data dump, in the same standardized N-Quad serialization format of the LOD Laundromat.

[16] See `http://data.semanticweb.org/dumps/conferences/iswc-2013-complete.rdf`
[17] See `http://lodlaundromat.org/resource/05c4972cf9b5ccc346017126641c2913`
[18] See `http://lodlaundromat.org/resource/05c4972cf9b5ccc346017126641c2913/metrics`
[19] See `http://lodlaundromat.org/resource/05c4972cf9b5ccc346017126641c2913/metricCalculation`
[20] `http://lodlaundromat.org/basket/`
[21] Publicly available at `https://github.com/LODLaundry/LODAnalysis`
[22] `http://lodlaundromat.org/sparql`

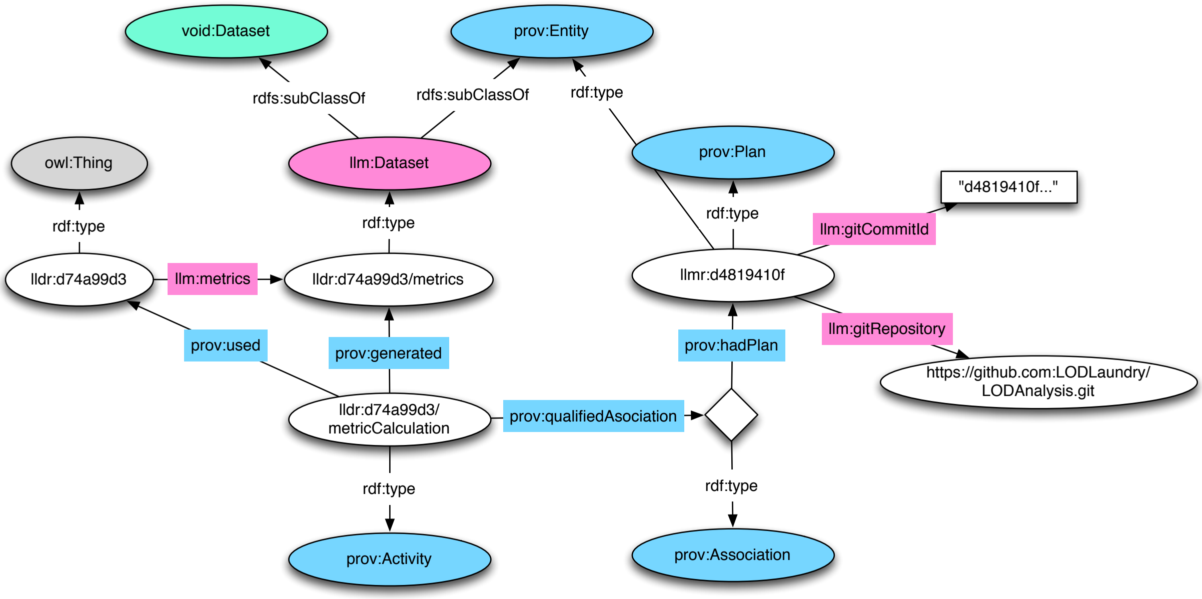

Fig. 4. Provenance model illustration

4.5. *Dataset Statistics*

Since the release of LOD Laundromat in September 2014 and the release of the meta-dataset in January 2015, we registered 2,119,218 document downloads, and 20,606,194 SPARQL queries on the meta-dataset. As mentioned before, the LOD Laundromat crawled and re-publishes over 650,000 documents containing over 38,000,000 triples. The meta-data of these crawled documents are published in the LOD Laundromat meta-dataset, and now contains over 110 million triples, accessible via a data dump and SPARQL endpoint.

5. Use Cases

The LOD Laundromat meta-dataset is intended to support a wide array of non-trivial use cases. The first use case we present is the evaluation of Semantic Web (SW) algorithms. In contemporary SW research novel algorithms are usually evaluated against only a handful of – often the same – datasets (i.e., mainly DBpedia, Freebase, and Billion Triple Challenge). The risk of this practice is that – over time – SW algorithms will be optimized for datasets with specific distributions, but not for others. In [18], we re-evaluate parts of three SW research papers using `Frank` [4], a bash interface interfacing with the LOD Laundromat. We showed how the LOD Laundromat meta-dataset can be used to relate datasets to their overall structural properties, and how SW evaluations can be performed on a much wider scale, leading to results that are more indicative of the *entire* LOD Cloud. For example, the re-evaluation of RDF HDT [8] (a binary compressed representation for RDF) showed a –previously unknown– relation between the degree of datasets and the RDF HDT compression ratio. This use case combines the strength of both the LOD Laundromat collection of documents and the LOD Laundromat meta-dataset. The following SPARQL query was used in the RDF HDT re-evaluation to find documents with a low average out degree[23]:

```
SELECT * WHERE {
  ?datadoc llm:metrics
          /llm:outDegree
          /llm:mean ?outDegree .
  FILTER(?outDegree < 5)
}
```

Similarly to *evaluating* SW algorithms, the LOD Laundromat meta-dataset can also be used to *tune* these SW algorithms or prune datasets with the desired

[23] For brevity, the presented queries do not contain the following LOD Laundromat namespaces:
```
PREFIX llm <http://lodlaundromat.org/metrics/ontology/>
PREFIX ll <http://lodlaundromat.org/ontology/>
```

property at an early stage, i.e., without having to load and interpret them. An example of this is PrefLabel[24], an online service that returns a human-readable label for a given resource-denoting IRI. The index behind the PrefLabel Web service is populated by streaming and analyzing LOD Laundromat datasets for RDFS label statements in datasets. PrefLabel uses the LOD Laundromat meta-dataset by pruning for datasets that do not contain RDF literals at all. This crude way of using the meta-dataset already excludes 20% of all the triples that are in the LOD Laundromat today, thereby significantly optimizing the algorithm. The following SPARQL query is used by PrefLabel to prune the list of documents:

```
SELECT ?doc WHERE {
  ?doc llm:metrics
      /llm:literals ?lit;
  FILTER(?lit = 0)
}
```

Another use case involves using the LOD Laundromat meta-dataset to analyze and compare datasets, e.g., in order to create an overview of the state of the LOD Cloud at a given moment in time. A common approach (see e.g. [10,13,19]) is to crawl Linked Data via dereferenceable URIs using tools such as LDspider [11], and/or to use catalogs such as datahub to discover the Linked Datasets. Both dereferenceable URIs and dataset catalogs come with limitations: most Linked Data URIs are not dereferenceable, and the dataset catalogs only cover a subset of the LOD Cloud. The LOD Laundromat on the other hand provides access to more than dereferenceable URIs only, and aims to provide a complete as possible dataset collection. The corresponding meta-dataset provides a starting point for e.g. finding datasets by Top Level Domain, serialization format, or structural properties such as number of triples. In [18] we re-evaluate (part of) exactly such a Linked Data Observatory paper [19], where we use the meta-dataset and LOD Laundromat to find the documents and extract namespace statistics.

Next to the *structural* meta-data properties, the *provenance* meta-data provides an interesting data source as well. Such provenance enables e.g. an analysis of common RDF serialization formats, as shown in figure 5. The following SPARQL query fetches the serialization information used by this figure:

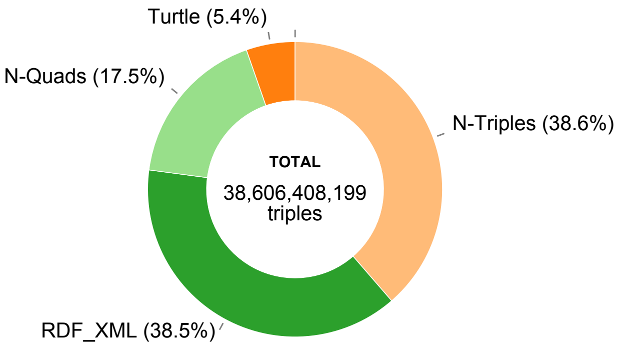

Fig. 5. Serialization formats

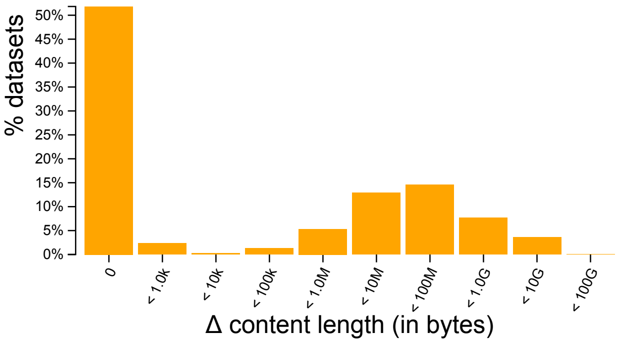

Fig. 6. Invalid HTTP Content Lengths

```
SELECT ?format (SUM(?t) AS ?count)
WHERE {
[] llo:serializationFormat ?format;
   llo:triples ?t .
}
GROUP BY ?format
```

The provenance information can be used to measure publishing best practices as well, such as whether the content length specified by the HTTP response matches the actual content length of the file. This is visualized in figure 5, which uses the following SPARQL query to fetch the data:

```
SELECT ?clength ?bcount ?t WHERE {
  [] llo:contentLength ?clength ;
     llo:byteCount ?bcount ;
     llo:triples ?t .
}
```

6. Conclusion

The dataset presented in this paper offers access to a large set of uniformly represented datasets descrip-

[24] http://preflabel.org

tions, acting as an enabler for large scale Linked Data research: finding or comparing linked datasets with certain structural properties is now as easy as executing a SPARQL query. And even better: because the dataset descriptions are linked to their uniform dataset representations, the access to the underlying data is extremely easy as well.

We are exploring the possibilities of storing snapshots of both the meta-dataset and the corresponding cleaned datasets, effectively creating snapshots of the state of the LOD Cloud. At this point, we consider this future work though.

Another future improvement we consider is to publish partitions of the datasets via more scalable and efficient ways than SPARQL. As explained in section 4.1, corner-cases in the LOD cloud can drastically increase the corresponding meta-data. Therefore, an efficient and scalable method is required for hosting such partitions. We consider publishing a selection of such partitions using non-SPARQL APIs with a stronger focus on scalability and efficiency.

Acknowledgements

This work was supported by the Dutch national program COMMIT.

References

[1] Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing Linked Datasets. In *Proceedings of the Linked Data on the Web Workshop (LDOW2009)*, 2009.

[2] Sören Auer, Jan Demter, Michael Martin, and Jens Lehmann. LODStats–an extensible framework for high-performance dataset analytics. In *Knowledge Engineering and Knowledge Management*, pages 353–362. Springer, 2012.

[3] Hamid R Bazoobandi, Steven de Rooij, Jacopo Urbani, Annette ten Teije, Frank van Harmelen, and Henri Bal. A Compact In-Memory Dictionary for RDF data. In *The Extended Semantic Web Conference – ESWC*, pages 205–220. Springer, 2015.

[4] Wouter Beek and Laurens Rietveld. Frank: The LOD Cloud at your Fingertips . In *Developers Workshop, The Extended Semantic Web Conference (ESWC)*, 2015.

[5] Wouter Beek, Laurens Rietveld, Hamid R Bazoobandi, Jan Wielemaker, and Stefan Schlobach. LOD Laundromat: A Uniform Way of Publishing Other People's Dirty Data. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 213–228. Springer, 2014.

[6] Carlos Buil-Aranda, Aidan Hogan, Jürgen Umbrich, and Pierre-Yves Vandenbussche. SPARQL Web-Querying Infrastructure: Ready for Action? In *The International Semantic Web Conference*, pages 277–293. Springer, 2013.

[7] Alison Callahan, José Cruz-Toledo, Peter Ansell, and Michel Dumontier. Bio2rdf Release 2: Improved Coverage, Interoperability and Provenance of Life Science Linked Data . In *The Semantic Web: Semantics and Big Data*, pages 200–212. Springer, 2013.

[8] Javier D Fernández, Miguel A Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. Binary RDF representation for publication and exchange (HDT). *Web Semantics: Science, Services and Agents on the World Wide Web*, 19:22–41, 2013.

[9] Olaf Hartig and Jun Zhao. Publishing and Consuming Provenance Metadata on the Web of Linked Data. In *Provenance and annotation of data and processes*, pages 78–90. Springer, 2010.

[10] Tom Heath and Christian Bizer. Linked data: Evolving the web into a global data space. *Synthesis lectures on the semantic web: theory and technology*, 1(1):1–136, 2011.

[11] Robert Isele, Jürgen Umbrich, Christian Bizer, and Andreas Harth. LDspider: An Open-Source Crawling Framework for the Web of Linked Data. In *Posters & Demos, 9th International Semantic Web Conference (ISWC2010)*, 2010.

[12] Krzysztof Janowicz, Pascal Hitzler, Benjamin Adams, Dave Kolas, and Charles Vardeman II. Five stars of Linked Data vocabulary use. *Semantic Web*, 5(3):173–176, 2014.

[13] Tobias Käfer, Jürgen Umbrich, Aidan Hogan, and Axel Polleres. Towards a Dynamic Linked Data Observatory. *Linked Data on the Web Workshop (LDOW2012)*, 2012.

[14] Timothy Lebo, Satya Sahoo, Deborah McGuinness, Khalid Belhajjame, James Cheney, David Corsar, Daniel Garijo, Stian Soiland-Reyes, Stephan Zednik, and Jun Zhao. Prov-o: The Prov Ontology. *W3C Recommendation, 30th April*, 2013.

[15] Eetu Mäkelä. Aether–Generating and Viewing Extended VoID Statistical Descriptions of RDF Datasets . In *The Semantic Web: ESWC 2014 Satellite Events*, volume 8465, pages 429–433. Springer, 2014.

[16] Tope Omitola, Landong Zuo, Christopher Gutteridge, Ian C Millard, Hugh Glaser, Nicholas Gibbins, and Nigel Shadbolt. Tracing the provenance of linked data using voiD. In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, page 17. ACM, 2011.

[17] Eyal Oren, Renaud Delbru, Michele Catasta, Richard Cyganiak, Holger Stenzhorn, and Giovanni Tummarello. Sindice.com: a Document-Oriented Lookup Index for Open Linked Data . *International Journal of Metadata, Semantics and Ontologies*, 3(1):37–52, 2008.

[18] Laurens Rietveld, Wouter Beek, and Stefan Schlobach. LOD Lab: Experiments at LOD Scale. In *Proceedings of the International Semantic Web Conference (ISWC)*. Springer, 2015.

[19] Max Schmachtenberg, Christian Bizer, and Heiko Paulheim. Adoption of the Linked Data Best Practices in Different Topical Domains. In *Proceedings of the International Semantic Web Conference (ISWC)*, pages 245–260. Springer, 2014.