

LODStats: The Data Web Census Dataset

Ivan Ermilov¹, Jens Lehmann², Michael Martin¹, and Sören Auer²

¹ AKSW, Institute of Computer Science, University of Leipzig, Leipzig, Germany
{iermilov, martin}@informatik.uni-leipzig.de

² University of Bonn and Fraunhofer IAIS, Bonn, Germany
{jens.lehmann, auer}@cs.uni-bonn.de

Abstract. Over the past years, the size of the Data Web has increased significantly, which makes obtaining general insights into its growth and structure both more challenging and more desirable. The lack of such insights hinders important data management tasks such as quality, privacy and coverage analysis. In this paper, we present the LODStats dataset, which provides a comprehensive picture of the current state of a significant part of the Data Web. LODStats is based on RDF datasets from *data.gov*, *publicdata.eu* and *datahub.io* data catalogs and at the time of writing lists over 9000 RDF datasets. For each RDF dataset, LODStats collects comprehensive statistics and makes these available in adhering to the *LDSO* vocabulary. This analysis has been regularly published and enhanced over the past five years at the public platform lodstats.aksw.org. We give a comprehensive overview over the resulting dataset.

1 Introduction

Over the past years, the size of the Data Web has increased significantly, which makes obtaining general insights into its growth and structure both more challenging and more desirable. The expansion of the Data Web can be to a large extent attributed to the efforts in the Semantic Web and Open Government communities. Both communities have a common goal: to provide 5-star³ RDF datasets to end-users. To achieve this goal, the Semantic Web community introduced a number of requirements for datasets, which should be fulfilled to be included into the *LOD Cloud*⁴. The Semantic Web community has a main dataset registry hub: the *datahub*⁵ data catalog, while Open Government initiatives usually distribute RDF datasets through their own data catalogs (e.g. *data.gov*, *publicdata.eu* and *open.canada.ca*).

All of the mentioned data catalogs utilize CKAN, an open-source data portal platform, which is a de-facto standard for Open Data. CKAN provides a solid framework to organize datasets and to expose metadata about them in various formats, including RDF. However, CKAN does not provide analytics over the

³ According to 5-star data model available at <http://5stardata.info>

⁴ <http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/>

⁵ <http://datahub.io/>

registered datasets and highly depends on the user input. Moreover, no single aggregation point exists. These factors limit the possibility to obtain general insights into the Data Web. The lack of such insights hinders important data management tasks such as quality, privacy and coverage analysis.

For this reason, attempts to analyze the Data Web were made previously. *SPARQL Endpoint Status*⁶ (SPARQLES) [4] addresses the problem of the availability of SPARQL endpoints over time. SPARQLES aggregates 553 SPARQL endpoints and exposes information on the availability and their features (e.g. support for SPARQL 1.0/1.1, availability of VoID/Service descriptions). *Linked Open Vocabularies*⁷ (LOV) [6] is a project for building an RDF vocabulary ecosystem, which can support reuse of vocabulary terms. LOV aggregates the vocabularies from various publishers and establish relationships between them using the VOAF vocabulary. The project collected 548 vocabularies (e.g. DCMI Metadata Terms, Friend of a Friend and others) and enabled vocabulary search by utilizing metrics derived from the analysis of the vocabularies and their relationships. The *vocab.cc* project attempted to fill the gap of vocabulary usage statistics. Being based on the Billion Triples Challenge (BTC) in 2012, vocab.cc introduced four metrics to evaluate the BTC dataset. However, the project has a limited scope (i.e. being restricted to the BTC dataset) and was a one-shot evaluation, and therefore does not provide sustainable statistics over time.

In this paper, we address the above-described gap in the Data Web analysis. We present the *LODStats* dataset, which provides a comprehensive picture of the current state of a significant part of the Data Web. At the time of writing, LODStats aggregates 9 960 RDF datasets from the data.gov, publicdata.eu and datahub.io data catalogs. For each RDF dataset, LODStats collects comprehensive statistics adhering to the RDF data model. This analysis has been regularly published and enhanced over the past five years at <http://lodstats.aksw.org>. We extend our previous work [3,5] as follows: (i) we include data.gov and publicdata.eu data catalogs, which account for 45% of the RDF datasets (ii) we publish the *LDSO* vocabulary, describing the LODStats data schema and (iii) we enrich the dataset with CKAN metadata. Overall, our contributions are as follows:

- We provide a 5-star RDF dataset containing statistical facts about the Data Web, which is interlinked with CKAN metadata.
- We showcase the usage of the dataset via five use case descriptions.
- We describe insights in the Data Web gained from the analysis of LODStats dataset.
- We maintain LODStats over the past five years, delivering sustainable solution to the Semantic Web community.

The rest of the paper is structured as follows: in section 2 we introduce the LODStats web application, section 3 outlines the design of the LODStats dataset, in section 4 we describe use cases supported by the dataset, section 5 exhibits the

⁶ <http://sparqles.ai.wu.ac.at/>

⁷ <http://lov.okfn.org/>

interfaces to access the dataset, we discuss the insights of the Data Web analysis in section 6, and finally conclude and outline future work in section 7.

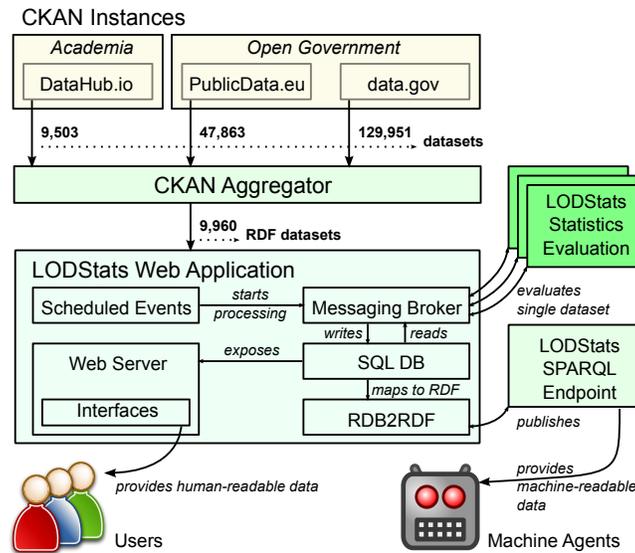
2 LODStats: Web Scale RDF Data Analytics

In this section, we briefly outline the inner workings of the LODStats application and show the evolution of the technical solution.

The general overview of the LODStats architecture is depicted in Figure 1. The *LODStats Statistics Evaluation* (LSE) module performs the execution of the statistical metrics on a dataset and is described in more detail in previous work [3,5].⁸ In this paper, we introduce the following new modules. To aggregate the datasets from the data catalogs we implemented the *CKAN Aggregator*⁹. The *Messaging Broker*¹⁰ allows to schedule processing and scale it horizontally (i.e. to distribute datasets processing between LSE modules running in parallel).

We provide interfaces both for human users and machine agents. The *RDB2RDF*¹¹ module provides virtual RDF views accessible through the *LODStats SPARQL Endpoint* for the consumption of machine agents. For human users, a web frontend is available at <http://lodstats.aksw.org>.

Moreover, we provide Docker image of the whole system publicly.¹² With LODStats Docker image, the application can be deployed on any Docker-enabled host with one command, namely `docker-compose up -d`.



⁸ For SPARQL endpoints, see <http://lodstats.aksw.org> for a complete list of endpoints.

⁹ <https://github.com/aksw/ckan-aggregator-py>

¹⁰ We use rabbitmq as a messaging broker <https://www.rabbitmq.com/>

¹¹ For RDB2RDF transformation we utilize Sparqlify <http://sparqlify.org/>

¹² <https://github.com/aksw/lodstats.docker>

3 Dataset Modelling

In this section, we describe the `LODStats Dataset Vocabulary (LDSO)`¹³, depicted in Figure 2. We designed LDSO as an extension of the *Data Catalog Vocabulary (DCAT)* [7] and *Vocabulary of Interlinked Datasets (VoID)* [1] according to the best practices of the vocabulary design, preservation and governance described in [6,2]. In the following, we describe the structure of the vocabulary.

The `ldso:Dataset` class is a representation of a dataset from a CKAN data catalog. Thus, to model `ldso:Dataset` we extend `dc:Dataset` by adding the `ldso:active` property and reusing general metadata properties such as `dc:identifier` and `dc:modified`. `ldso:active` is a boolean property, which separates up-to-date (i.e. existing in the CKAN data catalog) and out-dated datasets. `ldso:Dataset` connects to the data.gov, publicdata.eu and datahub.io data portals (`ldso:Ckan-Catalog`) via the `dc:isPartOf` property. Also, we interlink instances of `ldso:Dataset` to the corresponding RDF representations in the data portals using `owl:sameAs`. To process a `ldso:Dataset`, the LODStats application utilizes the value of the `dc:downloadURL` property to retrieve dumps. Subsequently, a `ldso:Dataset` is linked directly to the last evaluation result via `ldso:currentStats`. The modelling of `ldso:Dataset` instances, for example, supports the following queries: (i) *How many RDF datasets are in a particular CKAN data catalog?*, (ii) *What is the ratio between out-dated and up-to-date datasets?*, (iii) *Who is the dataset maintainer and what is her email address?*

A `ldso:StatResult` represents a single evaluation result for a `ldso:Dataset`. `ldso:StatResult` extends `void:Dataset` by adding set of statistical metrics in the LDSO namespace such as `ldso:literals`, `ldso:blanks`, `ldso:subclasses`. We connect `ldso:StatResult` to `ldso:Dataset` using the `foaf:primaryTopic` property. The VoID vocabulary introduces the concept of property and class partitions, which represent the subsets of a dataset utilizing particular properties/classes. We extend this design pattern by introducing new partitions, based on datatypes, vocabularies and languages. We interlink `ldso:StatResult` instances to the VoID description of the datasets, generated automatically on dataset evaluation. The modelling of `ldso:StatResult` allows, for example, the following queries: (i) *How many triples (literals, blanks, subclasses) are contained in the dataset?*, (ii) *How many triples in the dataset are adhering to the particular vocabulary (language, datatype)?*, (iii) *What is the size of the dataset dump (in bytes)?*

4 Relevance of the Dataset

Obtaining *comprehensive* statistical analysis about datasets made available on the Web of Data facilitates a number of important use cases (UC) and provides crucial benefits. These include:

Vocabulary Reuse (UC1). One of the advantages of semantic technologies is to simplify data integration via common vocabularies. However, it is often difficult

¹³ LDSO is published at <http://lodstats.aksw.org/ontology/ldso.owl>

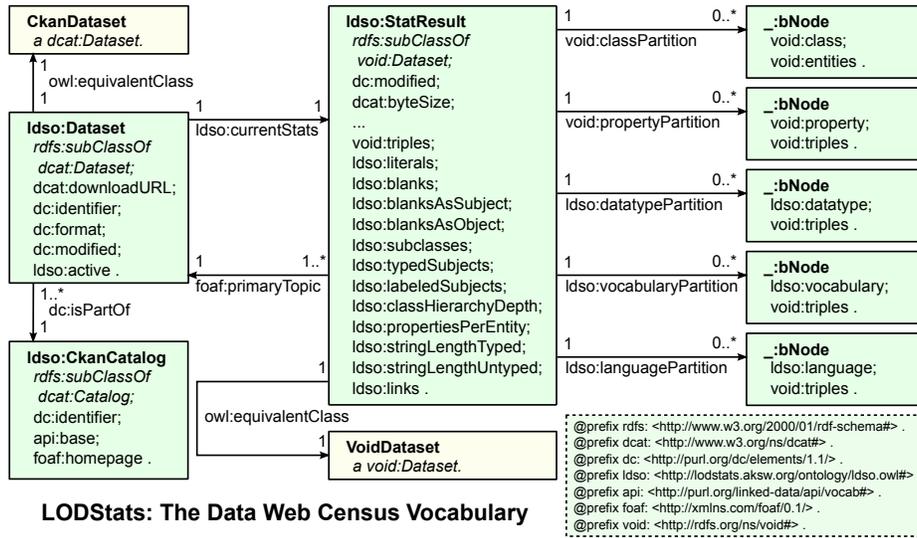


Fig. 2. LODStats Vocabulary Schema.

to identify relevant vocabulary elements. The LODStats web interface stores the usage frequency of vocabulary elements and provides search functionality. This allows knowledge engineers to find the most frequent schema elements, which can be used to model the task at hand. Having this functionality encourages reuse of schema elements and, therefore, simplifies data integration, which is one of the central advantages of semantic technologies. LODStats also provides a webservice for this functionality, such that third party tools can easily integrate search for similar classes and properties. For instance, Linked Open Vocabularies (LOV)¹⁴ utilizes vocabulary usage frequency as an indicator showing the users popularity of specific vocabulary inside the Linked Open Vocabularies catalogue.

Quality analysis (UC2). A major problem when using Web Data is quality. However, the quality of the datasets itself is not so much a problem as assessing and evaluating the expected quality and deciding whether it is sufficient for a certain application. Also, on the traditional Web we have very varying quality, but means were established (e.g. page rank) to assess the quality of information on the document web. In order to establish similar measures on the Web of Data it is crucial to assess datasets with regard to incoming and outgoing links, but also regarding the used vocabularies, properties, adherence to property range restrictions, their values etc. Hence, a statistical analysis of datasets can provide important insights with regard to the expectable quality.

Coverage analysis (UC3). Similarly important as quality is the coverage a certain dataset provides. LODStats can be used to compute several coverage dimensions. For instance, the most frequent properties for a particular dataset

¹⁴ <http://lov.okfn.org>

can be computed and allow to get an overview over instance data, e.g. whether it contains address information. Furthermore, the frequency of namespaces may also be an indicator for the domain of a dataset. The ranges of properties can give insights on whether spatial or temporal information is present in the dataset. In the case of spatial data, for example, we would like to know the region the dataset covers, which can be easily derived from minimum, maximum and average of longitude and latitude properties.

Privacy analysis (UC4). For quickly deciding whether a dataset potentially containing personal information can be published on the Data Web, we need to get a swift overview on the information contained in the dataset without looking at every individual data record. An analysis and summary of all the properties and classes used in a dataset can quickly reveal the type of information and thus prevent the violation of privacy rules.

Link target identification (UC5). Establishing links between datasets is a fundamental requirement for many Linked Data applications (e.g. data integration and fusion). However, as we learned the Web of Linked Data currently still lacks coherence (with less than 10% of the entities actually being linked). Meanwhile, there are a number of tools available which support the automatic generation of links (e.g. [8,9]). An obstacle for the broad use of these tools is, however, the difficulty to identify suitable link targets on the Data Web. By attaching proper statistics about the internal structure of a dataset (in particular about the used vocabularies, properties etc.) it will be dramatically simplified to quickly identify suitable target datasets for linking. For example, the use of longitude and latitude properties in a dataset indicates that this dataset might be a good candidate for linking spatial objects. If we additionally know the minimum, maximum and average values for these properties, we can even identify datasets which are suitable link targets for a certain region.

5 Availability, Interfaces and Sustainability

In this section, we describe the interfaces to access the dataset as well as how we support sustainability. We publish our dataset on datahub.io data catalog¹⁵. The datahub.io entry for LODStats includes:

- **VoID description.** Machine readable description of the dataset.
- **LDSO vocabulary.** LODStats Dataset Vocabulary.
- **LODStats SPARQL endpoint.** SPARQL endpoint for the application.
- **LODStats RDF dump.** The RDF dump of LODStats dataset (April 2016).
- **VoID descriptions RDF dump.** Automatically generated VoID descriptions from the LODStats application (April 2016).
- **Data.gov, PublicData.eu, Datahub.io RDF dumps.** RDF dumps of the crawled data catalogs (April 2015).

The SPARQL endpoint serves the last output of RDB2RDF module and exposes up-to-date data. We announce the LODStats dataset using public Semantic Web

¹⁵ Available at <http://datahub.io/dataset/lodstats>

lists and create a Web forum¹⁶ to support community feedback. The sustainability of LODStats is demonstrated through: (i) the LODStats project being running for over the last five years, (ii) a state of the Data Web evaluation being performed every 2 months or at least once per half a year during this period, (iii) the last evaluation was performed just recently.

6 Data Web Statistics Summary

In this section we provide brief overview of the insights into the Data Web, based on the statistics collected over the past five years for the RDF dumps. The general current statistics such as number of triples, entities, literals etc. are available on the LODStats web portal¹⁷.

Over the past five years, the number of the datasets has increased from 422 in 2011 to 9644 in 2015. The burst of the datasets number has occurred in 2015, when we included data catalogs from the Open Governments in LODStats. However, only a small part of the overall amount of triples: 1% for the PublicData.eu and 3% for the Data.gov portals, can be attributed to the governmental data catalogs. It can be explained by the fact, that Open Governments publish short documents such as monthly energy consumption or salary rates for the governmental facilities. The connectedness of the Data Web has increased to 40% since 2011, when only 3% of the overall amount of triples were links between different datasets.

The further Web Data statistics can be accessed from the LODStats SPARQL endpoint¹⁸. For instance, the datasets in 2011 can be requested as follows:

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX ldso: <http://lodstats.aksw.org/ontology/ldso.owl#>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>

SELECT ?datasetName ?evaluationDate ?ckanCatalogName {
  ?statResult a ldso:StatResult .
  ?statResult dc:modified ?evaluationDate .
  ?statResult foaf:primaryTopic ?dataset .
  ?dataset dc:identifier ?datasetName .
  ?dataset dc:isPartOf ?ckanCatalog .
  ?ckanCatalog dc:identifier ?ckanCatalogName .
  FILTER (?evaluationDate < "2012-01-01T00:00:00"^^xsd:dateTime
  && ?evaluationDate > "2011-01-01T00:00:00"^^xsd:dateTime)}
```

7 Conclusion and Future Work

We presented LODStats – The Data Web Census Dataset, which exposes statistics about the Data Web over the last five years. We exposed the dataset using

¹⁶ <https://groups.google.com/forum/#!forum/lodstats>

¹⁷ Statistics can be accessed at <http://lodstats.aksw.org/stats>

¹⁸ <http://lodstats.aksw.org/sparql>

SPARQL endpoint and as an RDF dump, providing the one point of access at the DataHub.io data catalog. We created a mailing list to collect the feedback from the community and announced the dataset on the major mailing lists.

In the future we will be processing very large datasets with more than hundreds of millions triples, which are expensive to process on a single machine. While LODStats supports horizontal scaling for a parallel processing of several datasets, it is an open question on how to distribute the processing of a single dataset. Also, the statistical evaluation for SPARQL endpoints are often not a feasible task due to request timeouts. Many SPARQL endpoints do not provide VoID descriptions for datasets or provide outdated ones. Therefore, the Semantic Web community will benefit from the solution, which can provide machine-readable statistics for large datasets using minimal resources.

References

1. Keith Alexander, Richard Cyganiak, Michael Hausenblas, and Jun Zhao. Describing linked datasets. In *LDOW*, 2009.
2. Dean Allemang and James Hendler. *Semantic web for the working ontologist: effective modeling in RDFS and OWL*. Elsevier, 2011.
3. 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.
4. Carlos Buil-Aranda, Aidan Hogan, Jürgen Umbrich, and Pierre-Yves Vandenbussche. Sparql web-querying infrastructure: Ready for action? In *The Semantic Web—ISWC 2013*, pages 277–293. Springer, 2013.
5. Ivan Ermilov, Michael Martin, Jens Lehmann, and Sören Auer. Linked open data statistics: Collection and exploitation. In *Knowledge Engineering and the Semantic Web*, pages 242–249. Springer, 2013.
6. Eva Mendez Rodriguez Greenberg, Jane Gema Bueno de la Fuente, Thomas Baker, Pierre-Yves Vandenbussche, and Bernard Vatant. Requirements for vocabulary preservation and governance. *Library Hi Tech*, 31(4):657–668, 2013.
7. Fadi Maali, John Erickson, and Phil Archer. Data catalog vocabulary (dcat). *W3C Recommendation*, 2014.
8. Axel-Cyrille Ngonga Ngomo and Sören Auer. Limes - a time-efficient approach for large-scale link discovery on the web of data. In *Proceedings of IJCAI*, 2011.
9. Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. *Discovering and maintaining links on the web of data*. Springer, 2009.