

What's up LOD Cloud

Observing The State of Linked Open Data Cloud Metadata

Ahmad Assaf^{1,2}, Aline Senart² and Raphaël Troncy¹

¹ EURECOM, Sophia Antipolis, France. <firstName.lastName@eurecom.fr>

² SAP Labs France. <firstName.lastName@sap.com>

Abstract. Linked Open Data (LOD) has emerged as one of the largest collections of interlinked datasets on the web. In order to benefit from this mine of data, one needs to access to descriptive information about each dataset (or metadata). However, the heterogeneous nature of data sources reflects directly on the data quality as these sources often contain inconsistent as well as misinterpreted and incomplete metadata information. Considering the significant variation in size, the languages used and the freshness of the data, one realizes that finding useful datasets without prior knowledge is increasingly complicated. Roomba is a tool we created to validate, correct and generate dataset metadata. In this paper, we present the results of running it on the LOD cloud hosted on the DataHub. The results demonstrate that the general state of LOD cloud needs more attention as most of the datasets suffer from bad quality metadata lacking some informative metrics needed to facilitate dataset search. The noisiest metadata values were access information such as licensing information, resource descriptions as well as resource reachability problems.

Keywords: Dataset Profile, Metadata, Data Quality, Data Portal

1 Introduction

From 12 datasets cataloged in 2007, the Linked Open Data (LOD) cloud³ has grown to nearly 1000 datasets containing more than 82 billion triples [2]. Data is being published by both public and private sectors and covers a diverse set of domains from life sciences to media or government data. The Linked Open Data cloud is potentially a gold mine for organizations and individuals who are trying to leverage external data sources in order to produce more informed business decisions [4]. However, the heterogeneous nature of data sources reflects directly on the data quality as these sources often contain inconsistent as well as misinterpreted and incomplete metadata information. Considering the significant variation in size, the languages used and the freshness of the data, one realizes that finding useful datasets without prior knowledge is increasingly complicated.

³ <http://datahub.io/dataset?tags=lod>

Data portals expose a set of information about each dataset as metadata. The model used varies across portals. However, a standard model should contain information about the datasets title, description, maintainer email, update and creation date, etc. We divided the metadata information into the following:

General information: General information about the dataset. e.g. title, description, ID, etc. This general information is manually filled by the dataset owner. In addition to that, tags and group information is required for classification and enhancing dataset discoverability. This information can be entered manually or inferred by modules plugged into a topical profiler. **Access information:** Information about accessing and using the dataset. This includes the dataset URL, license information i.e. license title and URL and information about the datasets resources. Each resource has as well a set of attached metadata e.g. resource name, URL, format, size, etc. **Ownership information:** Information about the ownership of the dataset. e.g. organization details, maintainer details, author, etc. The existence of this information is important to identify the authority on which the generated report and the newly corrected profile will be sent to. **Provenance information:** Temporal and historical information on the dataset and its resources. For example, creation and update dates, version information, version, etc. Most of this information can be automatically filled and tracked.

We have created Roomba, a tool⁴ that automatically validates, corrects and generates dataset metadata. In this paper, we target CKAN powered data portals and validate datasets against the CKAN standard model⁵. The results demonstrate that the general state of LOD cloud needs more attention as most of the datasets suffer from bad quality metadata lacking some informative metrics needed to facilitate dataset search. The noisiest metadata values were access information such as licensing information, resource descriptions as well as resource reachability problems.

2 Related Work

Data Catalog Vocabulary (DCAT) [6] and the Vocabulary of Interlinked Datasets (VoID) [5] are concerned with metadata about RDF datasets. There exist several tools aiming at exposing dataset metadata using these vocabularies like [3]. Few approaches tackle the issue of examining datasets metadata. The Project Open Data Dashboard⁶ validator analyzes machine readable files for automated metrics to check their alignment with the Open Data principles. Similarly on the LOD cloud, the Data Hub LOD Validator⁷ checks a dataset compliance for inclusion in the LOD cloud. However, it lacks the ability to give detailed insights about the completeness of the metadata and overview on the state of the whole LOD cloud group.

⁴ <https://github.com/ahmadassaf/opendata-checker>

⁵ http://demo.ckan.org/api/3/action/package_show?id=adur_district_spending

⁶ <http://labs.data.gov/dashboard/>

⁷ <http://validator.lod-cloud.net/>

The *State of the LOD Cloud Report* [8] measured the adoption of Linked Data best practices in 2011. More recently, the authors in [10] used LDSpider [7] to crawl and analyze 1014 different datasets in the web of Linked Data in 2014. While these reports expose important information about datasets like provenance, licensing and accessibility, they do not cover the whole set of metadata.

3 Experiments and Evaluation

In this section, we provide the experiments and evaluation of the proposed framework. All the experiments are reproducible by our tool and their results are available on the its Github repository.

We have run the framework on the LOD cloud. At the time of running this tool, the CKAN API returned 259 datasets for the tag name `lodcloud` with a total of 1068 resources. The current number of datasets in the LOD cloud (1000+ datasets) is the result of a crawl over the Web of Linked Data and not based on the datasets hosted in The Datahub [10]. We ran the instance and resource extractor in order to cache the metadata files for these datasets locally and ran the validation process which took around one and a half hour on a 2.6 Ghz Intel Core i7 processor with 16GB of DDR3 memory machine.

A CKAN dataset metadata describes three main sections in addition to the core dataset's properties. Those are the groups, tags and resources. Each section contains a set of metadata corresponding to one or more metadata type. For example, a dataset resource will have general information such as the resource name, access information such as the resource url and provenance information such as creation date. The framework generates a report aggregating all the problems in all these sections, fixing field values when possible. Errors can be the result of missing metadata fields, undefined field values or field value errors e.g., unreachable URL or incorrect email address.

Figures 1 and 2 show the percentage of errors found in metadata fields by section and by information type respectively. We found out that the most erroneous information for the dataset core information were ownership related as 41% were missing or undefined. Datasets resources have the poorest metadata. 64% of the general metadata, all the access information and 80% of the provenance information contained missing or undefined values. Table 1 shows the top metadata fields errors in each metadata information type.

We notice that 42.85% of the top metadata problems can be fixed automatically. 44.44% of these problems can be fixed by our tool while the others need tools that are plugged into the data portal. We further present and discuss the results grouped by metadata information type below.

3.1 General information

34 datasets (13.13%) did not have valid `notes` values. `tags` information for the datasets were complete except for the `vocabulary_id` as it was missing from

all the datasets' metadata. All the datasets `groups` information were missing `display_name`, `description`, `title`, `image_display_url`, `id`, `name`. After manual examination, we noticed a clear overlap between group and organization information. Many datasets like `event-media` used the organization field to show group related information (being in LOD Cloud) instead of the publishers details.

Metadata Field		Error %	Section	Error Type	Auto Fix
General	<code>group</code>	100%	Dataset	Missing	-
	<code>vocabulary_id</code>	100%	Tag	Undefined	-
	<code>url-type</code>	96.82%	Resource	Missing	-
	<code>mimetype_inner</code>	95.88%	Resource	Undefined	Yes
	<code>hash</code>	95.51%	Resource	Undefined	Yes
	<code>size</code>	81.55%	Resource	Undefined	Yes
Access	<code>cache_url</code>	96.9%	Resource	Undefined	-
	<code>webstore_url</code>	91.29%	Resource	Undefined	-
	<code>license_url</code>	54.44%	Dataset	Missing	Yes
	<code>url</code>	30.89%	Resource	Unreachable	-
	<code>license_title</code>	16.6%	Dataset	Undefined	Yes
Provenance	<code>cache_last_updated</code>	96.91%	Resource	Undefined	Yes
	<code>webstore_last_updated</code>	95.88%	Resource	Undefined	Yes
	<code>created</code>	86.8%	Resource	Missing	Yes
	<code>last_modified</code>	79.87%	Resource	Undefined	Yes
	<code>version</code>	60.23%	Dataset	Undefined	-
Ownership	<code>maintainer_email</code>	55.21%	Dataset	Undefined	-
	<code>maintainer</code>	51.35%	Dataset	Undefined	-
	<code>author_email</code>	15.06%	Dataset	Undefined	-
	<code>organization_image_url</code>	10.81%	Dataset	Undefined	-
	<code>author</code>	2.32%	Dataset	Undefined	-

Table 1: Top metadata fields error % by type

3.2 Access information

25% of the datasets access information (being the dataset URL and any URL defined in its groups) has issues related to them (missing or unreachable URLs). Three datasets (1.15%) did not have a URL defined (`tip`, `uniprotdatabases`, `uniprotcitations`) while 45 datasets (17.3%) defined URLs were not accessible at the time writing this paper. One dataset did not have resources information (`bio2rdfchebi`) while the other datasets had a total of 1068 defined resources. On the datasets resources level, we noticed wrong or inconsistent values in the `size` and `mimetype` fields. 20 (1.87%) resources had incorrect `mimetype` defined, while 52 (4.82%) had incorrect `size` values. These values have been automatically fixed based on the values defined in the HTTP response header. However, 44 datasets have valid `size` field values and 54 have valid `mimetype` field values where they were not reachable, thus providing incorrect information.

15 (68%) fields of all the other access metadata are missing or have undefined values. Looking closely, we noticed that most of these problems can be easily fixed automatically by tools that can be plugged to the data portal. For example, the top six missing fields are the `cache_last_updated`, `cache_url`, `urltype`, `webstore_last_updated`, `mimetype_inner` and `hash` which can be computed and filled automatically. However, the most important missing information which require manual entry are the dataset's `name` and `description` were missing from 817 (76.49%) and 98 (9.17%) resources respectively. A total of 334 resources (31.27%) URLs were not reachable, thus affecting highly the availability of these datasets. CKAN resources can be of various predefined types (*file*, *file.upload*, *api*, *visualization*, *codeanddocumentation*). The framework also breaks down these unreachable resources according to their types. 211 (63.17%) resources did not have valid `resource_type`, 112 (33.53%) were files, 8 (2.39%) and one (0.029%) metadata, example and documentation types.

To have more details about the resources URL types, we created a *key : objectmeta – fieldvalues* group level report on LOD cloud with `resources> format:title`. This will aggregate the resources format information for each dataset. We found out that only 161 (62.16%) of the datasets valid URLs have SPARQL endpoints defined by `api/sparql` resource format. 92.27% provided RDF example links and 56.3% provided direct links to RDF down-loadable dumps.

The noisiest part of the access metadata was license information. A total of 43 datasets (16.6%) did not have a defined `license_title` and `license_id` fields, where 141 (54.44%) had missing `license_url` field. However, we managed to normalize 123 (47.49%) of the datasets' license information using the manual mapping file.

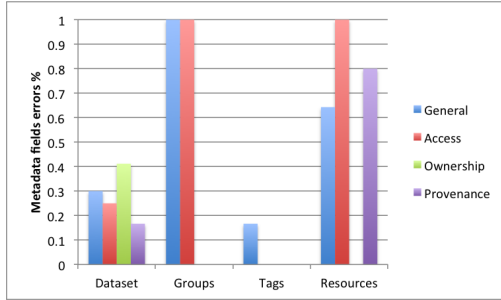


Fig. 1: Error % by section

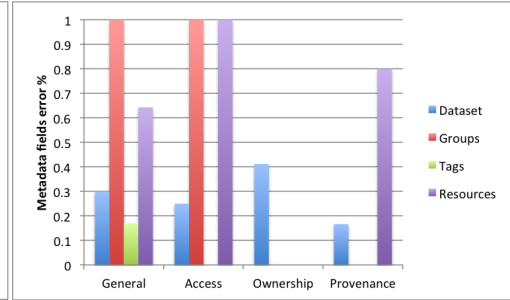


Fig. 2: Error % by information type

3.3 Ownership information

Ownership information is divided into direct ownership (author and maintainer) and organization information. Four fields (66.66%) of the direct ownership infor-

mation were missing or undefined. The breakdown for the missing information is: 55.21% `maintainer_email`, 51.35% `maintainer`, 15.06% `author_email`, 2.32% `author`. Moreover, our framework performs checks to validate existing email values. 11 (0.05%) and 6 (0.05%) of the defined `author_email` and `maintainer_email` fields were not valid email addresses respectively.

For the organization information, two field values (16.6%) were missing or undefined. 1.16% of the `organization_description` and 10.81% of the `organization_image_url` information with two out of these URLs were unreachable.

3.4 Provenance information

80% of the resources provenance information were missing or undefined. However, most of the provenance information e.g., `metadata_created`, `metadata_modified`) can be computed automatically by tools plugged into the data portal. The only field requiring manual entry is the `version` field which was found to be missing from 60.23% of the datasets.

4 Conclusion and Future Work

In this paper, we presented the results of running Roomba over the LOD cloud hosted in The Datahub. We discovered that the general state of the LOD datasets needs attention as most of them lack informative access information and their resources suffer low availability. These two metrics are of high importance for enterprises looking to integrate and use external linked data. We found out that the most erroneous information for the dataset core information were ownership related as 41% were missing or undefined. Datasets resources have the poorest metadata. 64% of the general metadata, all the access information and 80% of the provenance information contained missing or undefined values.

As part of our future work, we plan to run Roomba on various data portals and perform a detailed comparison to check the metadata health of LOD datasets against those in other prominent data portals.

References

1. C. Bizer. Evolving the web into a global data space. In *Proceedings of the 28th British National Conference on Advances in Databases*, 2011.
2. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 2009.
3. C. Böhm, J. Lorey, and F. Naumann. Creating void descriptions for web-scale data. *Web Semant.*, 2011.
4. D. Boyd and K. Crawford. Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, 2011.
5. R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander. Describing linked datasets with the VoID vocabulary. Technical report, 2011.
6. J. Erickson and F. Maali. Data catalog vocabulary (DCAT). Technical report, 2014. <http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/>.

7. R. Isele, J. Umbrich, C. Bizer, and A. Harth. Ldspider: An open-source crawling framework for the web of linked data. In *ISWC Posters & Demos*, CEUR Workshop Proceedings, 2010.
8. C. R. B. C. Jentzsch, A. State of the lod cloud.
9. H. Li. Data profiling for semantic web data. In *Web Information Systems and Mining*. 2012.
10. M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the linked data best practices in different topical domains. In *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*, 2014.