What's up LOD Cloud?

Observing The State of Linked Open Data Cloud Metadata

Ahmad Assaf[1,2], Aline Senart[2] and Raphaël Troncy[1]

[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 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 parts of the LOD cloud accessible via the DataHub API. The results demonstrate that the general state of examined datasets needs more attention as most of the datasets suffer from bad quality metadata lacking some informative metrics needed to facilitate dataset search. We also show that the automatic corrections done by Roomba increase the overall quality of the datasets metadata and highlight the need for manual efforts to correct some important missing information.

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

1 Introduction

The Linked Open Data (LOD) cloud[3] has grown significantly in the past years hosting various datasets covering a broad set of domains from life sciences to media and government data [3]. To maintain high quality data, publishers should comply with a set of best practices detailed in [2]. Metadata provisioning is one of those best practices requiring publishers to attach metadata needed to effectively understand and use datasets.

Data portals expose metadata via various models. A model should contain the minimum amount of information that conveys to the inquirer the nature and content of its resources [11]. It should contain information to enable data discovery, exploration and exploitation. We divided the metadata information into the following:

[3] http://datahub.io/dataset?tags=lod

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. **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.

Data portals are datasets' access points providing tools to facilitate data publishing, sharing, searching and visualization. CKAN[4] is the world's leading open-source data portal platform powering websites like the DataHub which hosts the LOD cloud. We have created Roomba [1], a tool that automatically validates, corrects and generates dataset metadata. Since we are evaluating the LOD cloud metadata, we validate the datasets against the CKAN standard model[5]. The results demonstrate that the general state of examined datasets 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 and resource descriptions in addition to large numbers of resource reachability problems. We also show that the automatic corrections of the tool increase the overall quality of the datasets metadata and highlight the need for manual efforts to correct some important missing information.

2 Related Work

Data Catalog Vocabulary (DCAT) [10] and the Vocabulary of Interlinked Datasets (VoID) [6] are concerned with metadata about RDF datasets. There exist several tools aiming at exposing dataset metadata using these vocabularies like [4]. Few approaches tackle the issue of examining datasets metadata. The Project Open Data Dashboard[6] 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 Validator[7] 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] http://ckan.org
[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 [12] 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 Roomba. All the experiments are reproducible by our tool and their results are available on the its Github repository[8].

3.1 Experimental Setup

The current state of the LOD cloud report [12] indicates that we have more than 1014 datasets available. These datasets were harvested via LDSpider crawler [7] seeded with 560 thousands URIs. However, since Roomba requires the datasets to be hosted in a data portal where either the dataset publisher or the portal administrator can attach relevant metadata to it, we relied on the information provided by the Datahub CKAN API. We examined two possible groups, the first tagged with "lodcloud" returned 259 datasets. The second tagged with "lod" returned only 75 datasets. We examined manually the two lists and found out the API result for the tag "lodcloud" is the correct one. The 259 datasets contained a total of 1068 resources. 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 two and a half hours on a 2.6 Ghz Intel Core i7 processor with 16GB of DDR3 memory machine.

3.2 Results and Evaluation

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.

[8] https://github.com/ahmadassaf/opendata-checker

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.

3.3 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	group	100%	Dataset	Missing	-
	vocabulary_id	100%	Tag	Undefined	-
	url-type	96.82%	Resource	Missing	-
	mimetype_inner	95.88%	Resource	Undefined	Yes
	hash	95.51%	Resource	Undefined	Yes
	size	81.55%	Resource	Undefined	Yes
Access	cahce_url	96.9%	Resource	Undefined	-
	webstore_url	91.29%	Resource	Undefined	-
	license_url	54.44%	Dataset	Missing	Yes
	url	30.89%	Resource	Unreachable	-
	license_title	16.6%	Dataset	Undefined	Yes
Provenance	cache_last_updated	96.91%	Resource	Undefined	Yes
	webstore_last_updated	95.88%	Resource	Undefined	Yes
	created	86.8%	Resource	Missing	Yes
	last_modified	79.87%	Resource	Undefined	Yes
	version	60.23%	Dataset	Undefined	-
Ownership	maintainer_email	55.21%	Dataset	Undefined	-
	maintainer	51.35%	Dataset	Undefined	-
	author_email	15.06%	Dataset	Undefined	-
	organization_image_url	10.81%	Dataset	Undefined	-
	author	2.32%	Dataset	Undefined	-

Table 1: Top metadata fields error % by type

3.4 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. 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.

3.5 Ownership information

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

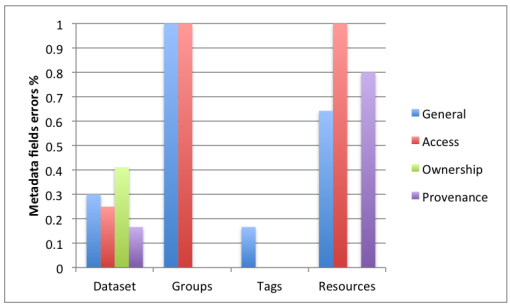

Fig. 1: Error % by section

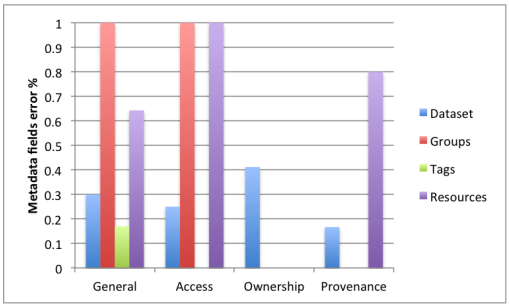

Fig. 2: Error % by information type

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.

3.6 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.

3.7 Enriched Profiles

Roomba can automatically fix when possible the license information (title, url and id) as well as the resources mimetype and size.

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.

We have noticed that most of the issues surrounding license information are related to ambiguous entries. To resolve that, we manually created a mapping file[9] standardizing the set of possible license names and urls using the open source and knowledge license information[10]. As a result, we managed to normalize 123 (47.49%) of the datasets' license information.

To check the impact of corrected fields, we seeded Roomba with the enriched profiles. Since Roomba uses file based cache system, we simply replaced all the datasets `json` files in the `\cache\datahub.io\datasets` folder with the those generated in `\cache\datahub.io\enriched`. After running Roomba again on the enriched profiles we noticed that the errors percentage for missing `size` fields decreased by 32.02% and for `mimetype` fields by 50.93%. We also noticed that the error percentage for missing `license_urls` decreased by 2.32%.

[9] https://github.com/ahmadassaf/opendata-checker/blob/master/util/licenseMappings.json
[10] https://github.com/okfn/licenses

4 Conclusion and Future Work

In this paper, we presented the results of running Roomba over the LOD cloud group hosted in the Datahub. We discovered that the general state of the examined 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.

We also showed that the automatic correction process can effectively enhance the quality of some information. We also noticed the need to have a community effort to manually correct missing important information like ownership information (maintainer, author, and maintainer and author emails).

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

Acknowledgments

This research has been partially funded by the European Union's 7th Framework Programme via the project Apps4EU (GA No. 325090).

References

1. A. Assaf, A. Sénart, and R. Troncy. Roomba: Automatic Validation, Correction and Generation of Dataset Metadata. In *24th World Wide Web Conference (WWW), Demos Track*, Florence, Italy, 2015.
2. C. Bizer. Evolving the Web into a Global Data Space. In *28th British National Conference on Advances in Databases*, 2011.
3. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems (IJSWIS)*, 2009.
4. C. BöHm, J. Lorey, and F. Naumann. Creating voiD Descriptions for Web-scale Data. *Journal of Web Semantics*, 9(3):339–345, 2011.
5. 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.
6. R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander. Describing Linked Datasets with the VoID Vocabulary. W3C Note, 2011. `http://www.w3.org/TR/void/`.
7. R. Isele, J. Umbrich, C. Bizer, and A. Harth. LDspider: An Open-source Crawling Framework for the Web of Linked Data. In *9th International Semantic Web Conference (ISWC), Posters & Demos Track*, 2010.
8. A. Jentzsch, R. Cyganiak, and C. Bizer. State of the lod cloud. `http://lod-cloud.net/state/`.

9. H. Li. Data Profiling for Semantic Web Data. In *International Conference on Web Information Systems and Mining(WISM)*, pages 472–479, Chengdu, China, 2012.
10. F. Maali and J. Erickson. Data Catalog Vocabulary (DCAT). W3C Recommendation, 2014. `http://www.w3.org/TR/vocab-dcat/`.
11. D. Nebert. Developing Spatial Data Infrastructures: The SDI Cookbook, 2004. `http://www.gsdi.org/docs2004/Cookbook/cookbookV2.0.pdf`.
12. M. Schmachtenberg, C. Bizer, and H. Paulheim. Adoption of the Linked Data Best Practices in Different Topical Domains. In *13th International Semantic Web Conference (ISWC)*, Riva del Garda, Italy, 2014.