

Observing The State of Linked Open Data Cloud Metadata

‡‡ Ahmad Assaf, ‡ Aline Senart, and † Raphaël Troncy
†EURECOM, Sophia Antipolis, France
‡ SAP Labs, Sophia Antipolis, France
†firstName.lastName@eurecom.fr, ‡firstName.lastName@sap.com

Keywords

Data Profiling, Metadata, Dataset Profile, LOD Cloud

1. INTRODUCTION

From 12 datasets cataloged in 2007, the Linked Open Data (LOD) cloud¹ has grown to nearly 1000 datasets containing more than 82 billion triples [1]. 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 [2]. 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.

We have created a tool² that automatically validates, corrects and generates dataset metadata for CKAN based data portals. In this poster, we present the results of running it on the LOD cloud hosted on the DataHub and comparing the results with those of Africa's largest open data portal, openAfrica. 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. METADATA PROFILING

¹<http://datahub.io/dataset?tags=lod>

²<https://github.com/ahmadassaf/opendata-checker>

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

WWW '15 Florence, Italy

Copyright 20XX ACM X-XXXXX-XX-X/XX/XX ...\$15.00.

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 modules plugged into the 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.

3. EXPERIMENT DETAILS

The results discussed below are based on the LOD cloud hosted on CKAN³. It contained 259 datasets at the time of writing this poster. For comparison, we have also profiled the openAfrica data portal⁴ containing 1653 datasets.

A CKAN dataset metadata describes three main sections in addition to the core datasets 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. Our tool 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.

³<http://ckan.org>

⁴<http://africaopendata.org/>

4. RESULTS AND EVALUATION

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. Figure 1 shows the percentage of errors found in metadata fields by section.

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.

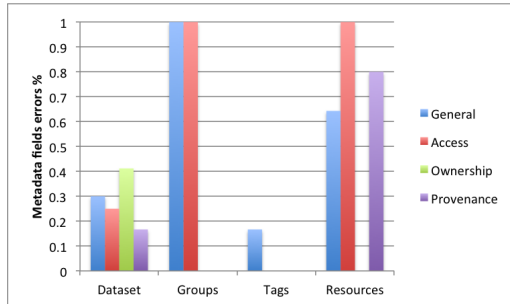


Figure 1: Error % by section

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`.

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 while 45 datasets (17.3%) defined URLs were not accessible at the time writing this poster.

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. 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. Their breakdown according to their type is: 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. 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.

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 tools 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.

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.

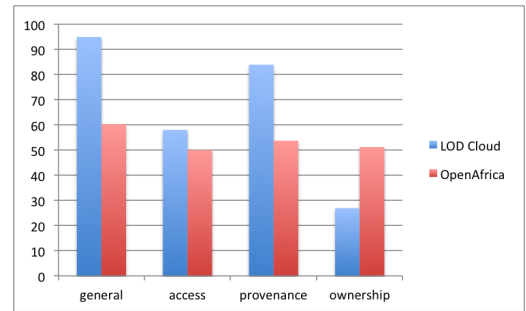


Figure 2: Average Error % by section LOD vs. openAfrica

Figure 2 shows the comparison between the average error rates in each metadata section in the LOD cloud and openAfrica. We notice that the quality of LOD cloud datasets is worse in the general, access and provenance sections. In the access information, the noted differences were in the percentage of unreachable URLs, `hash` and `size` fields. Moreover, provenance information had a clear gap as for example 86.8% of the LOD cloud `version` information was missing while it was complete in all the openAfrica datasets. However, the LOD cloud had better ownership and license information where 82.03% of the openAfrica datasets license information couldn't be normalized while it was only 52.5% in the LOD cloud.

5. REFERENCES

- [1] C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 5(3):1-22, 2009.
- [2] D. Boyd and K. Crawford. Six provocations for big data. *A Decade in Internet Time: Symposium on the Dynamics of the Internet and Society*, pages 1-17, 2011.