An Extensible Framework to Validate and Build Dataset Profiles

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 to descriptive information about each dataset (or metadata). This information can be used to delay data entropy, enhance datasets discovery, exploration and reuse as well as helping data portal administrators in detecting and eliminating spam. However, such metadata information is currently very limited to a few data portals where they are usually provided manually, thus being often incomplete and inconsistent in terms of quality. To address these issues, we propose a scalable automatic approach for extracting, validating, correcting and generating descriptive linked dataset profiles. This approach applies several techniques in order to check the validity of the metadata provided and to generate descriptive and statistical information for a particular dataset or for an entire data portal.

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

1 Introduction

From 12 datasets cataloged in 2007, the Linked Open Data cloud has grown to nearly 1000 datasets containing more than 82 billion triples[3] [6]. Data is being published by both the 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 [12]. This success lies in the cooperation between data publishers and consumers. Consumers are empowered to find, share and combine information in their applications easily. 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. This can be clearly noticed in the LOD Cloud where few datasets such as DBPedia [7], Freebase [11] and YAGO [42] are favored over less popular datasets that may include domain specific knowledge more suitable for the

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

tasks at hand. For example, for the task of building context-aware recommender systems in an academic digital library over LOD cloud, popular datasets like Semantic Web Dog Food[4], DBLP[5] or Yovisto[6] can be favored over lesser known but more specific datasets like VIAF[7] which links authority files of 20 national libraries, list of subject headings for public libraries in Spain[8] or the French dissertation search engine[9].

Dataset discovery can be done through public data portals like DataHub[10] and Europe's Public Data[11] or private ones like Quandl[12] or Engima[13]. Private portals harness manually curated data from various sources and expose them to users either freely or through paid plans. The data available is of higher quality but lesser quantity compared to what is available in public portals. Similarly, in some public data portals, administrators manually review datasets information, validate, correct and attach suitable metadata information. This information is mainly in the form of predefined tags such as *media, geography, life sciences* for organization and clustering purposes. However, the diversity of those datasets makes it harder to classify them in a fixed number of predefined tags that can be subjectively assigned without capturing the essence and breadth of the dataset [33]. Furthermore, the increasing number of datasets available makes the metadata review and curation process unsustainable even when outsourced to communities.

Data profiling is the process of creating descriptive information and collect statistics about that data. It is a cardinal activity when facing an unfamiliar dataset [1]. Data profiling reflects the importance of datasets without the need for detailed inspection of the raw data. It also helps in assessing the importance of the dataset, improving users' ability to search and reuse part of the dataset and in detecting irregularities to improve its quality. Data profiling includes typically several tasks:

- **Metadata profiling**: Provides general information on the dataset (dataset description, release and update dates), legal information (license information, openness), practical information (access points, data dumps), etc.
- **Statistical profiling**: Provides statistical information about data types and patterns in the dataset, i.e., properties distribution, number of entities and RDF triples, etc.
- **Topical profiling**: Provides descriptive knowledge on the dataset content and structure. This can be in form of tags and categories used to facilitate search and reuse.

[4] http://datahub.io/dataset/semantic-web-dog-food
[5] http://datahub.io/dataset/dblp
[6] http://datahub.io/dataset/yovisto
[7] http://datahub.io/dataset/viaf
[8] http://datahub.io/dataset/lista-encabezamientos-materia
[9] http://datahub.io/dataset/thesesfr
[10] http://datahub.io
[11] http://publicdata.eu
[12] https://quandl.com/
[13] http://enigma.io/

In this work, we address the challenges of automatic validation and generation of descriptive datasets profiles. This paper proposes Roomba, an extensible framework consisting of a processing pipeline that combines techniques for data portals identification, datasets crawling and a set of pluggable modules combining several profiling tasks. The framework validates the provided dataset metadata against an aggregated standard set of information. Metadata fields are automatically corrected when possible, e.g., adding a missing license URL reference. Moreover, a report describing all the issues highlighting those that cannot be automatically fixed is created to be sent by email to the dataset's maintainer. There exist various statistical and topical profiling tools for both relational and Linked Data. The architecture of the framework allows to easily add them as additional profiling tasks. However, in this paper, we focus on the task of dataset metadata profiling. We validate our framework against a manually created set of profiles and manually check its accuracy by examining the results of running it on various CKAN-based data portals.

The remainder of the paper is structured as follows. In Section 2, we present the motivation behind our framework. In Section 3, we review relevant related work. In Section 4, we describe our proposed framework's architecture and components that validate and generate dataset profiles. In Section 5, we evaluate the framework and we finally conclude and outline some future work in Section 6.

2 Motivation

Metadata provisioning is one of the Linked Data publishing best practices mentioned in [5]. Datasets should contain the metadata needed to effectively understand and use them. This information includes the dataset's license, provenance, context, structure and accessibility. The ability to automatically check this metadata helps in:

- **Delaying data entropy**: *Information entropy* refers to the degradation or loss limiting the information content in raw or metadata. As a consequence of information entropy, data complexity and dynamicity, the life span of data can be very short. Even when the raw data is properly maintained, it is often rendered useless when the attached metadata is missing, incomplete or unavailable. Comprehensive high quality metadata can counteract these factors and increase dataset longevity [32].
- **Enhancing data discovery, exploration and reuse**: Users who are unfamiliar with a dataset require detailed metadata to interpret and analyze accurately unfamiliar datasets. It is the first source checked when examining data sources. Moreover, several prominent data portals rely on metadata attached to datasets that enable search as they run a Solr index on the metadata schemas. Having missing or inconsistent information will affect the search results quality.
- **Enhancing spam detection**: Portals hosting public open data like Datahub allow anyone to freely publish datasets. Even with security measures like captchas and anti-spam devices, detecting spam is increasingly difficult. In

addition to that, the increasing number of datasets hinders the scalability of this process, affecting the correct and efficient spotting of datasets spam.

3 Related Work

Data Catalog Vocabulary (DCAT) [20] and the Vocabulary of Interlinked Datasets (VoID) [16] are concerned with metadata about RDF datasets. There exist several tools aiming at exposing dataset metadata using these vocabularies. In [9] authors generate VoID descriptions limited to a subset of properties that can be automatically deduced from resources within the dataset. However, it still provides data consumers with interesting insights. Quality Assessment of Data Sources (Flemming's Data Quality Assessment Tool)[14] provides basic metadata assessment as it calculates data quality scores based on manual user input. The user assigns weights to the predefined quality metrics and answer a series of questions regarding the dataset. These include, for example, the use of obsolete classes and properties by defining the number of described entities that are assigned disjoint classes, the usage of stable URIs and whether the publisher provides a mailing list for the dataset. The ODI certificate[15] on the other hand provides a description of the published data quality in plain English. It aspires to act as a mark of approval that helps publishers understand how to publish good open data and users how to use it. It gives publishers the ability to provide assurance and support on their data while encouraging further improvements through an ascending scale. ODI comes as an online and free questionnaire for data publishers focusing on certain characteristics about their data. Although these approaches try to perform metadata profiling, they are either incomplete or manual. In our framework, we propose a more automatized and complete approach.

Metadata profiling: The Project Open Data Dashboard[16] tracks and measures how US government websites implement the Open Data principles to understand the progress and current status of their public data listings. A validator analyzes machine readable files e.g., JSON files for automated metrics like the resolved URLs, HTTP status and content-type. However, deep schema information about the metadata is missing like description, license information or tags. Similarly on the LOD cloud, the Data Hub LOD Validator[17] gives an overview of Linked Data sources cataloged on the Data Hub. It offers a step-by-step validator guidance to check a dataset completeness level for inclusion in the LOD cloud. The results are divided into four different compliance levels from basic to reviewed and included in the LOD cloud. Although it is an excellent tool to monitor LOD compliance, it still lacks the ability to give detailed insights about the completeness of the metadata and overview on the state of the whole LOD cloud group and is very specific to the LOD cloud group rules and regulations.

[14] http://linkeddata.informatik.hu-berlin.de/LDSrcAss/datenquelle.php
[15] https://certificates.theodi.org/
[16] http://labs.data.gov/dashboard/
[17] http://validator.lod-cloud.net/

Statistical profiling: Calculating statistical information on datasets is vital to applications dealing with query optimization and answering, data cleansing, schema induction and data mining [28] [24] [33]. Semantic sitemaps [15] and RDFStats [34] where one of the first to deal with RDF data statistics and summaries. ExpLOD [30] creates statistics on the interlinking between datasets based on `owl:sameAs` links. In [1] the author introduces a tool that induces the actual schema of the data and gather corresponding statistics accordingly. LODStats [4] is a stream-based approach that calculates more general dataset statistics. ProLOD++ [2] is a Web-based tool that allows LOD analysis via automatically computed hierarchical clustering [10]. Aether [38] generates VoID statistical descriptions of RDF datasets. It also provides a Web interface to view and compare VoID descriptions. LODOP [22] is a MapReduce framework to compute, optimize and benchmark dataset profiles. The main target for this framework is to optimize the runtime costs for Linked Data profiling. In [29] authors calculate certain statistical information for the purpose of observing the dynamic changes in datasets.

Topical Profiling: Topical and categorical information facilitates dataset search and reuse. Topical profiling focuses on content-wise analysis at the instances and ontological levels. GERBIL [44] is a general entity annotation framework that provides machine processable output allowing efficient querying. In addition, there exist several entity annotation tools and frameworks [14] but none of those systems are designed specifically for dataset annotation. In [23], authors created a semantic portal to manually annotate and publish metadata about both LOD and non-RDF datasets. In [33], authors automatically assigned Freebase domains to extracted instance labels of some of the LOD Cloud datasets. The goal was to provide automatic domain identification, thus enabling improving datasets clustering and categorization. In [8], authors extracted dataset topics by exploiting the graph structure and ontological information, thus removing the dependency on textual labels. In [21] authors generate VoID and VoL descriptions via a processing pipeline that extracts dataset topic models ranked on graphical models of selected DBpedia categories.

Dataset Search:Dataset search can be done without relying on attached metadata (tags and categories). For example, there exist several approaches to create LOD indexes. In [3], authors used VoID descriptions to optimize query processing by determining relevant query-able datasets. In [26], authors created an approximate index structure (QTree) and an algorithm for answering conjunctive queries over Linked Data. SchemEX [31] is a stream-based approach leveraging type and property information of RDF instances to create schema-level indexes.

Semantic search engines like Sindice [18], Swoogle [19] and Watson [17] help in entities lookup but are not designed specifically for dataset search. In [39], authors utilized the sig.ma index [43] to identify appropriate data sources for interlinking. Dataset search and discovery is currently done via data portals that rely on attached metadata to provide dataset search features as they run a Solr index on the metadata schemas. Having missing or inconsistent information will affect the search results quality.

Although the above mentioned tools are able to provide various types of information about a dataset, there exists no approach that aggregates this information and is extensible to combine additional profiling tasks. To the best of our knowledge, this is the first effort towards extensible automatic validation and generation of descriptive dataset profiles.

4 Profiling Data Portals

In this section, we provide an overview of Roomba's architecture and the processing steps for validating and generating dataset profiles. Figure 1 shows the main steps which are the following: (i) data portal identification; (ii) metadata extraction; (iii) instance and resource extraction; (iv) profile validation (v) profile and report generation.

Roomba is built as a Command Line Interface (CLI) application using Node.js. Instructions on installing and running the framework are available on its public Github repository[18]. The various steps are explained in detail below.

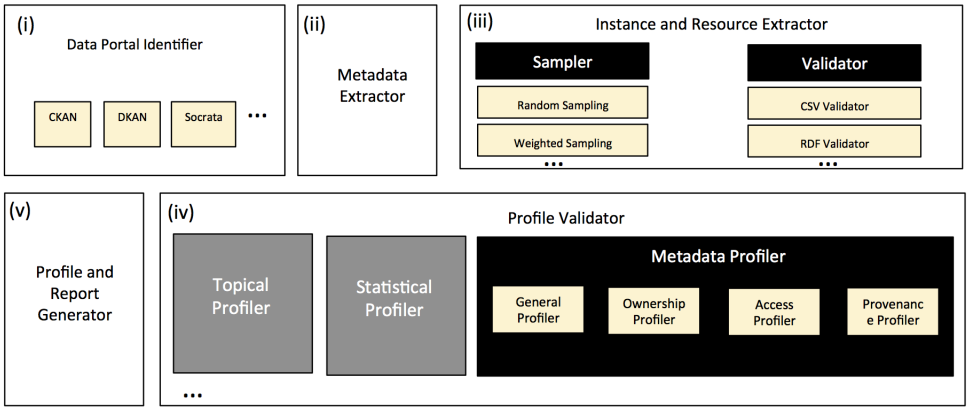

Fig. 1. Processing pipeline for validating and generating dataset profiles

4.1 Data Portal Identification

Data portals can be considered as data access points providing tools to facilitate data publishing, sharing, searching and visualization. CKAN[19] is the world's leading open-source data portal platform powering websites like DataHub, Europe's Public Data and the U.S Government's open data. Modeled on CKAN, DKAN[20] is a standalone Drupal distribution that is used in various public data

[18] https://github.com/ahmadassaf/opendata-checker
[19] http://ckan.org
[20] http://drupal.org/project/dkan

portals as well. Socrata[21] helps public sector organizations improve data-driven decision making by providing a set of solutions including an open data portal. In addition to these tradition data portals, there is a set of tools that allow exposing data directly as RESTful APIs like Datatank[22] and Database-to-API[23].

Roomba should be extensible to any data portal. Since every portal has its own API and data model, identifying the software powering data portals is a vital first step. We rely on several Web scraping techniques in the identification process which includes a combination of the following:

- **URL inspection**: Various CKAN based portals are hosted on subdomains of the `http://ckan.net`. For example, CKAN Brazil (`http://br.ckan.net`). Checking the existence of certain URL patterns can detect such cases.
- **Meta tags inspection**: The `<meta>` tag provides metadata about the HTML document. They are used to specify page description, keywords, author, etc. Inspecting the `content` attribute can indicate the type of the data portal. We use CSS selectors to check the existence of these meta tags. An example of a query selector is `meta[content*="ckan]` (all meta tags with the attribute content containing the string $CKAN$). This selector can identify CKAN portals whereas the `meta[content*="Drupal"]` can identify DKAN portals.
- **Document Object Model (DOM) inspection**: Similar to the meta tags inspection, we check the existence of certain DOM elements or properties. For example, CKAN powered portals will have DOM elements with class names like `ckan-icon` or `ckan-footer-logo`. A CSS selector like `.ckan-icon` will be able to check if a DOM element with the class name `ckan-icon` exists. The list of elements and properties to inspect is stored in a separate configurable object for each portal. This allows the addition and removal of elements as deemed necessary.

The identification process for each portal can be easily customized by overriding the default function. Moreover, adding or removing steps from the identification process can be easily configured.

After those preliminary checks, we query one of the portal's API endpoints. For example, DataHub is identified as CKAN, so we will query the API endpoint on `http://datahub.io/api/action/package_list`. A successful request will list the names of the site's datasets, whereas a failing request will signal a possible failure of the identification process.

4.2 Metadata Extraction

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 dataset's title, description, maintainer email, update and creation date, etc. We divided the metadata information into the following types:

[21] http://www.socrata.com
[22] http://thedatatank.com
[23] https://github.com/project-open-data/db-to-api

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 dataset's 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.

Building a standard metadata model is not the scope of this paper, and since we focus on CKAN-based portals, we validate the extracted metadata against the CKAN standard model[24].

After identifying the underlying portal software, we perform iterative queries to the API in order to fetch datasets metadata and persist them in a file-based cache system. Depending on the portal software we can issue specific extraction jobs. For example, in CKAN based portals, we are able to crawl and extract the metadata of a specific dataset, all the datasets in a specific group e.g., LOD Cloud or all the datasets in the portal.

4.3 Instance and Resource Extraction

From the extracted metadata we are able to identify all the resources associated with that dataset. They can have various types like a SPARQL endpoint, API, file, visualization, etc. However, before extracting the resource instance(s) we perform the following steps:

- **Resource metadata validation and enrichment**: Check the resource attached metadata values. Similar to the dataset metadata, each resource should include information about its mimetype, name, description, format, valid de-referenceable URL, size, type and provenance. The validation process issue an HTTP request to the resource and automatically fills up various missing information when possible, like the mimetype and size by extracting them from the HTTP response header. However, missing fields like name and description that needs manual input are marked as missing and will appear in the generated summary report.

[24] http://demo.ckan.org/api/3/action/package_show?id=adur_district_spending

– **Format validation**: Validate specific resource formats against a linter or a validator. For example, node-csv[25] for CSV files and n3[26] to validate N3 and Turtle RDF serializations.

Considering that certain datasets contain large amounts of resources and the limited computation power of some machines on which the framework might run on, a sampler module can be introduced to execute various sample-based strategies detailed as they were found to generate accurate results even with comparably small sample size of 10%. These strategies introduced in [21] are:

– **Random Sampling**: Randomly selects resources instances.
– **Weighted Sampling**: Weighs each resources as the ratio of the number of datatype properties used to define a resource over the maximum number of datatype properties over all the datasets resources.
– **Resource Centrality Sampling**: Weighs each resource as the ration of the number of resource types used to describe a particular resource divided by the total number of resource types in the dataset. This is specific and important to RDF datasets where important concepts tend to be more structured and linked to other concepts.

However, the sampler is not restricted only to these strategies. Strategies like those introduced in [35] can be configured and plugged in the processing pipeline.

4.4 Profile Validation

A dataset profile should include descriptive information about the data examined. In our framework, we have identified three main categories of profiling information. However, the extensibility of our framework allows for additional profiling techniques to be plugged in easily i.e., a quality profiling module reflecting the dataset quality. In this paper, we focus on the task of metadata profiling.

Metadata validation process identifies missing information and the ability to automatically correct them. Each set of metadata (general, access, ownership and provenance) is validated and corrected automatically when possible. Each profiler task has a set of metadata fields to check against. The validation process check if each field is defined and if the value assigned is valid.

There exist a bunch of special validation steps for various fields. For example, the email addresses and urls should be validated to ensure that the value entered is syntactically correct. In addition to that, for urls, we issue an HTTP `HEAD` request in order to check if that URL is reachable. We also use the information contained in a valid `content-header` response to extract, compare and correct some resources metadata values like `mimetype` and `size`.

Despite the legal issues surrounding Linked Data licenses [40], it is still considered a gold mine for organizations who are trying to leverage external data

[25] https://github.com/wdavidw/node-csv
[26] https://github.com/RubenVerborgh/N3.js

sources in order to produce more informed business decisions [12]. In [36] the authors see the potential economic effect unfolding in education, transportation, consumer products, electricity, oil and gas, health care and consumer finance. They estimate the potential annual value enabled by Open Data in these domains to be 3 trillion US Dollars across seven domains. As a result, validating license related information is vital. However, from our experiments, we found out that datasets' license information is noisy. The license names if found are not standardized. For example, Creative Commons CCZero can be also CC0 or CCZero. Moreover,the license URI if found and if de-referenceable can point to different reference knowledge bases e.g., `http://opendefinition.org`. To overcome this issue, we have manually created a mapping file standardizing the set of possible license names and the reference knowledge base[27]. In addition, we have also used the open source and knowledge license information[28] to normalize the license information and add extra metadata like the domain, maintainer and open data conformance.

```
{
        "license_id" : ["ODC-PDDL-1.0"],
        "disambiguations" : ["Open Data Commons Public Domain Dedication and License
             (PDDL)"]
},
{
        "license_id" : ["CC-BY-SA-4.0", "CC-BY-SA-3.0"],
        "disambiguations" : ["cc-by-sa", "CC BY-SA","Creative Commons Attribution
             Share-Alike"]
}
```

Listing 1.1. License mapping file sample

4.5 Profile and Report Generation

The validation process highlights the missing information and presents them in a human readable report. The report can be automatically sent to the dataset maintainer email if exists in the metadata.

In addition to the generated report, the enhanced profiles are represented in JSON using the CKAN data model and are publicly available[29].

Data portal administrators need an overall knowledge of the portal datasets and their properties. Our framework has the ability to generate numerous reports of all the datasets by passing formated queries. There are two main sets of aggregation tasks that can be run:

- **Aggregating meta-field values**: Passing a string that corresponds to a valid field in the metadata. The field can be flat like `license_title` (aggregates all the license titles used in the portal or in a specific group) or nested like `resource>resource_type` (aggregates all the resources types for all the datasets). Such reports are important to have an overview of the possible values used for each metadata field.

[27] https://github.com/ahmadassaf/opendata-checker/blob/master/util/licenseMappings.json
[28] https://github.com/okfn/licenses
[29] https://github.com/ahmadassaf/opendata-checker/tree/master/results

– **Aggregating key:object meta-field values**: Passing two meta-field values separated by a colon `:` e.g., `resources>resource_type:resources>name`. These reports are important as you can aggregate the information needed when also having the set of values associated to it printed.

For example, the meta-field value query `resource>resource_type` run against the LODCloud group will result in an array containing $[file, api, documentation...]$ values. These are all the resource types used to describe all the datasets of the group. However, to be able to know also what are the datasets containing resources corresponding to each type, we issue a key:object meta-field query `resource>resource_type:name`. The result will be a JSON object having the `resource_type` as the key and an array of corresponding datasets titles that has a resource of that type.

```
                    Metadata Report
group information is missing. Check organization information as they
    can be mixed sometimes
organization_image_url field exists but there is no value defined
                    Tag Statistics
There is a total of: 21 [undefined] vocabulary_id fields  100.00%
                    License Report
License information has been normalized !
                  Resource Statistics
There is a total of: 10 [missing] url-type fields  100.00%
There is a total of: 9 [missing] created fields  90.00%
There is a total of: 10 [undefined] cache_last_updated fields  100.00%
There is a total of: 10 [undefined] size fields  100.00%
There is a total of: 10 [undefined] hash fields  100.00%
There is a total of: 10 [undefined] mimetype_inner fields  100.00%
There is a total of: 7 [undefined] mimetype fields  70.00%
There is a total of: 10 [undefined] cache_url fields  100.00%
There is a total of: 6 [undefined] name fields  60.00%
There is a total of: 9 [undefined] webstore_url fields  90.00%
There is a total of: 9 [undefined] last_modified fields  90.00%
There is one [undefined] format field  10.00%
              Resource Connectivity Issues
There are 2 connectivity issues with the following URLs:
  - http://dbpedia.org/void/Dataset
                  Un-Reachable URLs Types
There are: 1 unreachable URLs of type [file]
```

Listing 1.2. Excerpt of the DBpedia validation report

5 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 in its Github repository.

A CKAN dataset metadata describes four main sections in addition to the core dataset's properties. These sections are:

- **Resources**: The distributable parts containing the actual raw data. They can come in various formats (JSON, XML, RDF, etc.) and can be downloaded or accessed directly (REST API, SPARQL endpoint).
- **Tags**: Provide descriptive knowledge on the dataset content and structure. They are used mainly to facilitate search and reuse.
- **Groups**: A dataset can belong to one or more group that share common semantics. A group can be seen as a cluster or a curation of datasets based on shared categories or themes.
- **Organizations**: A dataset can belong to one or more organization controlled by a set of users. Organizations are different from groups as they are not constructed by shared semantics or properties, but solely on their association to a specific administration party.

Each of these sections contains a set of metadata corresponding to one or more type (general, access, ownership and provenance). 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 addresses.

5.1 Experimental Setup

We ran our tool on two CAKN-based data portals. The first is the Datahub targeting specifically the LOD cloud group. The current state of the LOD cloud report [41] indicates that the LOD cloud contains 1014 datasets. They were harvested via an LDSpider crawler [27] seeded with 560 thousands URIs. Roomba on the other hand, fetches datasets hosted in data portals where datasets have attached relevant metadata. As a result, we relied on the information provided by the Datahub CKAN API. Examining the tags available, we found two candidate groups. The first tagged with "lodcloud" returned 259 datasets, while the second tagged with "lod" returned only 75 datasets. After manually examining the two lists, we found out the datasets grouped with the tag "lodcloud" are the correct ones. To qualify other CKAN-based portals for the experiments, we used `dataportals.org`, which contains a comprehensive list of Open Data portals from around the world. In the end, we chose the Amsterdam data portal [30]. The portal was commissioned in 2012 by the Amsterdam Economic Board Open Data Exchange (ODE), and covers a wide range of information domains (energy, economy, education, urban development, etc.) about Amsterdam metropolitan region.

[30] http://data.amsterdamopendata.nl/

We ran the instance and resource extractors in order to cache the metadata files for these datasets locally and ran the validation process. The experiments were executed on a 2.6 Ghz Intel Core i7 processor with 16GB of DDR3 memory machine. The approximate execution time alongside the summary of the datasets' properties are presented in table 1.

Data Portal	No. Datasets	No. Groups	No. Resources	Processing Time
LOD Cloud	259	N/A	1068	140 mins
Amsterdam Open Data	172	18	480	35 mins

Table 1. Summary of the experiments details

In our evaluation, we focused on two aspects: i)*profiling correctness* which assesses the validity of the errors generated in report manually, and ii)*profiling completeness* which assesses if the profilers cover all the errors in the datasets metadata.

5.2 Profiling Correctness

To measure profile correctness, we need to make sure that the issues reported by Roomba are valid on the dataset, group and portal levels.

On the dataset level, we choose three datasets from both the LOD Cloud and the Amsterdam data portal. The datasets details are shown in table 2.

Dataset Name	**Data Portal**	**Group ID**	**Resources**	**Tags**
dbpedia	Datahub	lodcloud	10	21
event-media	Datahub	lodcloud	9	15
bbc-music	Datahub	lodcloud	2	14
bevolking_cijfers_amsterdam	Amsterdam	bevolking	6	12
bevolking-prognoses-amsterdam	Amsterdam	bevolking	1	3
religieuze_samenkomstlocaties	Amsterdam	bevolking	1	8

Table 2. Datasets chosen for the correctness evaluation

To measure the profiling correctness on the groups level, we selected four groups from the Amsterdam data portal containing a total of 25 datasets. The choice was made to cover groups in various domains that contain a moderate number of datasets that can be checked manually (between 3-9 datasets). Table 3 summarizes the groups chosen for the evaluation.

After running Roomba and examining the results on the selected datasets and groups, we found out that our framework provides 100% correct results on the individual dataset level and on the aggregation level over groups. Since our portal level aggregation is extended from the group aggregation we can infer that the portal level aggregation also produces complete correct profiles. However, the

Group Name	Domain	Datasets	Resources	Tags
bestuur-en-organisatie	Management	9	45	101
bevolking	Population	3	8	23
geografie	Geography	8	16	56
openbare-orde-veiligheid	Public Order & Safety	5	19	34

Table 3. Groups chosen for the correctness evaluation

lack of a standard way to create and manage collections of datasets was the source of some errors when comparing the results from these two portals. For example, in Datahub, we noticed that all the datasets `groups` information were missing, while in the Amsterdam Open Data portal, all the `organisation` information was missing. Although the error detection is correct, the overlap in the usage of group and organization can give a false indication about the metadata quality.

5.3 Profiling Completeness

We analyzed the completeness of our framework by manually constructing a set of profiles that act as a golden standard. These profiles cover the range of uncommon problems that can occur in a certain dataset[31]. These errors are:

- Incorrect `mimetype` or `size` for resources
- Invalid number of tags or resources defined
- Check if the license information can be normalized via the `license_id` or the `license_title` as well as the normalization result.
- Syntacticly invalid `author_email` or `maintainer_email`.

After running our framework at each of these profiles, we measured the completeness and correcteness of the results. We found out that our framework covers indeed all the metadata problems that can be found in a CKAN standard model correctly.

6 Conclusion and Future Work

In this paper, we proposed a scalable automatic approach for extracting, validating, correcting and generating descriptive linked dataset profiles. This approach applies several techniques in order to check the validity of the metadata provided and to generate descriptive and statistical information for a particular dataset or for an entire data portal. Based on our experiments running the tool on the LOD cloud, we discovered that the general state of the 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.

It has been noticed that the issues surrounding metadata quality affect directly dataset search as data portals rely on such information to power their

[31] https://github.com/ahmadassaf/opendata-checker/tree/master/test

search index. We noted the need for tools that are able to identify various issues in this metadata and correct them automatically. We evaluated our framework manually against two prominent data portals and proved that we can automatically scale the validation of datasets metadata profiles completely and correctly.

As part of our future work, we plan to introduce workflows that will be able to correct the rest of the metadata either automatically or through intuitive manually-driven interfaces. We also plan to integrate statistical and topical profilers to be able to generate full comprehensive profiles. We also intend to suggest a ranked standard metadata model that will help generate more accurate and scored metadata quality profiles. We also plan to run this tool on various CKAN based data portals, schedule periodic reports to monitor the evolvement of datasets metadata. Finally, at some stage, we plan to extend this tool for other data portal types like DKAN and Socrata.

Acknowledgments

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

References

1. Data profiling for semantic web data. In *Web Information Systems and Mining*. Springer Berlin Heidelberg, 2012.
2. Z. Abedjan, T. Gruetze, A. Jentzsch, and F. Naumann. Profiling and mining rdf data with prolod++. In *Data Engineering (ICDE), 2014 IEEE 30th International Conference on*, 2014.
3. K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets - on the design and usage of void, the 'vocabulary of interlinked datasets'. In *WWW 2009 Workshop: Linked Data on the Web (LDOW2009)*.
4. S. Auer, J. Demter, M. Martin, and J. Lehmann. LODStats - an Extensible Framework for High-performance Dataset Analytics. In *18^{th} International Conference on Knowledge Engineering and Knowledge Management (EKAW)*, pages 353–362, Galway, Ireland, 2012.
5. C. Bizer. Evolving the web into a global data space. In *Proceedings of the 28th British National Conference on Advances in Databases*, 2011.
6. C. Bizer, T. Heath, and T. Berners-Lee. Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, 2009.
7. C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann. Dbpedia - a crystallization point for the web of data. *Web Semant.*, 2009.
8. C. Böhm, G. Kasneci, and F. Naumann. Latent topics in graph-structured data. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, 2012.
9. C. BöHm, J. Lorey, and F. Naumann. Creating void descriptions for web-scale data. *Web Semant.*, 2011.
10. C. Bohm, F. Naumann, Z. Abedjan, D. Fenz, T. Grutze, D. Hefenbrock, M. Pohl, and D. Sonnabend. Profiling linked open data with prolod. In *Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on*, 2010.

11. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, 2008.
12. 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.
13. M. Brümmer, C. Baron, I. Ermilov, M. Freudenberg, D. Kontokostas, and S. Hellmann. DataID: Towards semantically rich metadata for complex datasets. In *Proceedings of the 10th International Conference on Semantic Systems*, 2014.
14. M. Cornolti, P. Ferragina, and M. Ciaramita. A framework for benchmarking entity-annotation systems. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, 2013.
15. R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, and G. Tummarello. Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web. In *5^{th} European Semantic Web Conference (ESWC)*, pages 690–704, Tenerife, Spain, 2008.
16. R. Cyganiak, J. Zhao, M. Hausenblas, and K. Alexander. Describing linked datasets with the VoID vocabulary. Technical report, 2011.
17. M. d'Aquin and E. Motta. Watson, more than a semantic web search engine. *Semant. web*, 2011.
18. R. Delbru, N. Toupikov, and M. Catasta. Hierarchical link analysis for ranking web data. *The Semantic Web: Research and Applications*, 2010.
19. L. Ding, T. Finin, A. Joshi, R. Pan, and R. Cost. Swoogle: A semantic web search and metadata engine. *CIKM04*, 2004.
20. J. Erickson and F. Maali. Data catalog vocabulary (DCAT). Technical report, 2014. http://www.w3.org/TR/2014/REC-vocab-dcat-20140116/.
21. B. Fetahu, S. Dietze, B. Pereira Nunes, M. Antonio Casanova, D. Taibi, and W. Nejdl. A scalable approach for efficiently generating structured dataset topic profiles. In *The Semantic Web: Trends and Challenges*. Springer International Publishing, 2014.
22. B. Forchhammer, A. Jentzsch, and F. Naumann. LODOP - Multi-Query Optimization for Linked Data Profiling Queries. In *International Workshop on Dataset PROFIling and fEderated Search for Linked Data (PROFILES)*, Heraklion, Greece, 2014.
23. M. Frosterus, E. Hyvönen, and J. Laitio. DataFinland - A Semantic Portal for Open and Linked Datasets. In *8^{th} Extended Semantic Web Conference (ESWC)*, pages 243–254, 2011.
24. M. Frosterus, E. Hyvönen, and J. Laitio. Creating and publishing semantic metadata about linked and open datasets. In *Linking Government Data*. 2011.
25. M. Frosterus, E. Hyvönen, and J. Laitio. Datafinland - a semantic portal for open and linked datasets. In *ESWC (2)'11*, 2011.
26. A. Harth, K. Hose, M. Karnstedt, A. Polleres, K.-U. Sattler, and J. Umbrich. Data summaries for on-demand queries over linked data. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, 2010.
27. 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.
28. A. Jentzsch. Profiling the Web of Data. In *13^{th} International Semantic Web Conference (ISWC), Doctoral Consortium*, Trentino, Italy, 2014.

29. T. Käfer, A. Abdelrahman, J. Umbrich, P. O'Byrne, and A. Hogan. Observing Linked Data Dynamics. In *10^{th} European Semantic Web Conference (ESWC)*, 2013.
30. S. Khatchadourian and M. P. Consens. ExpLOD: Summary-based Exploration of Interlinking and RDF Usage in the Linked Open Data Cloud. In *7^{th} Extended Semantic Web Conference (ESWC)*, pages 272–287, Heraklion, Greece, 2010.
31. M. Konrath, T. Gottron, S. Staab, and A. Scherp. Schemex - efficient construction of a data catalogue by stream-based indexing of linked data. *Web Semant.*, 2012.
32. Kovács-Láng. Global Terrestrial Observing System. Technical report, GTOS Central and Eastern European Terrestrial Data Management and Accessibility Workshop, 2000.
33. S. Lalithsena, P. Hitzler, A. Sheth, and P. Jain. Automatic domain identification for linked open data. In *Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on*, 2013.
34. A. Langegger and W. Woss. RDFStats - An Extensible RDF Statistics Generator and Library. In *20^{th} International Workshop on Database and Expert Systems Application (DEXA)*, pages 79–83, 2009.
35. J. Leskovec and C. Faloutsos. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.
36. J. Manyika and E. A. Doshi. Open data: Unlocking innovation and performance with liquid information. Technical report, 2013.
37. P. N. Mendes, M. Jakob, A. García-Silva, and C. Bizer. Dbpedia spotlight: Shedding light on the web of documents. In *Proceedings of the 7th International Conference on Semantic Systems*, I-Semantics '11, 2011.
38. E. Mäkelä. Aether – generating and viewing extended void statistical descriptions of rdf datasets. In *Proceedings of the ESWC 2014 demo track, Springer-Verlag*, 2014.
39. A. Nikolov, M. d'Aquin, and E. Motta. What should i link to? identifying relevant sources and classes for data linking. In *The Semantic Web*, volume 7185 of *Lecture Notes in Computer Science*. Springer Berlin Heidelberg, 2012.
40. K. J. Prateek Jain, Pascal Hitzler and C. Venkatramani. There's no money in linked data. 2013.
41. 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.
42. F. M. Suchanek, G. Kasneci, and G. Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, 2007.
43. G. Tummarello, S. Danielczyk, R. Cyganiak, R. Delbru, M. Catasta, E. P. Federale, and S. Decker. Sig.ma: Live views on the web of data. In *In Proc. WWW-2010*. ACM Press.
44. R. Usbeck, M. Röder, A.-C. Ngonga-Ngomo, C. Baron, A. Both, M. Brümmer, D. Ceccarelli, M. Cornolti, D. Cherix, B. Eickmann, P. Ferragina, C. Lemke, A. Moro, R. Navigli, F. Piccinno, G. Rizzo, H. Sack, R. Speck, R. Troncy, J. Waitelonis, and L. Wesemann. GERBIL – general entity annotation benchmark framework. In *Submitted to the 24th WWW conference*, 2015.