ANALYSIS CERTIFICATE

magister
by compilatio.net

Account : **Bernard Merialdo**
ID : **g3xu5sj1**
Title : **Thesis ahmad assaf-rapporteurs.pdf**
Folder : **Ahmad Assaf**
Comments : *Not available*
uploaded on the :10/07/2015 4:17 PM

Similarity document :

2%

Similarities section 2 :

1%

DETAILED INFORMATION

Title : Thesis Ahmad Assaf-Rapporteurs.pdf
Description : Ahmad Assaf
Analysed on : 10/07/2015 4:38 PM
Login ID : 7c89gku1
uploaded on the : 10/07/2015 4:17 PM
Upload type : manual submission
File name : Thesis Ahmad Assaf-Rapporteurs.pdf
File type : pdf
Word count : 11033
Character count : 75726

TOP PROBABLE SOURCES- *AMONG 3 PROBABLE SOURCES*

1. **2014.eswc-conferences.org**/.../papers/paper_84.pdf <1%

SIMILARITIES FOUND IN THIS DOCUMENT/SECTION

Matching similarities : **<1 %**
Assumed similarities : **<1 %**
Accidental similarities : **<1 %**

Highly probable sources - 3
Less probable sources - 1

Accidental sources- 3 Sources
Ignored sources - 3

HIGHLY PROBABLE SOURCES

3 Sources	*Similarity*
1. **project-open-data.cio.gov**/.../v1.1/schema	<1%
2. **2014.eswc-conferences.org**/.../papers/paper_84.pdf	<1%
3. **ceur-ws.org**/.../Vol-1426/paper-03.pdf	<1%

LESS PROBABLE SOURCES

1 Source	*Similarity*
1. **dublincore.org**/.../documents/dcmi-terms	<1%

ACCIDENTAL SOURCES

3 Sources	*Similarity*
1. Document: ehjp18 - belongs to another user	<1%
2. **labs.data.gov**/.../dashboard/docs	<1%
3. **validator.lod-cloud.net**/.../	<1%

IGNORED SOURCES

3 Sources	*Similarity*
1. **ceur-ws.org**/.../Vol-1362/PROFILES2015_paper1.pdf	29%
2. **ceur-ws.org**/.../Vol-1362/PROFILES2015_paper3.pdf	20%
3. **www.eurecom.fr**/.../Publications/Assaf_Troncy-www15.pdf	8%

SIMILARITIES FOUND IN THIS DOCUMENT/SECTION

Matching similarities : **<1 %**

Assumed similarities : **<1 %**

Accidental similarities : **<1 %**

TEXT EXTRACTED FROM THE DOCUMENT

CKAN helps users from dierent domains (national and regional governments, companies and organizations) to easily publish their data through a set of workows to publish, share, search and manage datasets. CKAN is the portal powering web sites like Datahub, the Europe's Public Data portal or the U.S Government's open data portal5 . CKAN is a complete catalog system with an integrated data storage and powerful RESTful JSON API. It oers a rich set of visualization tools (e.g. maps, tables, charts) as well as an administration dashboard to monitor datasets usage and statistics. CKAN allows publishing datasets either via an import feature or through a web interface. Relevant metadata describing the dataset and its resources as well as organization related information can be added. A Solr6 index is built on top of this metadata to enable search and ltering. The CKAN data model7 contains information to describe a set of entities (dataset, resource, group, tag and vocabulary). CKAN keeps the core metadata restricted as a JSON le, but allows for additional information to be added via "extra" arbitrary key/value elds. CKAN supports Linked Data and RDF as it provides a complete and functional mapping of its model to Linked Data formats. An extension called ckanext-dcat8 provides plugins that allow CKAN to expose and consume metadata from other catalogs using DCAT as their model. The Open Data Companion Kit9 is a mobile application the provides a unied data access point for over 100 of open data portals. The application basically aims at CKAN-based portals providing a unique experience to mobile users.

5 6

http://data.gov http://lucene.apache.org/solr/ 7 http://docs.ckan.org/en/ckan-1.8/domain-model.html 8 https://github.com/ckan/ckanext-dcat 9 http://www.socrata.com/open-data-eld-guide/open-data-eld-kit/

3.1.7 DKAN

DKAN10 is a Drupal-based DMS with a full suite of cataloging, publishing and visualization features. Built over Drupal, DKAN can be easily customized and extended. The actual datasets in DKAN can be stored either within DKAN or on external sites. DKAN users are able to explore, search and describe datasets through the web interface or a RESTful API. The DKAN data model11 is very similar to the CKAN one, containing information to describe datasets, resources, groups and tags.

3.1.8 Socrata

Socrata12 is a commercial platform to streamline data publishing, management, analysis and reusing. It empowers users to review, compare, visualize and analyze data in real time. Datasets hosted in Socrata can be accessed using RESTful API that facilitates search and data ltering. Socrata allows exible data management by implementing various data governance models and ensuring compliance with metadata schema standards. It also enables administrators to track data usage and consumption through dashboards with realtime reporting. Socrata is very exible when it comes to customizations. It has a consumer-friendly experience giving users the opportunity to tell their story with data. Socrata's data model is designed to represent tabular data: it covers a basic set of metadata properties and has good support for geospatial data.

3.1.9 Junar

Junar13 adopts the Software-as-a-Service (SaaS) approach for data collection, enrichment, analysis and collaboration. Junar provides various functionalities that allow collaboration with colleagues to manage Open Data projects. Users are allowed to attach metadata to the information they publish to enhance search and discoverability.

3.1.10 INSPIRE metadata

The Infrastructure for Spatial Information in the European Community directive (INSPIRE)14 aims at ensuring a compatible and usable spatial data infrastructure across the European Union.

http://nucivic.com/dkan/ http://docs.getdkan.com/dkan-documentation/dkan-developers/datasettechnical-field-reference/ 12 http://socrata.com 13 http://junar.com/ 14 http://inspire.ec.europa.eu/index.cfm

11 10

The directive proposes a framework using a common metadata specication for data sharing, monitoring and reporting. The framework also denes rules to describe datasets and a set of implementation rules. For metadata schema, these include rules for the description of data sets, which could be adopted by open data publishers.

3.1.11

Schema.org

Schema.org15 is a collection of schemas used to markup HTML pages with structured data. This structured data allows many applications, such as search engines, to understand the information contained in Web pages, thus improving the display of search results and making it easier for people to nd relevant data. Schema.org covers many domains. We are specically interested in the Dataset schema. However, there are many classes and properties that can be used to describe organizations, authors, etc.

3.1.12

Common Core Metadaa Schema (CCMS)

Project Open Data (POD)16 is an online collection of best practices and case studies to help data publishers. It is a collaborative project that aims to evolve as a community resource to facilitate adoption of open data practices and facilitate collaboration and partnership between both private and public data publishers. The POD metadata model (CCMS)17 is based on DCAT. Similarly to DCAT-AP, POD denes three types of metadata elements: Required, Required-if (conditionally required) and Expanded (optional). The metadata model is presented in the JSON format and encourages publishers to extend their metadata descriptions using elements from the “Expanded Fields” list, or from any well-known vocabulary.

3.2

Metadata Model Classication

A dataset metadata model must contain sucient information so that consumers can easily understand and process the data that is described. After analyzing the most prominent models described in section 3.1, we nd out that a dataset can contain four main sections: • Resources: The actual raw data that can be downloaded or accessed directly via queryable endpoints. Resources can come in various formats such as JSON, XML or RDF.

15 16

http://schema.org http://project-open-data.cio.gov/ 17 https://project-open-data.cio.gov/v1.1/schema/

• Tags: Descriptive knowledge about the dataset content and structure. This can range from simple textual representation to semantically rich controlled terms. Tags are the basis for datasets search and discovery. • Groups: Groups act as organizational units that share common semantics. They can be seen as a cluster or a curation of datasets based on shared categories or themes. • Organizations: Organizations are another way to arrange datasets. However, they dier from groups as they are not constructed by shared semantics or properties, but solely on the dataset’s association to a specic administration party. Upon close examination of the various data models, we grouped the metadata information into eight main types. Each section discussed above should contain one or more of these types. For example, resources have general, access, ownership and provenance information while tags have general and provenance information only. The eight information types are: • General information: The core information about the dataset (e.g., title, description, ID). The most common vocabulary used to describe this information is Dublin Core18 . • Access information: Information about dataset access and usage (e.g., URL, license title and license URL). In addition to the properties in the models discussed above, there are several vocabularies designed specially to describe data access rights, e.g., Linked Data Rights19 , the Open Digital Rights Language (ODRL)20 . • Ownership information: Authoritative information about the dataset (e.g., author, maintainer and organization). The common vocabularies used to expose ownership information are Friend-of-Friend (FOAF)21 for people and relationships, vCard [73] for people and organizations and the Organization ontology [127] designed specically to describe organizational structures. • Provenance information: Temporal and historical information about the dataset creation and update records, in addition to versioning information (e.g., creation data, metadata update data, latest version). Provenance information coverage varies across the modeled surveyed. However, its great importance

18 19

http://dublincore.org/documents/dcmi-terms/ http://oeg-dev.dia.fi.upm.es/licensius/static/ldr/ 20 http://www.w3.org/ns/odrl/2/ 21 http://xmlns.com/foaf/spec/

lead to the development of various special vocabularies like the Open Provenance Model22 and PROV-O [95]. DataID [31] is an eort to provide semantically rich metadata with focus on providing detailed provenance, license and access information. • Geospatial information: Information reecting the geographical coverage of the dataset represented with coordinates or geometry polygons. There are several additional models and extensions specically designed to express geographical information. The Infrastructure for Spatial Information in the European Community (INSPIRE) directive23 aims at establishing an infrastructure for spatial information. Mappings have been made between DCAT-AP and the INSPIRE metadata. CKAN provides as well a spatial extension24 to add geospatial capabilities. It allows importing geospatial metadata from other resources and supports various standards (e.g., ISO 19139) and formats (e.g., GeoJSON). • Temporal information: Information reecting the temporal coverage of the dataset (e.g., from date to date). There has been some notable work on extending CKAN to include temporal information. govdata.de is an Open Data portal in Germany that extends the CKAN data model to include information like temporal granularity, temporal coverage to and temporal granularity from. • Statistical information: Statistical information about the data types and patterns in datasets (e.g., properties distribution, number of entities and RDF triples). This information is particularly useful to explore a dataset as it gives detailed insights about the raw data when provided properly. VoID is the only model that provides statistical information about a dataset. VoID denes properties to express dierent statistical characteristics of datasets like the total number of triples, total number of entities, total number of distinct classes, etc. However, there are other vocabularies such as SCOVO [69] that can model and publish statistical data about datasets. • Quality information: Information that indicates the quality of the dataset on the metadata and instance levels. In addition to that, a dataset should include an openness score that measures its alignment with the Linked Data publishing standards [16]. Quality information is only expressed in the POD metadata. However, govdata.de extends the CKAN model also to include a ratings average eld. Moreover, there are various other vocabularies like

22 23

http://open-biomed.sourceforge.net/opmv/ http://inspire.ec.europa.eu/ 24 https://github.com/ckan/ckanext-spatial

daQ [43] that can be used to express datasets quality. The RDF Review Vocabulary25 can also be used to express reviews and ratings about the dataset or its resources. Figure 3.1 summarizes the information grouping. Each dataset describes one or more information section (resources, tags, groups or organizations) which can contain one more information type.

Figure 3.1: Information sections and groups across data models

3.3

Mapping Metadata Models

Since establishing a common vocabulary or model is the key to communication, we identied the need for an harmonized dataset metadata model containing sucient information so that consumers can easily understand and process datasets. To create the mappings between the dierent models, we performed various steps: • Examine all the models and vocabularies specications and documentations. • Examine existing datasets using these models and vocabularies. Data Portals26 provides a comprehensive list of Open Data Portals from around the world. It was our entry point to nd out portals using CKAN or DKAN as their underlying DMS. We also investigated portals known to be using specic DMS. Socrata, for example, maintains a list of Open Data portals using their software on their homepage such as http://pencolorado.org and http://data. maryland.gov. • Examine the source code of some portals. This was specically the case for Socrata as their API returns the raw data serialized as JSON rather than the

25 26

http://vocab.org/review/ http://dataportals.org

CKAN resources tags groups organization DKAN resources tags groups organization POD distribution keyword theme publisher DCAT dcat:Distribution dcat:Dataset :keyword dcat:Dataset :theme dcat:Dataset :publisher VoID void:Dataset void:dataDump void:Dataset :keyword void:Dataset :publisher

Schema.org Dataset:distribution CreativeWork:keywords CreativeWork:about Socrata attachments tags category -

Table 3.1: Data models sections mapping

dataset's metadata. As a consequence, we had to investigate the Socrata Open Data API (SODA) source code27 and check the dierent classes and interfaces. The rst task is to map the four main information sections (resources, tags, groups and organization) across those models. Table 3.1 shows our proposed mappings. For the ontologies (DCAT, VoID), the rst part represents the class and the part after represents the property. For Schema.org, the rst part refers to the schema and the second part after : refers to the property. Table 3.2 presents the full mappings between the models across the information groups. Entries in the CKAN marked with are properties from CKAN extensions and are not included in the original data model. Similar to the sections mappings, for the ontologies (DCAT, VoID), the rst part represents the class and the part after represents the property. However, sometimes the part after refers to another resource. For example, to describe the dataset's maintainer email in DCAT, the information should be presented in the dcat:Dataset class using the dcat:contactPoint property. However, the range of this property is a resource of type vcard which has the property hasEmail. For Schema.org, similar to the sections mapping, the rst part refers to the schema and the second part after : refers to the property. However, if the property is inherited from another schema we denote that by using a as well. For example, the size of a dataset is a property for a Dataset schema specied in its distribution property. However, the type of distribution is dataDownload which is inherited from the MediaObject schema. The size for MediaObject is dened in its contentSize property which makes the mapping string Dataset:distribution DataDownload MediaObject:contentSize.

27 https://github.com/socrata/soda-java/tree/master/src/main/java/com/ socrata/model

Table 3.2: Harmonized Dataset Models Mappings

Data Model CKAN id private state type name isopen notes title num resources num tags DKAN id private state type name notes title POD identier accessLevel DCAT dcat:Dataset dct:identier VoID Schema.org Socrata id/externalId privateMetadata publicationStage name description name

Thing:additionalType Thing:name description title dcat:Dataset dct:description dcat:Dataset dct:title void:Dataset dct:description void:Dataset dc:title void:Dataset void:documents void:Dataset dct:conformsTo void:Dataset dct:language void:Dataset dct:accuralPeriodicity void:Dataset dct:license CreativeWork:license Thing:url void:Dataset dct:rights Thing:description Thing:name

General Information

conformsTo language accuralPeriodicity license title license id license url url attribution text version revision id metadata created metadata modied revision timestamp license title license

dcat:Dataset dct:conformsTo dcat:Dataset dct:language dcat:Dataset dct:accuralPeriodicity dcat:Distribution dct:license

CreativeWork:inLanguage

access information

license name licenseId license termsLink

url

landingPage rights

dcat:Dataset dcat:landingPage dcat:Distribution dct:rights

attribution attributionLink CreativeWork:version metadata created metadata modied revision timestamp dcat:Distribution dct:created dcat:Distribution dct:modied dcat:Distribution dct:issued dcat:Dataset dct:temporal dcat:Dataset dcat:contactPoint vcard:fn dcat:Dataset dcat:contactPoint vcard:hasEmail dcat:Dataset dct:creator foaf:Person:givenName author email bureauCode programCode dcat:Dataset foaf:Person:mbox dct:creator void:Dataset dct:creator foaf:Person:givenName void:Dataset foaf:Person:mbox dct:creator void:Dataset dct:created void:Dataset dct:modied void:Dataset dct:issued void:Dataset dct:temporal CreativeWork:dateCreated CreativeWork:dateModied CreativeWork:datePublished Dataset:temporal CreativeWork:producer Thing:name CreativeWork:producer Person:email owner displayName / owner ScreenName

provenance

modied issued temporal

maintainer maintainer email

maintainer maintainer email

contactPoint fn contactPoint hasEmail

owner org ownership author author email

CreativeWork:sourceOrganization:LegalName CreativeWork:author Thing:name CreativeWork:author Person:email

description isPartOf systemOfRecords describedBy describedByType spatial

CreativeWork:sourceOrganization Thing:description CreativeWork:isPartOf CreativeWork:hasPart

spatial-text geographical granularity GeoSpatial

dcat:Dataset dct:spatial

void:Dataset dct:spatial

Dataset:spatial bbox

3.3. Mapping Metadata Models

Table 3.2 Harmonized Dataset Models Mappings DCAT VoID

33

Data Model

CKAN

DKAN

POD

Schema.org

Socrata layers bboxCrs namespace

temporal Temporal temporal granularity temporal coverage to temporal coverage from ratings average

dcat:Dataset dct:temporal

void:Dataset dct:temporal

Dataset:temporal

Quality

dataQuality Organization void:Dataset dct:creator foaf:Organization:givenName

CreativeWork:aggregateRating

title description General Information id type name image url state is organization approval status

name

dcat:Dataset dct:creator foaf:Organization:givenName

CreativeWork:sourceOrganization:LegalName CreativeWork:sourceOrganization Thing:description CreativeWork:sourceOrganization Thing:additionalType CreativeWork:sourceOrganization Thing:name

subOrganizationOf provenance revision timestamp revision id Resources resource group id id size state hash general description format mimetype mimetype inner name position resource type describedBy describedByType conformsTo cache url url-type url url downloadURL accessURL webstore url cache last updated revision timestamp provenance dcat:Distribution dcat:downloadURL dcat:Distribution dcat:accessURL void:Dataset void:dataDump name title dcat:Distribution dct:title resource group id id size state description format mimetype description format mediaType dcat:Distribution dct:description dcat:Distribution dct:format dcat:Distribution dcat:mediaType void:Dataset dct:format dcat:Distribution dcat:byteSize

CreativeWork:sourceOrganization:subOrganization

blobId Dataset:distribution diaObject:contentSize DataDownload Me-

Dataset:distribution Thing:description

DataDownload Me-

Dataset:distribution DataDownload diaObject:encodingFormat

Dataset:distribution Thing:name

DataDownload

lename / name

Dataset:distribution DataDownloadThing:additionalType

access information

Dataset:distribution Thing:url Dataset:distribution diaObject:contentUrl

DataDownload DataDownload Me-

accessPoints

revision timestamp

Table 3.2 Harmonized Dataset Models Mappings DCAT VoID

Data Model

CKAN webstore last updated created last modied revision id display name description title image display url id name subgroups vocabulary id

DKAN created last modied revision id

POD

Schema.org Dataset:distribution DataDownload ativeWork:dataCreated Dataset:distribution DataDownload ativeWork:dataModied CreCre-

Socrata created at updated at

Groups display name description title image display url id name Tags dcat:Dataset dcat:theme skos:ConceptScheme dcat:Dataset dcat:keyword dcat:Dataset dcat:theme skos:Concept

General

vocabulary id

General

display name name state id revision timestamp

name

id

Provenance

3.4

Towards A Harmonized Model (HDL)

Examining the dierent models and their mappings in Table 3.2, we noticed a lack of a complete model that covers all the information types. There is an abundance of extensions and application proles that try to ll in those gaps, but they are usually domain specic addressing specic issues like geographic or temporal information.

To the best of our knowledge, there is still no

complete model that encompasses all the described information types. In this section, we present HDL, a harmonized dataset model that aims at lling this gap by taking the best from these models. In addition to the core dataset metadata, HDL describes the four common sections of datasets described in Section 3.2 (see Figure 3.2).

Figure 3.2: CKAN data model Table 3.3 describes the required elds across all the sections of a dataset and its core metadata. For example, a dataset resource, group, organization as well as to

the dataset itself will have an id, name, etc.

Field id name title Label Unique Identier Name Title Description A dataset unique identication Machine-readable name of the asset Human-readable name of the asset. Should be in plain English and include sucient detail to facilitate search and discovery Human-readable description (e.g., an abstract) with sucient detail to enable a user to quickly understand whether the asset is of interest Date on which the dataset was created Most recent date on which the dataset was changed, updated or modied Required Yes Yes Yes

description

Description

Yes

created modied

Creation Date Last Modication Date

Yes Yes

Table 3.3: Common required metadata elds for all the datasets sections Table 3.4 describes the authorship information that can be included in dierent sections. For example, a group has a required administrator eld. A group administrator inherits all the elds mentioned in this table, meaning that he must have an id, name, email and an optional role within the organization.

Field id name email role Label Unique Identier Name E-mail Role Description A person unique identication Human-readable name of the person A valid electronic mail address for the person Human-readable name of the asset. Should be in plain English and include sucient detail to facilitate search and discovery Required Yes Yes Yes No

Table 3.4: Metadata elds for ownership information

3.4.1

Resources

Resources are the main data containers of a dataset, they are a vital part of the dataset metadata as they are the facade on which users will interact with. Many of the core dataset metadata as we will see in Section 3.4.5 have an aggregate value of some resources elds. In addition to the common core metadata eld described in Table 3.3, Table 3.5 described the resources metadata elds.

Field type Label Type Description The human-readable format of the resource Required Yes Continued on next page

Table 3.5 Metadata elds for resources information section Label Description Download URL providing direct access to a resource, for example via API URL or a graphical interface Access URL URL providing indirect access to a resource. For example, the Web page on which the download url is available at Format A human-readable description of the le format of a distribution Hash Automatically generated unique md5 or sha-1 hash. Mainly used for indexing purposes. State The state of the current resource e.g. published, draft, under revision Access Level The degree to which this resource could be made publiclyavailable, e.g., public, restricted public, private MIME-type Machine-readable le format that conforms to the IANA Media Types 28 Size Actual size (content-length) of the resource in bytes Described By URL to the

data dictionary for the distribution found at the

Field download url access url format hash state access level mimetype size described by conforms to rating data quality cache url temporal granularity temporal coverage from

Required Yes Yes Yes Yes Yes Yes Yes Yes Yes No Yes Yes Yes If-Applicable If-Applicable

download url

Conforms To Rating Data Quality Cache URL Temporal Granulairty Temporal Coverage Starting Range Temporal Coverage End Range Spatial Text Spatial Granularity Bounding Box Layers URI used to identify a standardized specication the distribution conforms to Normalized score of the resource rating by users The resource objective quality score A URL of the resource cached version (used for portals with build in cloud storage) The detail levels associated with the temporal information of the dataset Start date of applicability for the data

temporal coverage to

End date of applicability for the data

If-Applicable

spatial text spatial granularity bbox layers

cache modied revision id

Cache Modied Revision ID

A textual information about the range of spatial applicability of a dataset. e.g., named place like London, United Kingdom. The detail levels associated with the spatial coverage of the dataset An area dened by two longitudes and two latitudes e.g., 0.489—51.28—0.236—51.686 A slice of the geographic coverage in a particular area. For example, on a road map roads, national parks, and rivers might be considered as dierent layers. Most recent date on which the resource cache was changed, updated or modied Latest revision ID for the resource Continued

If-Applicable If-Applicable If-Applicable If-Applicable

Yes Yes on next page

28

http://www.iana.org/assignments/media-types/media-types.xhtml

Table 3.5 Metadata elds for resources information section Label Description Revision Latest timestamp for the resource revision Timestamp License ID The normalized license ID with which the resource has been published. If the license is open, the ID should conform to one available at https://github.com/okfn/licenses License Title The normalized human-readable title of the resource license. If the license is open, the title should conform to one available at https://github.com/okfn/licenses License URL The normalized URL of the resource license. If the license is open, the URL should conform to one available at https:// github.com/okfn/licenses Attribution The attribution text that should be inserted based on the acText companying license guidelines if applicable.,The text is provided by the original author. Attribution The attribution link to the original source if applicable Link Rights Information regarding access or restrictions based on privacy, security, or other policies. If the access is restricted, should also include information on how to ask for access information.

Field revision timestamp license id

Required Yes Yes

license title

Yes

license url

Yes

attribution text

If-Applicable

attribution link rights

If-Applicable Yes

Table 3.5: Metadata elds for resources information section

3.4.2 Groups

In addition to the metadata elds in Table 3.3, a group must also include information about an author in an administrator eld. This means that he inherits all the elds mentioned in Table 3.4. In addition to that, a group can be part of a larger group, thus a subGroupOf eld is required when applicable to denote the id of the parent group.

3.4.3 Tags

One extra eld is required in addition to those mentioned in Table 3.3 which is vocabulary id. This elds represents a unique identier referring to the vocabulary (if used) controlling the tag. For example, if a dataset denes a geographical coverage, then a possible tag vocabulary would be to add a Country Code eld with values such as en, fr, ar, etc. This eld is optional, however, its existence enforce restrictions and provide semantic grouping and clustering of datasets in portals.

3.4.4 Organization

Table 3.6 describes the required eld to describe the organization information section in addition to those in Table 3.3. Those elds are mainly inspired by the Organization Ontology [127].

Field sub organization of Label Sub Organization Of Based At Description Represents hierarchical containment of organizations by Indicating if an organization is a sub-part or child of another organization Indicates the site at which an organization is based. This does not restrict the possibility for an organization to be at multiple sites human-readable address for the company’s site Location description for the organization e.g. lat, long coordinates Required If-Applicable

based at

Yes

has site location

Has Site Location

Yes

Table 3.6: Metadata elds for organization information section

3.4.5 Core Metadata

In addition to the common metadata elds described in Table 3.3, Table 3.7 describes the core metadata elds of every dataset. In addition to those, two authorship related elds are also required: maintainer and owner. Both elds inherit the authorship properties described in Table 3.4.

Field access Label Download URL Access URL State Access Level Rating Data Quality Revision ID Revision Timestamp

License ID Description URL providing direct access to a dataset, for example via API or a graphical interface. The access method should aggregate all the dataset resources available.

URL providing indirect access to a dataset.

For example, the Web page on which the download url is available at The state of the current dataset e.g. published, draft, under revision The degree to which this dataset could be made publiclyavailable, e.g., public, restricted public, private Normalized score of the average resources rating The average quality score of the dataset resources Latest revision ID for the resource Latest timestamp for the resource revision Required Yes

access url state access level rating data quality revision id revision timestamp license id

Yes Yes Yes Yes Yes Yes Yes

The normalised license ID(s) with which the dataset resources Yes has been published. If the license is open, the ID should conform to one available at https://github.com/okfn/licenses Continued on next page

Table 3.7 Dataset core metadata elds Description The normalised human-readable title(s) of the dataset resources licenses. If the license is open, the title should conform to one available at https://github.com/okfn/licenses License URL The normalised URL of the license used. If the license is open, the URL should conform to one available at https://github. com/okfn/licenses Attribution The attribution text that should be inserted based on the acText companying license guidelines if applicable.,The text is provided by the original author. Attribution The attribution link to the original source if applicable Link Rights An aggregate information regarding the dataset access or

restrictions based on privacy, security, or other policies.

If the access is restricted, should also include information on how to ask for access information. Language The aggregate set of languages used in the dataset resources Language The aggregate set of machine-readable language codes used in Code the dataset resources, e.g., en, fr Metadata The creation date of the dataset metadata Creation Date Metadata Most recent date on which the dataset metadata was changed, Modication updated or modied Date Is Part of The unique identier of a dataset of which the dataset is a subset Has Part The unique identier of a dataset which is a part of the current dataset Number of Total number of resources for the dataset Resources Number of Total number of tags for the dataset Tags Label License Title

Field license title

Required Yes

license url

Yes

attribution text

If-Applicable

attribution link rights

If-Applicable Yes

language language code metadata created

Yes Yes Yes

metadata modied

Yes

is part of has part number of resources number of tags

Yes Yes Yes Yes

Table 3.7: Dataset core metadata elds

3.4.6

Controlling Field Values

Various models control the set of values used to describe some of the model's properties. For example, CKAN model controls values for the resource type property and restrict them to: le: direct accessible bitstream, file.upload, api, visualization, code and documentation. However, dataset publishers do not always conform to these predened values and can add additional values. In order to know the set of values in these elds we examined the models of several CKAN datasets with a tool called Roomba. Roomba is a scalable automatic approach for extracting, validating, correcting and generating descriptive linked dataset proles (see Chapter 4). We created two main reports with Roomba. The rst aims to list the le types specied for resources using the query string resources>resource type:resources>name (see Listing 3.3) and the second one to collect the list of extras values using the query string extras>key:extras>value (see Listing 3.1 and Listing 3.2). We ran the report generation process on two prominent data portals: the Linked Open Data (LOD) cloud hosted on the Datahub containing 259 datasets and the Africa's largest open data portal, OpenAfrica29 that contains 1653 datasets.

namespace with t o t a l count o f : 1169 t r i p l e s with t o t a l count o f : 1193 p u b l i s h i n g I n s t i t u t i o n with t o t a l count o f : 17 shortname with t o t a l count o f : 753 l i n k s : d b p e d i a with t o t a l count o f : 768 l i n k s : l c s h with t o t a l count o f : 42

Listing 3.1: Excerpt of the extras aggregation report for the LOD Cloud

a c c e s s c o n s t r a i n t s with t o t a l count o f : 890 bboxe a s t l o n g with t o t a l count o f : 890 bboxwestl o n g with t o t a l count o f : 890 s p a t i a l with t o t a l count o f : 890 s p a t i a l datas e r v i c e t y p e with t o t a l count o f : 890 s p a t i a l r e f e r e n c e system with t o t a l count o f : 890

Listing 3.2: Excerpt of the extras eld aggregation report for OpenAfrica portal

f i l e with t o t a l count o f : 157 a p i with t o t a l count o f : 91 metadata with t o t a l count o f : 13 example with t o t a l count o f : 26 f i l e . u pl oa d with t o t a l count o f : 8 documentation with t o t a l count o f : 8 api , a p i / sparq , r d f with t o t a l count o f : 5 P u b l i c a t i o n with t o t a l count o f : 1 D a t a s e t with t o t a l count o f : 1

Listing 3.3: Result for aggregating resource type eld values on the LOD Cloud After examining the results, we noticed that for OpenAfrica, 53% of the datasets contained additional information about the geographical coverage of the dataset (e.g., spatial-reference-system, spatial harvester, bbox-east-long, bbox-north-long, bbox-south-long, bbox-west-long). In addition, 16% of the datasets have additional provenance and ownership information (e.g., frequency -of-update, dataset-reference-date). For the LOD cloud, the main information embedded in the extras elds are about the structure and statistical distribution of the dataset (e.g., namespace, number of triples and links). The OpenAfrica

29

http://africaopendata.org/

resources did not specify any extra resource types. However, in the LOD cloud, we observe that multiple resources dene additional types (e.g., example, api/sparql, publication, example). At the moment, HDL does not control the metadata eld values. However, restricting those values to a nite set as shown above pave the way to achieve better data harmonization across portals.

3.5

Summary

Data models vary across data portals. In this chapter, we surveyed the landscape of various models and vocabularies that described datasets on the web. As a result, we did not nd any that oers enough granularity to completely describe complex datasets facilitating search, discovery and recommendation. For example, the Datahub uses an extension of the Data Catalog Vocabulary (DCAT) [104] which prohibits a semantically rich representation of complex datasets like DBpedia30 that has multiple endpoints and thousands of dump les with content in several languages [31]. From our survey, we found that a proper integration of Open Data into businesses requires datasets to include the following information: • Access information: a dataset is useless if it does not contain accessible data dumps or query-able endpoints; • License information: businesses are always concerned with the legal implications of using external content. As a result, datasets should include both machine and human readable license information that indicates permissions, copyrights and attributions; • Provenance information: depending on the dataset license, the data might not be legally usable if there are no information describing its authoritative and versioning information. Current models under-specify these aspects limiting the usability of many datasets. Since establishing a common vocabulary or model is the key to communication, we identied the need for a harmonized dataset metadata model containing sucient information so that consumers can easily understand and process datasets. We have identied four main sections that should be included in the model: resources, groups, tags and organizations. Furthermore, we have classied the information to be included into eight types. Our main contribution is a set of mappings between each properties of those models. This has lead to the design of HDL, a harmonized dataset model, that takes the best out of these models to ensure complete metadata coverage to enable data discovery, exploration and reuse.

30

http://dbpedia.org

Chapter 4

Dataset Proles Generation and Validation

4.1 Introduction

The heterogeneous nature of data sources reects directly on the data quality as they often contain inconsistent as well as misinterpreted and incomplete metadata information. Moreover, the signicant variation in size, formats and freshness of the data, makes it more dicult to nd useful datasets without prior knowledge. This can be clearly noticed in the LOD Cloud where few datasets such as DBPedia [23], Freebase [27] and YAGO [136] are favored over less popular datasets that may include domain specic knowledge more suitable for the tasks at hand. For example, for the task of building context-aware recommender systems in an academic digital library over the LOD cloud, popular datasets like the Semantic Web Dog Food1 , DBLP2 or Yovisto3 can be favored over lesser known but more specic datasets like VIAF4 which links authority les of 20 national libraries, list of subject headings for public libraries in Spain5 or the French dissertation search engine6 . Users explore datasets in data portals relying on the metadata information attached by either the dataset owner or the data portal administrator. This information is mainly in form of predened tags such as media, geography, life sciences that are used for organization and clustering purposes. However, the increasing diversity of those datasets makes it harder to classify them in a xed number of tags that are subjectively assigned without capturing the essence and breadth of the dataset [91]. Furthermore, the increasing number of datasets available makes the manual review and curation of metadata unsustainable even when outsourced to communities. In this chapter, we address the challenges of automatic validation and generation of descriptive datasets proles. We describe Roomba, an extensible framework consisting of a processing pipeline that combines techniques for data portals identication, datasets crawling and a set of pluggable modules combining several proling

1 2

http://datahub.io/dataset/semantic-web-dog-food http://datahub.io/dataset/dblp 3 http://datahub.io/dataset/yovisto 4 http://datahub.io/dataset/viaf 5 http://datahub.io/dataset/lista-encabezamientos-materia 6 http://datahub.io/dataset/thesesfr

tasks. The framework validates the provided dataset metadata against an aggregated standard set of information. Metadata elds are automatically corrected when possible (e.g., adding a missing license URL reference). Moreover, a report describing all the issues that cannot be automatically xed is created to be sent by email to the dataset's maintainer. There exist various statistical and topical proling tools for both relational and Linked Data. The architecture of the framework allows to easily add them as additional proling tasks. However, in this chapter, we focus on the task of dataset metadata proling, ignoring the tasks of statistical and topical proling. We validate our framework against a manually created set of proles and manually check the accuracy by examining the results of running it on various CKAN-based data portals.

4.2 Motivation

Metadata provisioning is one of the Linked Data publishing best practices mentioned in [20]. Datasets should contain the metadata needed to eectively 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 [89]. • Enhancing data discovery, exploration and reuse: Users who are unfamiliar with a dataset require detailed metadata to interpret and analyze accurately unfamiliar datasets. A study conducted by the European Union commission [147] found that both business and users are facing diculties in discovering, exploring and reusing public data. due to missing or inconsistent metadata information. • 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 dicult. In addition to that, the increasing number of datasets hinders the scalability of this process, aecting the correct and ecient spotting of datasets spam.

4.3 Related Work

Data Catalog Vocabulary (DCAT) [104] and the Vocabulary of Interlinked Datasets (VoID) [37] are concerned with metadata about RDF datasets. There exist several tools aiming at exposing dataset metadata using these vocabularies. In [25], the 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. Flemming's Data Quality Assessment Tool7 provides basic metadata assessment as it computes data quality scores based on manual user input. The user assigns weights to the predened quality metrics and answers a series of questions regarding the dataset. These include, for example, the use of obsolete classes and properties by dening 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 certicate8 , 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 proling, they are either incomplete or manual. In our framework, we propose a more automatized and complete approach. Metadata proling: The Project Open Data Dashboard9 tracks and measures how US government web sites implement the Open Data principles

to understand the progress and current status of their public

data listings. A validator analyzes machine readable les: e.g., JSON les 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 Datahub LOD Validator10 gives an overview of Linked Data sources cataloged on the Datahub. It oers a step-by-step validator guidance to check a dataset

completeness level for inclusion in the LOD cloud.

The results are divided into four dierent 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 entire LOD cloud group and it is very specic to the LOD cloud group rules and regulations.

http://linkeddata.informatik.hu-berlin.de/LDSrcAss/datenquelle.php https://certificates.theodi.org/ 9 http://labs.data.gov/dashboard/ 10 http://validator.lod-cloud.net/

8 7

Statistical proling: Calculating statistical information on datasets is vital to applications dealing with query optimization and answering, data cleansing, schema induction and data mining [79, 56, 91]. Semantic sitemaps [36] and RDFStats [92] are one of the rst to deal with RDF data statistics and summaries. ExpLOD [83] creates statistics on the interlinking between datasets based on owl:sameAs links. In [100], the author introduces a tool that induces the actual schema of the data and gathers corresponding statistics accordingly. LODStats [13] is a stream-based approach that calculates more general dataset statistics. ProLOD++ [1] is a Web-based tool that allows LOD analysis via automatically computed hierarchical clustering [26]. Aether [106] generates VoID statistical descriptions of RDF datasets.

It also provides a Web interface to view and compare

VoID descriptions. LODOP [53] is a MapReduce framework to compute, optimize and benchmark dataset proles. The main target for this framework is to optimize the runtime costs for Linked Data proling. In [80] authors calculate certain statistical information for the purpose

of observing the dynamic changes in datasets.

Topical Proling: Topical and categorical information facilitates dataset search and reuse. Topical proling focuses on content-wise analysis at the instances and ontological levels. GERBIL [144] is a general entity annotation framework that provides machine processable output allowing ecient querying. In addition, there exist several entity annotation tools and frameworks [35] but none of those systems are designed specically for dataset annotation. In [57], the authors created a semantic portal to manually annotate and publish metadata about both LOD and non-RDF datasets. In [91], the authors automatically assigned Freebase domains to extracted instance labels of some of the LOD Cloud datasets. The goal was to provide automatic domain identication, thus enabling improving datasets clustering and categorization. In [24], the authors extracted dataset topics by exploiting the graph structure and ontological information, thus removing the dependency on textual labels. In [49], the 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], the authors used VoID descriptions to optimize query processing by determining relevant query-able datasets. In [64], the authors created an approximate index structure (QTree) and an algorithm for answering conjunctive queries over Linked Data. SchemEX [86] is a stream-based approach leveraging type and property information of RDF instances to create schema-level indexes. Semantic search engines like Sindice [45], Swoogle [46] and Watson [39] help in

Figure 4.1: Processing pipeline for validating and generating dataset proles entities lookup but they are not designed specically for dataset search. In [115], the authors utilized the sig.ma index [142] 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 aect 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 proling tasks. To the best of our knowledge, this is the rst eort towards extensible automatic validation and generation of descriptive dataset proles.

4.4 Proling Data Portals

In this section, we provide an overview of Roomba's architecture and the processing steps for validating and generating dataset proles. Figure 4.1 shows the main steps which are the following: (i) data portal identication; (ii) metadata extraction; (iii) instance and resource extraction; (iv) prole validation (v) prole and report generation. Roomba is built as a Command Line Interface (CLI) application (see Figure 4.2) using Node.js and is available on the tools Github repository11 . Roomba allows data portal administrators like Dan to: • Fetch information about the portal's data management system

11

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

• Fetch all the information about datasets from a data portal • Fetch all the groups information from a data portal • Crawl, fetch and cache datasets (a specic dataset, datasets in a specic group, datasets in the whole portal) • Execute aggregation report on a specic group or on the whole data portal • Prole a specic dataset, a whole group or the whole data portal

Figure 4.2: Screenshot for Roomba command line tool Appendix A.1 details the instructions for installing and running the framework. The various steps are explained in detail below.

4.4.1

Data Management System Identication

Data portals are considered to be data access points providing tools to facilitate data publishing, sharing, searching and visualization. Section 3.1 highlights the main data management systems powering those data portals and the various dataset models used. In addition to these traditional data management systems, there is a set of tools that allow exposing data directly as RESTful APIs like Datatank12 and Databaseto-API13 . Roomba is extensible to any data portal. Since every portal has its own API and data model, identifying the software powering data portals is a vital rst step. The

12 13

http://thedatatank.com https://github.com/project-open-data/db-to-api

Data Portal Identier (component (i)) relies on several Web scraping techniques in the identication 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. The Data Portal Identier uses 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, the Data Portal Identier checks the existence of certain DOM elements or properties. For example, CKAN-powered portals 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 congurable object for each portal. This allows the addition and removal of elements as deemed necessary. The identication process for each portal can be easily customized by overriding the default function. Moreover, adding or removing steps from the identication process can be easily congured. After those preliminary checks, the Data Portal Identier issues a query to one of the portal's API endpoints. For example, DataHub is identied 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 identication process.

4.4.2

Metadata Extraction

Data portals expose a set of information about each dataset as metadata. The model used varies across portals. However, a standard model (see section 3.2) must contain information about the dataset's title, description, maintainer email, update and creation date, etc. Since Roomba operates on CKAN-based data portals, the Metadata Extractor (component (ii)) validates the extracted metadata against the CKAN standard

50 model14 (see Listing 4.1).

{

license_title : License not specified , maintainer : , relationships_as_object : [], private : false, maintainer_email : , num_tags : 4, id : 7e4d4ef3-f452-4c35-963d-9c6e582374b3 , metadata_created : 2015-07-22T14:29:55.490069 , metadata_modified : 2015-07-22T14:30:18.584924 , author : Lucy Chambers , author_email : , state : active , version : , creator_user_id : 01b3756a-e1ca-4d4a-b8f1-6880a00095d6 , type : dataset }

Listing 4.1: Excerpt of a dataset prole in CKAN standard model After identifying the underlying portal software, The

Metadata Extractor performs iterative queries to the API in order to fetch datasets metadata and persist them in a le-based cache system. Depending on the portal software, The Metadata Extractor can issue specic extraction jobs. For example, in CKAN-based portals, The Metadata Extractor is able to crawl and extract the metadata of a specic dataset, all the datasets in a specic group (e.g., LOD cloud) or all the datasets in the portal.

4.4.3

Instance and Resource Extraction

From the extracted metadata, the Instance and Resource Extractor (component (iii)) is able to identify all the resources associated with that dataset. They can have various types like a SPARQL endpoint, API, le, visualization, etc. However, before extracting the resource instance(s), the extractor performs 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 MIME-type, name, description, format, valid dereferenceable URL, size, type and provenance. The validation process issues an HTTP request to the resource and automatically lls up various missing

14 http://demo.ckan.org/api/3/action/package_show?id=adur_district

spending

4.4. Proling Data Portals

51

information when possible, like the MIME-type and size by extracting them from the HTTP response header. However, missing elds like name and description that needs manual input are marked as missing and will appear in the generated summary report. • Format validation: Validate specic resource formats against a linter or a validator. For example, node-csv15 for CSV les and n316 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 submodule is introduced to execute various sample-based strategies as they were found to generate accurate results even with comparably small sample size of 10% [49]. The sampling strategies introduced are: • Random Sampling: Randomly selects resources instances.

• Weighted Sampling: Weighs each resource as the ratio of the number of datatype properties used to dene a resource over the maximum number of datatype properties over all the datasets resources. • Resource Centrality Sampling: Weighs each resource as the ratio 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 specic 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 that we oer by default. Strategies like those introduced in [99] can be congured and plugged in the processing pipeline.

4.4.4

Prole Validation

A dataset prole should include descriptive information about the data examined. In Roomba, we have identied three main categories of proling information. However, the extensibility of our framework allows for additional proling techniques to be plugged in easily (Section 5.5 describes an extension to measure the objective qualities of datasets). The Prole Validator (component (iv)) identies missing information and the ability to automatically correct them. Each set of metadata (general, access, ownership

15 16

https://github.com/wdavidw/node-csv https://github.com/RubenVerborgh/N3.js

52

Chapter 4. Dataset Proles Generation and Validation

and provenance) is validated and corrected automatically when possible. Each proler task has a set of metadata elds to check against. The validation process check if each eld is dened and if the value assigned is valid. There exist many special validation steps for various elds. 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, the Prole Validator issues an HTTP HEAD request in order to check if that URL is reachable. The Prole Validator also uses the information contained in a valid content-header response to extract, compare and correct some resources metadata values like mimetype and size. Having valid license information is vital for organization looking to integrate external data. However, from our experiments, we found out that datasets' license information is often missing or noisy. The license names if found are not standardized. For example, Creative Commons CCZero can also be CC0 or CCZero. Moreover, the license URI if found and if de-referenceable can point to dierent reference knowledge bases e.g., http://opendefinition.org. To overcome this issue, we have manually created a mapping le standardizing the set of possible license names and the reference knowledge base (see Listing E.1). In addition, we have also used the open source and knowledge license information17 to normalize the license information and add extra metadata like the domain, maintainer and open data conformance. The Prole Validator uses this mapping le to valdiate and normalize datasets license information.

{ 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 4.2: License mapping le sample

4.4.5

Prole and Report Generation

The validation process highlights the missing information and presents them in a human readable report (see appendix B). The report can be automatically sent to the dataset maintainer email if exists in the metadata. In addition to the generated report, the enhanced proles are represented in JSON using the CKAN data model

17

https://github.com/okfn/licenses

and are publicly available18 .

== Metadata Report == group i n f o r m a t i o n i s m i s s i n g . Check o r g a n i z a t i o n i n f o r m a t i o n a s t h e y can be mixed sometimes o r g a n i z a t i o n i m a g e u r l f i e l d e x i s t s but t h e r e i s no v a l u e d e f i n e d == Tag S t a t i s t i c s == There i s a t o t a l o f : 21 [u n d e f i n e d] v o c a b u l a r y i d f i e l d s 100.00% == L i c e n s e Report == L i c e n s e i n f o r m a t i o n has been n o r m a l i z e d ! == Resource S t a t i s t i c s == There i s a t o t a l o f : 10 [m i s s i n g] u r l t y p e f i e l d s 100.00% There i s a t o t a l o f : 9 [m i s s i n g] c r e a t e d f i e l d s 90.00% There i s a t o t a l o f : 10 [u n d e f i n e d] c a c h e l a s t u p d a t e d f i e l d s 100.00% There i s a t o t a l o f : 10 [u n d e f i n e d] s i z e f i e l d s 100.00% There i s a t o t a l o f : 10 [u n d e f i n e d] hash f i e l d s 100.00% There i s a t o t a l o f : 10 [u n d e f i n e d] m i m e t y p e i n n e r f i e l d s 100.00% There i s a t o t a l o f : 7 [u n d e f i n e d] mimetype f i e l d s 70.00% There i s a t o t a l o f : 10 [u n d e f i n e d] c a c h e u r l f i e l d s 100.00% There i s a t o t a l o f : 6 [u n d e f i n e d] name f i e l d s 60.00% There i s a t o t a l o f : 9 [u n d e f i n e d] w e b s t o r e u r l f i e l d s 90.00% There i s a t o t a l o f : 9 [u n d e f i n e d] l a s t m o d i f i e d f i e l d s 90.00% There i s one [u n d e f i n e d] f o r m a t f i e l d 10.00% == Resource Connectivity I s s u e s == There a r e 2 c o n n e c t i v i t y i s s u e s with t h e f o l l o w i n g URLs : \ u r l { http : / / dbpedia . org / void / Dataset } == UnR e a c h a b l e URLs Types == There a r e : 1 u n r e a c h a b l e URLs o f t y p e [f i l e]

Listing 4.3: Excerpt of the DBpedia validation report Data portal administrators like Paul 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 formatted queries. There are two main sets of aggregation tasks that can be run: • Aggregating meta-eld values: Passing a string that corresponds to a valid eld in the metadata. The eld can be at like license title (aggregates all the license titles used in the portal or in a specic group) or nested

18

https://github.com/ahmadassaf/opendata-checker/tree/master/results

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 eld. • Aggregating key:object meta-eld values: Passing two meta-eld 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-eld value query resource>resource type run against the LODCloud group will result in an array containing [f ile, 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-eld 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.

4.5

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 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 dierent from groups as they are not constructed by shared semantics or properties, but solely on their association to a specic 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, xing eld values when possible. Errors can be the result of missing metadata elds, undened eld values or eld value errors (e.g., unreachable URL or incorrect email addresses).

4.5.1

Experimental Setup

We ran our tool on two CKAN-based data portals. The rst is the Datahub targeting specically the LOD cloud group. The current state of the LOD cloud report [130] indicates that the LOD cloud contains 1014 datasets. They were harvested via an LDSpider crawler [76] 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 rst 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 as they contained more recent and accurate metadata. 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. We chose the Amsterdam data portal 19 as it is updated frequently and highly maintained. 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. 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 4.1. Data Portal LOD Cloud Amsterdam Open Data No. Datasets 259 172 No. Groups N/A 18 No. Resources 1068 480 Processing Time 140 mins 35 mins

Table 4.1: Summary of the experiments details In our evaluation, we focused on two aspects: i)proling correctness which manually assesses the validity of the errors generated in the report, and ii)proling completeness which assesses if the prolers cover all the errors in the datasets metadata.

19

http://data.amsterdamopendata.nl/

4.5.2

Proling Correctness

To measure prole 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 4.2.

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

Table 4.2: Datasets chosen for the correctness evaluation

To measure the proling 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 4.3 summarizes the groups chosen for the evaluation. Group Name bestuur-en-organisatie bevolking geograe openbare-orde-veiligheid Domain Management

Population Geography Public Order & Safety Datasets 9 3 8 5 Resources 45 8 16 19 Tags 101 23 56 34

Table 4.3: Groups chosen for the correctness 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 proles. However, the 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.

4.5.3 Proling Completeness

We analyzed the completeness of our framework by manually constructing a synthetic set of proles. These proles cover the range of uncommon problems that can occur in a certain dataset[20] . These errors are: • Incorrect mimetype or size for resources; • Invalid number of tags or resources dened; • Check if the license information can be normalized via the license id or the license title as well as the normalization result; • Syntactically invalid author email or maintainer email. After running our framework at each of these proles, we measured the completeness and correctness 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.

4.6 Analyzing Proling Results

In this section, we describe our experiments when running the Roomba tool on the LOD cloud. Figures 4.3 and 4.4 show the percentage of errors found in metadata elds by section and by information type respectively. We observe that the most erroneous information for the dataset core information is related to ownership since this information is missing or undened for 41% of the datasets. Datasets resources have the poorest metadata. 64% of the general metadata, all the access information and 80% of the provenance information contain missing or undened values. Table 4.4 shows the top metadata elds errors for each metadata information type. We notice that 42.85% of the top metadata problems can be xed automatically. Among them, 44.44% of these problems can be xed 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 in the following sub-sections.

4.6.1 General Information

34 datasets (13.13%) do not have valid notes values. tags information for the datasets are complete except for the vocabulary id as this is missing from all the datasets' metadata. All the datasets groups information are missing display name, description, title, image display url, id, name. After manual examination, we observe a clear overlap between group and organization information. Many

20 https://github.com/ahmadassaf/opendata-checker/tree/master/test

Metadata Field group vocabulary id url-type General mimetype inner hash size cache url webstore url Access license url url license title cache last updated webstore last updated Provenance created last modied version maintainer email maintainer Ownership author email organization image url author

Error % 100% 100% 96.82% 95.88% 95.51% 81.55% 96.9% 91.29% 54.44% 30.89% 16.6% 96.91% 95.88% 86.8% 79.87% 60.23% 55.21% 51.35% 15.06% 10.81% 2.32%

Section Dataset Tag Resource Resource Resource Resource Resource Resource Dataset Resource Dataset Resource Resource Resource Resource Dataset Dataset Dataset Dataset Dataset Dataset

Error Type Missing Undened Missing Undened Undened Undened Undened Undened Missing Unreachable Undened Undened Undened Missing Undened Undened Undened Undened Undened Undened Undened

Auto Fix Yes Yes Yes Yes Yes Yes Yes Yes Yes -

Table 4.4: Top metadata elds error % by type

datasets like event-media use the organization eld to show group related information (being in the LOD Cloud) instead of the publishers details.

4.6.2

Access Information