Challenges of mapping current CKAN metadata to DCAT

Sebastian Neumaier, Jürgen Umbrich, and Axel Polleres

Vienna University of Economics and Business

1 Introduction

This report describes our experiences using the mapping of the metadata in CKAN powered Open Data portals to the DCAT model. CKAN is the most prominent portal software framework used for publishing Open Data and used by several governmental portals including `data.gov.uk` and `data.gov`. We studied the actual usage of DCAT in 133 existing Open Data portals and report the following key findings:

– The CKAN model allows for additional/optional ("extra") metadata keys which have portal specific mappings to the DCAT model. This potentially results in differing DCAT representations or even in a loss of metadata information as we discuss in Section 2.

– We observed issues regarding the use of the current DCAT specification for modelling certain CKAN datasets: CKAN datasets commonly group semantically related but not identical content. Representing differing content within a single dataset is currently not supported in DCAT (cf. Section 3).

– Eventually, we discuss possible adaptions and suggestions for DCAT (Section 4).

As underlying data pool for this analysis serves the Open Data Portal Watch project,[1] a framework for monitoring and assessing the quality of metadata on Open Data portals, currently monitoring 133 active CKAN portals.[2]

1.1 DCAT export for CKAN metadata.

The CKAN software provides an extension to export and harvest RDF serializations of CKAN datasets based on DCAT.[3] This extension is maintained by the Open Knowledge Foundation and is also gathered in the Portal Watch project, however, in a slightly adapted form[1]). The extension defines the mapping of metdata for CKAN datasets and resources to the corresponding DCAT classes `dcat:Dataset` and `dcat:Distribution`.

With DCAT-AP, the DCAT application profile for European data portals, there is a DCAT specification for describing public sector datasets, which is also

[1] Neumaier, S., Umbrich, J., Polleres, A.: Automated quality assessment of metadata across open data portals. ACM Journal of Data and Information Quality (2016)

[2] `http://data.wu.ac.at/portalwatch`

[3] `https://github.com/ckan/ckanext-dcat`

supported by the recent version of the CKAN-to-DCAT extension.[4] However, in general we cannot assume that this extension is already deployed in all CKAN portals: we were able to retrieve the DCAT descriptions of datasets for 93 of the 133 active CKAN portals.

In the following we analyse current limitations of this CKAN-to-DCAT extension wrt. to optional "extra" metadata keys and the problem of modelling dataset resources as DCAT distributions.The analyses presented in this paper are based on a metadata snapshot of the 133 CKAN portals in week 29/2016 (third week of July).

2 Mapping of "extra" keys

The CKAN software allows portal providers to include additional metadata fields in the metadata schema. When retrieving the metadata description for a dataset via the API, these keys are included in the resulting JSON under the key "`extras`". However, it is not guaranteed that the DCAT conversion of the CKAN metadata contains these extra keys. Depending on the version and the configuration of the export-extension there are three different cases:

- *Predefined mapping:* In recent versions of the extension the portal provider can define a mapping for certain CKAN fields to a specific RDF property. For instance, a CKAN extra field `contact-email` (which is not by default part of the CKAN schema and is not defined in the extension's mapping) could be mapped to an RDF graph using the property vcard:hasEmail from the vCard ontology, e.g.:

```
<http://example.com/example-dataset>
  dcat:contactPoint [
    vcard:hasEmail "example@email.com"
  ] ;
```

- *Default mapping:* A pattern for exporting all available extra metadata keys which can be observed in several data portals[5] is the use of the dct:relation (Dublin Core vocabulary) property to describe just the label and value of the extra keys, e.g.:

```
<http://example.com/example-dataset>
  dct:relation [
    rdfs:label "contact-email" ;
    rdf:value "example@email.com"
  ] ;
```

- *No mapping:* The retrieved DCAT description returns no mapping of these keys and the information is therefore not available.

[4] The extension is DCAT-AP compatible since January 2016: `https://github.com/ckan/ckanext-dcat/commit/b0bf250926390d65e26c2bc2faaf3b8ee8db62f7`

[5] E.g., see `http://data.nsw.gov.au/data/dataset/2b20a392-5e1a-48f4-a85a-0d5e39661c1e.rdf`

> For instance, the Irish Open Data portal provides a set of extra keys, e.g., `contact-email`, `contact-name` or `date_updated` for their datasets.[6] However, the mapped DCAT representation omits these keys in the output.

Analysing "extra" keys in CKAN portals. We analysed the metadata of 514k datasets over all 133 CKAN portals and extracted a total of 3607 extra keys. Table 1 lists the most frequently used keys sorted by the number of datasets they appear in. Many of these keys occur in multiple portals (most frequent `spatial` in 33 portals) which is based on the fact that these keys are generated by widely used CKAN extensions. The keys in Table 1 are all generated by the harvesting[7] and spatial extension.[8]

Table 1: Most frequent extra keys

Extra key	**Datasets**	**Portals**	**Origin**	**DCAT key**
`harvest_object_id`	268241	30	Harvesting extension	—
`spatial`	245211	33	Spatial extension	dct:spatial
`harvest_source_id`	243136	29	Harvesting extension	—
`harvest_source_title`	243043	29	Harvesting extension	—
`guid`	150811	20	Spatial extension	—
`resource-type`	148843	15	Spatial extension	—
`contact-email`	148671	17	Spatial extension	dcat:contactPoint
`metadata-date`	141758	15	Spatial extension	dct:issued/modified
`spatial-reference-system`	140191	16	Spatial extension	—
`dataset-reference-date`	139874	15	Spatial extension	—

Additionally, we suggest in Table 1 DCAT conform properties which could be used to map these frequent extra keys (cf. "DCKAT key" column).[9] In case of an empty cell, we were not able to choose an appropriate property.

Looking into more detail of these 3607 extra keys, we discovered that 1502 unique keys are of the form `links:`{*dataset-id*}, e.g., `links:air-temperature` or `links:air-pressure`. All these `links:`-keys originate from the `datahub.io` portal, which provides references to Linked Data as CKAN datasets. The portal uses these keys to encode links between two datasets within the portal.

Table 2 lists the number of keys occurring in multiple portals: 35 of all extra keys occur in more than 10 of the 133 CKAN portals, which mainly originate

[6] E.g., see `https://data.gov.ie/api/rest/dataset/2a4912e7-4e8e-40b4-95a5-91d6ef505443`

[7] `http://extensions.ckan.org/extension/harvest/`

[8] `http://docs.ckan.org/projects/ckanext-spatial/en/latest/`

[9] Note, that the CKAN-to-DCAT extension maps the extra key `spatial_uri` to dct:spatial. However, `spatial_uri` is only present in 61 datasets.

Table 2: Extra keys appearing in multiple portals

Portals	1	2	$3-9$	$10-19$	≥ 20
Extra keys	2269	1131	172	30	5

from popular CKAN extensions. The majority of the extra keys occur only in one or two portals.[10]

3 Modelling CKAN resources using DCAT

The CKAN software allows data providers to add multiple *resources* to a dataset description. These resources are basically links to the actual data and some additional corresponding metadata (e.g., format, title, mime-type, ...).

This concept of resources relates to *distributions* in DCAT. A DCAT distribution is defined in the following way: "*Represents a specific available form of a dataset. Each dataset might be available in different forms, these forms might represent different formats of the dataset or different endpoints. [...]*"[11] This means that distributions of a dataset should consist of the same data in different representations. We applied the following two heuristics in order to find out if CKAN resources are used as distributions, i.e., if CKAN resources represent the same content in different formats:

– *Title similarity:* We compared the titles of resources of a dataset using Ratcliff-Obershelp string similarity used in the Python difflib library. In case any two resource-titles have a measure of higher than 0.8 (with a maximum similarity of 1) we consider the resources as "distributions". For instance, two resources with titles "air-temperature.csv" and "air-temperature.json" most likely contain the same data in CSV and JSON format.

– *Formats:* We looked into the file formats of the resources and report the number of datasets where (1) all formats differ and (2) some formats appear multiple times (e.g., a dataset consisting of two CSVs).

We analysed 514,675 datasets published on 133 CKAN data portals. Out of these 514k datasets 265,732 datasets hold more than one resource (cf. Table 3). Out of the 265k multi-resource datasets, for 100k datasets all corresponding file formats are different, indicating that these are possibly distributions of the dataset. Using string similarity we encountered similar titles for at least two resources in 145k out of the 265k datasets.

These numbers indicate that there is no common agreement on how to use resources of datasets in CKAN. On the one hand there is a high number of datasets where resources are published as "distributions" (see *different file formats* and *similar titles* in Table 3) while on the other hand the remaining datasets group

[10] The high number of keys occurring in two portals is potentially due to the fact that many portals harvest datasets, i.e. the metadata descriptions, of other portals (see the number of portals using the harvesting extension in Table 1).

[11] `https://www.w3.org/TR/vocab-dcat/#class-distribution`

Table 3: Distributions vs. resources in CKAN datasets

Total	**> 1 resources**	**File format**		**Similar titles**
		all different	*multi-appearance*	
514,675	265,732	100,170	165,562	145,171

resources by other aspects (see *multi-appearance*); e.g., a dataset consisting of the resources "air-temperature-2013.csv", "air-temperature-2014.csv", "air-temperature-2015.csv".

4 Conclusions & Suggestions

By analyzing metadata of 514k datasets of over 133 CKAN portals we have provided insights into the use of CKAN metadata keys in current Open Data portals and have highlighted potential issues when mapping these keys to DCAT.

Mapping produces different output for "extra" keys depending on the version of the extension: Differently mapped output of DCAT impedes integration and search across data portals. In order to overcome this issue CKAN data portals should export DCAT-AP conform metadata. Therefore the portals have to install/activate/update their CKAN-to-DCAT extensions.

No mapping defined for frequently used "extra" keys: Some widely-used CKAN extensions (e.g., for metadata harvesting) create/add additional metadata keys. For some of these keys there exists no proper mapping, even though there is a corresponding DCAT property (e.g., `spatial` key). However, there are also keys where the DCAT specification lacks suitable properties. For instance, the CKAN harvesting extension adds references to the source-datasets. This keys could be modelled in DCAT by using for example a owl:sameAs property.

"Extra" keys potentially duplicate existing CKAN keys: Some of the most common extra keys do not add additional information to the metadata since there is existing, semantically equivalent metadata keys, e.g., `contact-email` and the existing `maintainer_email` key. In this case, additional DCAT constraints to avoid redundant/duplicated mappings would increase the quality of the output metadata.

DCAT specification defines distributions as versions of the same content in different representations: However, certain CKAN datasets group resources by other (reasonable) semantic relations and a direct mapping to DCAT is therefore not possible. To model these datasets appropriate, DCAT has to allow other dimension descriptions for the distribution level. For instance, in order to model distributions within a dataset for differing time spans, the distribution level should allow a dct:temporal property.

Acknowledgements This work has been supported by the Austrian Research Promotion Agency (FFG) under the project ADEQUATe (grant no. 849982).