VU Research Portal

Towards a standard-based open data ecosystem: analysis of DCAT-AP use at national and European level

Barthelemy, Florian; Cochez, Michael; Dimitriadis, Iraklis; Karim, Naila; Loutas, Nikolaos; Magnisalis, Ioannis; Comet, Lina Molinas; Peristeras, Vassilios; Wyns, Brecht

published in
Electronic Government
2022

DOI (link to publisher)
10.1504/EG.2022.121856

document version
Publisher's PDF, also known as Version of record

document license
Article 25fa Dutch Copyright Act

Link to publication in VU Research Portal

citation for published version (APA)
Barthelemy, F., Cochez, M., Dimitriadis, I., Karim, N., Loutas, N., Magnisalis, I., Comet, L. M., Peristeras, V., & Wyns, B. (2022). Towards a standard-based open data ecosystem: analysis of DCAT-AP use at national and European level. *Electronic Government*, *18*(2), 137-180. https://doi.org/10.1504/EG.2022.121856

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:
vuresearchportal.ub@vu.nl

Download date: 14. Oct. 2023

Towards a Standard-based Open Data Ecosystem: Analysis of DCAT-AP use at National and European Level

Florian Barthélémy[a], Michael Cochez[b], Iraklis Dimitriadis[b], Naila Karim[b], Nikolaos Loutas[a], Ioannis Magnisalis[c], Lina Molinas Comet[b], Vassilios Peristeras[c], Brecht Wyns[a]

[a]*PwC Belgium*
[b]*Fraunhofer/FIT, Germany*
[c]*International Hellenic University, Greece*

Abstract

In Europe, an open government data ecosystem is being developed. This ecosystem is implemented using various technologies and platforms. In fact, the use of a common metadata standard for describing datasets and Open Data portals, i.e. the DCAT-AP specification, appears as the *lingua franca* that connects an, otherwise, fragmented environment. In this context, the standard-based consolidation of Open Data promotes the subsidiarity principle, allowing Open Data portal owners to choose platforms and internal representations based on their specific requirements. However, the portal owners must provide an export with DCAT-AP compliant metadata about the dataset they store. In this paper we provide a detailed study of how the DCAT-AP specification is used in practice, both at the national and the European level. Consequently, we also identify issues, challenges, and opportunities for improvements that can be used as input for the next revision cycle of the standard. Essentially, our goal is to contribute towards the enrichment of a growing and promising European Open Data ecosystem.

Keywords: open data, metadata, data standards, open data portals, interoperability, DCAT, DCAT-AP

1. Introduction and Scope

In this part, we first provide some background information for the dynamic Open Data ecosystem in Europe along with an introduction to the current fragmentation challenge. Commonly used metadata standards can facilitate Open Data interoperability and create a more coherent Open Data environment. Therefore, we present existing standards and their adoption by governments and developers. Last, we present the goals and the scope of this paper.

1.1. Background

Governments possess a large amount of basic data which is of critical value for society, not only with respect to the economic (e.g. [1] [2]) but also to the social aspect (e.g. [3] [4]). In this context, governments of different countries all around the world are developing policies to release this kind of data as Open (Government) Data [5] [6]. In 2003, the European Union (EU) adopted legislation to promote the re-use of Public Data in Member States via the Public Sector Information (PSI) Directive 2003/98/EC [7], which was lately revised in 2013 (Directive 2013/37/EU) [8]. As a result of the revision, the main amendments are the adoption of the "open by default" principle; the breakaway from cost-based charging for PSI towards a marginal cost-oriented fee (therefore the transparency regarding calculation of the fees is increased); the inclusion of certain cultural institutions as public sector bodies (previously outside the scope); and support for machine-readable documents and open formats.

In response to the modifications on the revised PSI directive, European public administrations have set up cross-domain and cross-organisational Open Data portals. Therefore, these portals have significantly contributed to the establishment of the necessary foundation for a European Open Data ecosystem, creating value and paving the way towards data-driven governments [5], while some of them already support and publish Linked Data [9] [10]. Additionally, research efforts also exist in order to lift open government portals to the web, specifically with the use of web data standards [11].

Nevertheless, in real implementations of data portals, limitations appeared soon enough [12] [13]. As a matter of fact, the inherent political, cultural, and linguistic diversity in Europe was not the only challenge to overpass. In addition to this, the development of Open Data portals has not always been coordinated neither within or across countries. Moreover, the use of different platforms and the lack of common semantics, metadata [14] and data models have resulted in a fragmented landscape of Open Data portals as disconnected information silos was created. Hence, making it hard to exchange metadata between the portals and limiting their ability to interoperate [15]. What is more important from the user perspective, citizens and businesses have to query over 150 separate Open Data portals in Europe if they want to find and combine data and information referring to Europe as a whole. This situation leads, not only to duplication of information and inconsistencies, but it also hampers cross-portal search and limits the discoverability of datasets. Datasets described with different metadata is a hindrance for interoperability across catalogs.

Overcoming the challenges described above, while respecting the subsidiarity principle, was only possible if the different portals with different descriptions of metadata would adhere to a common metadata language. Indeed, a common metadata language.

1.2. Vocabularies and Metadata Standards for Open Data

1.2.1. DCAT-AP

To improve the fragmented open data environment in Europe, a common metadata language, DCAT-AP (DCAT Application Profile), was developed under the Interoperability Solutions for European Public Administrations, Businesses and Citizens (ISA[2]) Programme of the European Commission. More specifically under its action on promoting semantic interoperability among the European Union Member States (SEMIC). It is mainly based on the Data Catalog Vocabulary (DCAT), which was developed initially at the Digital Enterprise Research Institute in Ireland [16] [17] and became later a W3C recommendation under the responsibility of the Government Linked Data Working Group[1]. DCAT is an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web. In this respect, an Application Profile is a specification that re-uses terms from one or more base standards, and adds more specificity. That is done by identifying mandatory, recommended, and optional elements to be used for a particular application, as well as recommendations for controlled vocabularies to be used.

Furthermore, it is possible to extend the DCAT-AP in order to meet local needs of implementers. In this context, the use of such extensions is indicated by [18], where the authors use several web vocabularies to generate metadata from Open Data portals which are available as Linked Data. Likewise, extensions have already been developed by the EC/ISA[2] Programme for the fields of geospatial [19] and statistical [20] data. The first one, the GeoDCAT-AP specification, was developed under the coordination of the Joint Research Center team responsible for the implementation of the INSPIRE Directive. While the second one, StatDCAT-AP specification, was part of the work under the coordination of EUROSTAT. Generally speaking, any DCAT-AP extension should follow strict guidelines. Essentially, the main aspect of these guidelines is that any data created according to a DCAT-AP extension must also be valid DCAT-AP data. Besides, properties and classes added should not have names which can be easily confused with those from DCAT-AP. Finally, the DCAT-AP "optional" or "recommended" properties can be redefined as optional, recommended, or mandatory in the extension, or they can be completely removed. In fig. 1 we present a general overview of DCAT-AP as a UML diagram showing the complete structure of its classes and corresponding properties in order to give a better understanding of the underlying specification and its structure.

Essentially, DCAT-AP allows the exchange of descriptions of data sets among data portals with the purpose of favouring data sets explorations in a cross-portal environment.[3] DCAT-AP has become a de facto standard for open data catalogs in Europe and beyond. By now, it is implemented by the European Data Portal [21], 15 national data portals, and also by several other portals at regional and local level [22]. Implementations of DCAT can also be found outside Europe, e.g. in Canada.[4]

[1] W3C. Government Linked Data (GLD) Working Group http://www.w3.org/2011/gld/wiki/Main_Page

[2] Precise guidelines on how to create a valid DCAT-AP extension are available https://joinup.ec.europa.eu/node/150345/

[3] https://joinup.ec.europa.eu/release/dcat-application-profile-data-portals-europe-final

[4] https://open.canada.ca/data/en

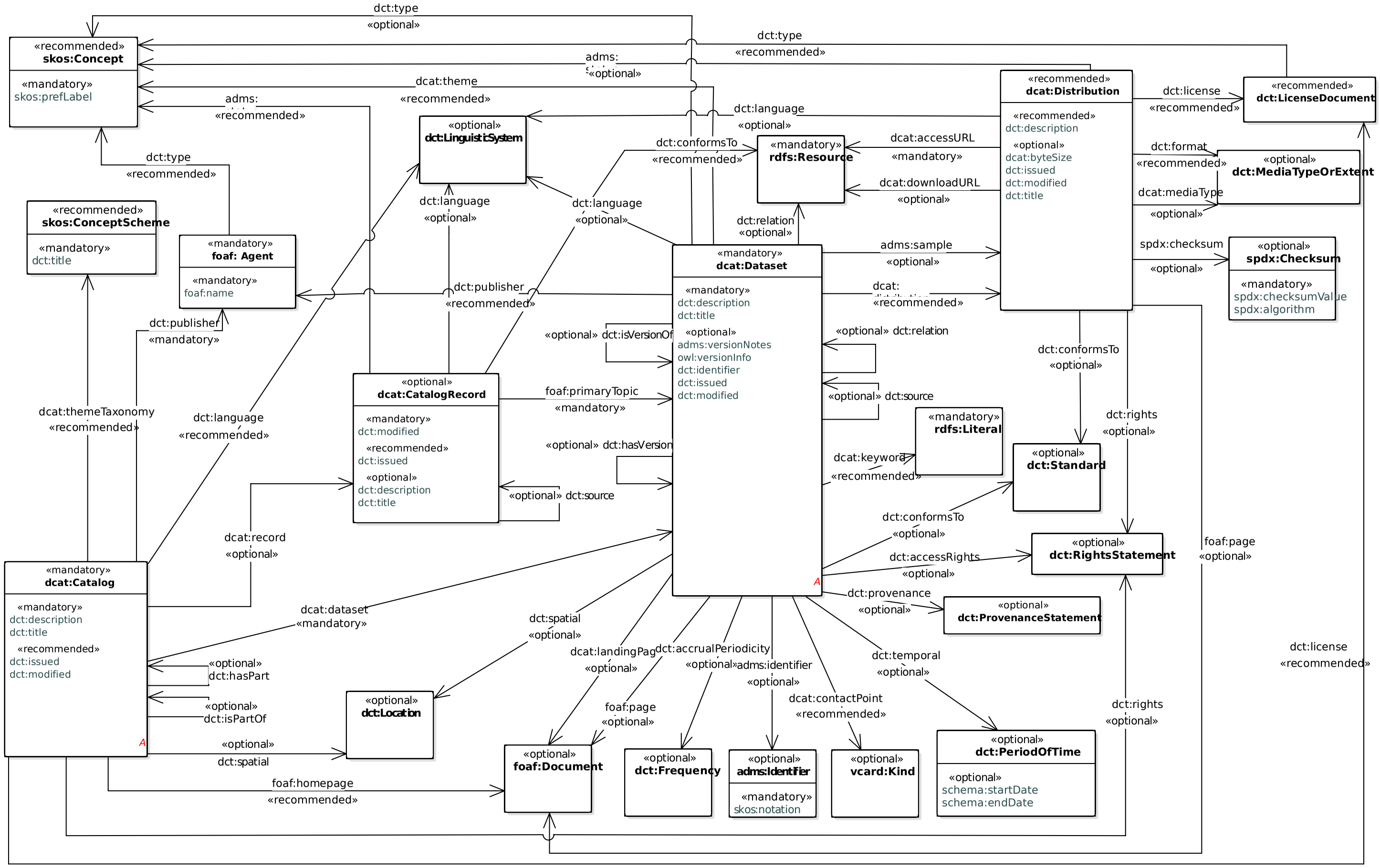

Figure 1: Classes and corresponding properties of DCAT-AP Adapted from the DCAT-AP 1.1 specification document `https://joinup.ec.europa.eu/release/dcat-ap-v11`

1.2.2. Other vocabularies and metadata standards

The other two specifications that are in widespread use are the CKAN platform native metadata schema [23] and the schema.org [24] dataset description vocabulary. These two specifications are discussed below. Additionally, other standards currently used for open data include, among others, the ISO 19115 [25] and INSPIRE metadata standards for geographic information and services [26]; the Data Documentation Initiative specification[5] (DDI) for data produced by surveys; the SDMX standard for statistics [27]; CERIF for research data [28]; VoID for linked datasets [29] and, in the healthcare and life sciences domain; the Dataset Description vocabulary and DATS [30]. Moreover, other Open Data Catalogs (ODC) tools[6] (i.e. similar to CKAN) define and natively use their own metadata to describe the datasets they store. The leading open data platforms, in addition to CKAN include DKAN, Junar, OGPL, and Socrata, all of them using their own metadata formats.

From these tools, CKAN is probably the most popular ODC. As CKANs harvesting functionality can be used to pull in metadata from other data portals, CKAN can be used to create a federated network of data portals which share data between each other. This is useful if, for example, a national portal aggregates information from CKAN instances of local governments, or if a topic-specific CKAN instance aggregates a subset of datasets from other CKAN sources. By default, CKAN provides a rich set of its own metadata for each dataset[7]. Moreover, CKAN supports DCAT[8], so when using DCAT, metadata can be federated from other non-CKAN but DCAT-compliant catalogs. For example, the European Data Portal has deployed its own extensions upon CKAN[9,10,11] which harvests data from various ODCs. Furthermore, even though CKAN is an open source project, DCAT main advantage upon the CKAN native metadata format is that DCAT is an open standard. This means that DCAT is technology neutral and agnostic. Besides, it provides a common language for connecting and aggregating data coming from diverse platforms. This was also the main reason why Germany decided to use a DCAT-based Application Profile as their national open data standard, contrary to initial thoughts to use CKAN [31].

Even though CKAN is widely used by government portals, it introduces some problems in the mapping of data when exporting to the DCAT standard. The main problems observed [32] are the following:

- Metadata particular to the government portals implementations are supported by CKAN, as additional keys and are used to map data from CKAN

[5] Data documentation initiative specification (2012) http://www.ddialliance.org/Specification

[6] Technology Options for Open Government Data Platforms. World Bank http://opendatatoolkit.worldbank.org/en/technology.html

[7] CKAN metadata

[8] Via an extension ckanext-dcat `https://github.com/ckan/ckanext-dcat`

[9] GitHub - ckanext-edp

[10] Understanding the European Data portal (report)

[11] Source Code of the European Data Portal

to the DCAT model. Yet, assuring that the conversion from the CKAN metadata to the DCAT version will contain the supplementary metadata is not possible.

- The additional metadata used by CKAN, and necessary for the mapping to DCAT, are duplicating some metadata in CKAN without increasing the value of the corresponding data.
- While DCAT does not support clustering of different content in only one dataset, CKAN allows this by linking content that is not exactly the same but which is semantically alike.
- There does not exist a complete mapping for the commonly used additional CKAN keys to the corresponding DCAT property even if that property already exists.

The other extensively used tool is schema.org, which is a vocabulary project initiated by the search engines Google, Microsoft, Yahoo and Yandex. The schema.org vocabularies are advertised to be developed following an open community process, e.g., using a mailing list and through GitHub. In addition, schema.org has recently offered an extension to facilitate semantic annotation of datasets [24]. It also claims to be based upon W3C DCAT work, and benefit from collaboration around the DCAT, ADMS and VoID vocabularies. Particularly, the focus of this effort is to provide metadata annotation to be used by data publishers on the web with the purpose of improving the performance of search engines. However, the overall standard development process in schema.org is not compatible with existing definitions of "open standards"[12] process, mainly due to the closed-door initial development and the semi-open decision process and specification ownership.

In relevant literature, there are studies that investigate the uptake and use of vocabularies and metadata standards. The uptake in embedding annotations from schema.org into web pages is driven by the fact that this data is consumed by search engines, increasing with this the discoverability of the annotated webpages [33]. In this context, a study in 2014 revealed that only 0.3% of websites are making use of the Google approved schema.org tools which help webmasters and online marketers present extra information to the search engines, despite the fact that over a third of Google search results could benefit from rich snippets supported by schema.org[13]. We do suspect that the uptake has increased significantly since, but also that still far from all web resources are annotated. In another study [34], the authors analyzed how the tourism industry and specifically hotels are using schema.org. More specifically, they study the usage and distribution of schema.org, how it is applied and whether or not the classes and properties of the vocabulary are used in a syntactically and semantically correct way.

[12] Definition of "Open Standards" https://www.itu.int/en/ITU-T/ipr/Pages/open.aspx

[13] Schema.org analysis report 2014

However, studies that compare use and uptake of metadata standards explicitly for open data are scarce[14]. An exception is the International Benchmark Open Data and Use of Standards conducted in 2012, in a pre-DCAT era where "Many Open Data proponents resist advocating specific data formats, feeling that standardization at that level may be premature". Obviously, the same discussion has become very relevant nowadays with the plethora of Open Data platforms and implementations all over the world.

More generally, in the process of moving open government data towards Linked Data on the Web, bottlenecks include not only the discovery of data, but also ontological alignment and enhancement, interfaces and tools for moving Open Data towards Linked Data, and dataset consumption methods from citizens or developers[15] (e.g., APIs).

1.3. Scope of This Work

In the current work, we investigate the way in which the DCAT- AP specification, the de facto standard for describing open datasets and open data catalogs, is put into practice and used in Europe. Primarily, we try to understand better the real-world usage as well as the challenges related to this important initiative to create a European Open Data ecosystem based on an open standard. We do this not only from a standardization outlook but also from a practical usage perspective. Our analysis looks at both national Open Data portals, with a focus on those which have adapted the DCAT-AP recommendation; and the European Data Portal, which collects data from many national portals. Overall, we found that DCAT-AP is used by already many and an increasing number of portals. Moreover, the data collected from the portals are reasonably in accordance with the standard. However, we also found that further work is still needed in order to unify or avoid fragmentation which may be caused by national portals and their extensions. Another aspect observed is that often pragmatics win over being interoperable. This is for example the case when portals present their data in various incompatible formats.

This paper is structured as follows: in the next section we present the methodology used for collecting the data for this study. Then, in Section 3, we describe our findings in both, national portals and the European Data Portal. Afterwards, in 4 we discuss the findings explaining the properties' usage on different classes. At last in Section 5 we present our conclusions as well as the eventual work that can be done.

Additionally, it is also important to mention that part of this work was published earlier as a technical reports.[16]

[14] Towards a better supply and distribution process for Open Data: Case study, International Standards

[15] Linked Open Government Data: Lessons from Data.gov.uk https://eprints.soton.ac.uk/340564/2/Linked%2520OGD.pdf

[16] `https://joinup.ec.europa.eu/document/national-extensions-analysis-dcat-ap` and `https://joinup.ec.europa.eu/document/report-dcat-ap-use`

2. Methodology

In this section we shed light on the methodology we followed in order to analyze the usage of DCAT-AP concerning two aspects, namely:

- usage in real-world of DCAT-AP meta-data elements (i.e. class, properties) in living datasets marked-up. More precisely, on the European Data Portal, which is in charge of collecting (i.e., harvests) the metadata from national portals in order to improve the accessibility and increase the value of Open Data.[17].
- degree of divergence of modelling national profiles when compared to EU published DCAT-AP v1.1. Furthermore, the analysis of the techniques used by the data portals to publish their metadata using these profiles.

To deal with these two dimensions we followed a mixed-methods [35] approach. Our methodology follows aspects of exploratory data analysis [36] in an attempt to verify results by triangulation [37] and eventually synthesize conclusions under a specific frame of reference. In figure 2 we showcase in detail this methodology.

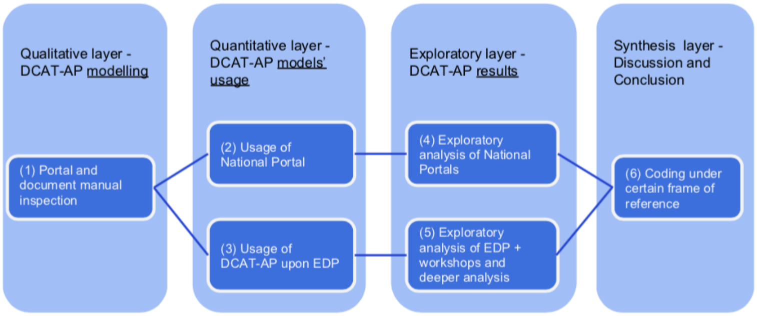
Qualitative layer - DCAT-AP modelling
(1) Portal and document manual inspection
Quantitative layer - DCAT-AP models' usage
(2) Usage of National Portal
(3) Usage of DCAT-AP upon EDP
Exploratory layer - DCAT-AP results
(4) Exploratory analysis of National Portals
(5) Exploratory analysis of EDP + workshops and deeper analysis
Synthesis layer - Discussion and Conclusion
(6) Coding under certain frame of reference

Figure 2: Aspects of our mixed-methods based methodology

[17] About the European Data Portal https://www.europeandataportal.eu/en/what-we-do/our-activities

The qualitative analysis of national portals was conducted on a manual basis ((1) of figure 2). For the purpose, specific portals were selected on the basis of claiming DCAT-AP compliance and availability of descriptions concerning modeling of national extensions, whether documented or just implemented. Thus, in this stage we manually inspected national portals, examined documentation concerning DCAT-AP modeling, downloaded national DCAT-AP descriptions in any available format (in all cases investigated, the most complete description was in textual format) and checked for differences with the original DCAT-AP v1.1 specification. A team of three people performed the actual work of analyzing the available information within a period of 6 months (1/1/2017-30/6/2017). Further details on this part of the work performed are given in Section 2.2.

Additionally, quantitative analysis was performed in order to collect and evaluate data for the other part of our work, namely the actual usage of the identified DCAT-AP national profiles. Two parallel paths of work were steered in order to collect relevant data. Firstly ((2) of figure 2), we selected portals with a national profile and where a data dump in a standard RDF format was available. In order to avoid inefficient visualization of data, we depict charts for meta-data properties which were used by at least one portal on at least a threshold (i.e. 10%) that statistically depicts a sample of importance and which is free of outliers. Descriptive statistics and exploratory analysis were employed, resulting in simple chart visualizations and percentages of class and properties usage. A team of three people performed the actual work of analyzing the available information within a period of 6 months (1/1/2017-30/6/2017). The team used a tool to examine RDF dumps to count usage of element items from the nationally proposed schema.

In this context of quantitative analysis, ((3) of figure 2), we have followed a specific method of data collection following the methodology described in detail within Section 2.1. In short, we have gathered data from the EDP portal on (harvested) semantic descriptions of datasets throughout Europe. A team of three people performed the actual work of analyzing the available information to 10/10/2017. Particularly, the outcomes of this analysis raised a number of questions for the experts of the DCAT-AP working group. We discuss them in Section 2.1.

During collection and exploratory analysis of the data we coded results found about DCAT-AP usage from class and properties usage viewpoint. Thus, we could derive a holistic view, presented in discussion section, where we organize findings structured according to the most prominent DCAT classes, namely Catalog(Record), Dataset and Distribution. Further, we present findings which involve multiple classes. Triangulating results from our quantitative and qualitative analysis allowed us to identify the most important and verified findings and reach relevant proposals, as we present them in Section 5.

2.1. Analysis of DCAT-AP Use via the European Data Portal

In order to analyse the real use of DCAT-AP classes and properties on the European data Portal (EDP), SPARQL queries were developed and executed on its SPARQL endpoint i.e. `https://www.europeandataportal.eu/sparql`.

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dct: <http://purl.org/dc/terms/>
SELECT COUNT(DISTINCT ?s)
WHERE {
    ?s a dcat:Dataset .
    ?s dct:description ?o
}
```

Figure 3: Query used to answer 'How many datasets have a description?'. This particular query resulted in 728196 datasets with a description.

Given the high number of properties to analyse, the creation of queries was semi-automated using a spreadsheet [18]. Moreover, the queries were also transformed into URLs and results were automatically loaded. The queries can be used to monitor the DCAT-AP use or similar models on every portal that has a SPARQL endpoint. Figure 3 illustrates the query we used to answer the question 'How many datasets have a description?'.

Particularly, the outcomes of this analysis raised a number of questions for the experts of the DCAT-AP working group. As a consequence, these questions were presented and discussed during two webinars on 19/10/2017[19] and on 08/12/2017[20]. The comments received during those webinars are included in the explanations in section 4 in subsections on vocabulary use. Moreover, in order to further continue the discussion online, topics were created in the DCAT-AP repository on GitHub[21], where members of the DCAT-AP working group provide their views on the use of different properties and classes and, consequently, on how these elements should be tackled in the next version of DCAT-AP. Besides the analysis of the use of DCAT-AP classes and properties, a more detailed analysis was conducted for the properties having a controlled vocabulary specified in DCAT-AP. For those properties, targeted at large catalogs that use the properties, a subset of metadata was extracted from the EDP SPARQL endpoint as N-triples. Specifically, the subset represents 5552 datasets from 6 catalogs harvested by the EDP. The catalogs were selected based on several criteria: they should have a significant size to ensure a good representation of all properties that use code lists, they should represent different content types to avoid an over-representation of geospatial data sets and they should be geographically spread across the European Union. For each property supposed to

[18] The spreadsheet with queries and results is available at `https://joinup.ec.europa.eu/sites/default/files/document/2018-04/SPARQL%20query%20results%20from%20the%20European%20Data%20Portal.xlsx`

[19] https://joinup.ec.europa.eu/event/dcat-ap-change-management-release-policy-webinar-19-october-2017-1400-cet

[20] https://joinup.ec.europa.eu/event/change-and-release-management-policy-dcat-ap-final-webinar-8-december-2017-1000-cet

[21] https://github.com/SEMICeu/DCAT-AP

take a value from a controlled vocabulary, two validation checks were done:

- confirm the use/non-use of the properties analysed;
- if the property is used, verify if the value is correctly using the controlled vocabulary or not.

For the second check, the DCAT-AP validator[22] was updated to run validation rules on the metadata. The validation rules analysed automatically if properties start with a specific code list URI, such as `http://publications.europa.eu/resource/authority/data-theme/` for the property *dcat:theme* or if the URI contains the correct term category for *dcat:mediaType* from the list maintained by IANA[23]: `application/`, `audio/`, `font/`, `image/`, etc. For the full overview of properties and code lists analysed, refer to section 3.2.

2.2. Analysis of National Extensions

In the first part we investigated the changes proposed in the formal specifications of national profiles. After that, we examine the actual use of these extensions.

The change analysis of formal specifications is performed on formal specifications and standardization documents collected from the National DCAT-AP portals shown in Table 1. These countries were selected in our analysis because they have published a DCAT-AP specification.

Table 1: National DCAT-AP profiles

Belgium - Fedict, OpenKnowledgeBE, Web Address: `http://dcat.be/` The information in this analysis is based on communication with Fedict. There is no specific information on the website.
Germany - Finanzbehrde - Geschfts- und Koordinierungsstelle GovData, Web Address: `http://dcat-ap.de/def/`, Version/Update Date: V1.0 2017-06-21
Ireland-Open Data Unit-Dept of Public Expenditure & Reform, Web Address: `https://data.gov.ie/technical-framework`, Version/Update Date: 2015-06-01
Italy - AgID - Agenzia per l'Italia Digitale, Web Address: `https://linee-guida-cataloghi-dati-profilo-dcat-ap-it.readthedocs.io/it/latest/`, Version/Update Date: Release 1.0 2017-04-09 Revision 4e3c5e31
The Netherlands - Kennis- en Exploitatiecentrum Officile Overheidspublicaties (KOOP), Web Address: `http://dcat-nl.info/nl/latest/`, Version/Update Date: V 1.1 2017-06-01 Revision 120bc7b7

[22] http://dcat-ap.semic.eu/dcat-ap_validator.html

[23] https://www.iana.org/assignments/media-types/media-types.xhtml

Norway - Agency for Public Management and eGovernment (Difi), Web Address: `https://doc.difi.no/dcat-ap-no/`, Version/Update Date: 2016-10-11
Spain - APORTA INITIATIVE, Web Address: `http://datos.gob.es/es/documentacion/guia-de-aplicacion-de-la-norma-tecnica-de-interoperabilidad-\de-reutilizacion-de`, Version/Update Date: 2016-07-28
Sweden VINNOVA, Web Address: `https://docs.google.com/document/d/17-vEfZXlu9kykcmjXZo1_Z8QKkr7-Prgwd6YUKLRrjk/edit`,(restricted access), Version/Update Date: 2016-06-07
Switzerland - Open Government Data Switzerland, Web Address: `https://handbook.opendata.swiss/en/library/ch-dcat-ap`, Version/Update Date: 2016-02-09

The change analysis had to be done manually as for many portals no automatically analyzable data formats were available, e.g., a schema file. Moreover, we noticed that even when a formal specification was available as an OWL ontology, there were still more restrictions in the textual specification. Some of these inconsistencies appear due to limitations in the expressive power of OWL. Another finding is that even though some countries maintain a list of updated, added and deleted properties to DCAT-AP, others have added additional properties or made updates to the standard DCAT-AP, but did not explicitly document the changes. As a consequence, the change identification was rather a tedious task.

As a result of our investigation, we created an analysis report by comparing each class and property from the national profiles with the usage notes, cardinality, and status (optional, recommended, or mandatory) in DCAT-AP. Moreover, from this change analysis report, we identified updates in properties ranges as well as the removal of properties and classes.

Before presenting the discussion on the property updates, we present a brief overview of the identified changes per country. In this context, we also identify and comment properties updated in many extensions, for example language, licensing, media types, and format. Furthermore, we also inspect whether updates are in accordance with the DCAT-AP extension rules. As for the non-compliant updates, we discussed them in part 4.

Afterwards, we analysed the actual use of DCAT-AP in these portals. This proved more difficult than anticipated. The main reason of the difficulty was that some of these portals do not provide their metadata in any of the standardized RDF serialization formats, but rather in either a platform specific or own customized format. We consider this fact as an interesting finding. As we focus on interoperability, we decided to only analyse the portals which provide

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?prop, count(?prop)
FROM <http://localhost:8890/DCAT/Norway> WHERE {{
  SELECT distinct ?DS ?prop WHERE {{
    SELECT distinct ?dataset AS ?DS WHERE {{
      ?catalog a dcat:Catalog.
      ?catalog dcat:dataset ?dataset .
    } UNION {
      ?dataset a dcat:Dataset.
    }}
  } ?DS ?prop ?val}
}} group by ?prop
```

Figure 4: The query used to find the properties used for dcat:Dataset in the portal of Norway. The result of the query is all properties found and the number of dcat:Datasets having this property at least once. This particular query resulted in, for example, *dcat:keyword* – 512, meaning that 512 datasets (out of 550) specified at least one keyword. Note that if we would just count the number of dcat:keyword properties for all datasets, we would have obtained 1670, but it would not be clear how these are distributed over the datasets.

their metadata in a standardized way. That was the case for Ireland, Norway, Spain, Sweden, and Switzerland.

For these countries, we downloaded the available metadata and measured how often properties on the DCAT-AP Catalog, Dataset, and Distribution are used. To be specific, we measured the fraction of instances which had at least one occurrence of a property. So, when we show the statistic for the *dct:title* property on the Dataset class, it means that that fraction of the Datasets had at least one value for that property. For DCAT-AP properties which were used by at least one portal on at least 10% of the instances, we show detailed charts for all portals, also for these which have less than 10% usage of the property. If there is no bar for a given property, that means that the country does not use the property at all. On the contrary, other properties are merely listed. Then, for the additional properties used in the respective portals, we also show statistics.

Figure 4 illustrates the query we used to answer the question ‘How many datasets have a description?’.

2.3. Data and Tool Availability

The tools and data used in this work consists of

- The spreadsheet used to collect the changes made in national profile specifications.
- The SPARQL queries used to gather the use of properties on national profiles

- The spreadsheet used to collect the outcomes of these queries
- The SPARQL queries used to find out the use of properties on the EDP. These are included in a spreadsheet with their results.
- The Controlled vocabularies used for comparison, which are listed in table 2.

All of these resources are made available for future investigation from `http://datalab.rwth-aachen.de/DCAT/tools.zip`[24]

2.4. Limitations

The chosen methodology allows to analyse the use of DCAT-AP from an external viewpoint, based on openly available information. This leads to a number of limitations, which need to be taken into account when interpreting the results presented in this paper.

As the national extension analysis is based on formal descriptions of DCAT-AP extensions, the study does not cover countries that do not explicitly specify an extension. A national implementation of DCAT-AP in which additional properties are used, is thus not included in this study, unless a formal description of the extension exists.

The EDP harvests the metadata from data portals in Europe. These data portals use DCAT-AP in different ways: some are not using DCAT-AP at all, some follow a national DCAT-AP profile and some claim to be fully-compliant with DCAT-AP. The EDP maps these varied uses of DCAT-AP to their own use of DCAT-AP, and sometimes populates the values automatically. These mappings and additions influence the analysis conducted in this study. By querying the EDP endpoint, this analysis assesses only partially the use of DCAT-AP for all data portals harvested by the EDP, as some properties and values from national portals might not be harvested, and some nationally used metadata schemas are transformed into the EDP's DCAT-AP schema. Moreover, the study does not investigate how metadata is stored "under the hood" of the data portals, for example when the CKAN schema is used in the software, and DCAT-AP is used on top as a representation layer.

3. Overview of Collected Data and Empirical Findings

3.1. Use of DCAT-AP Classes and Properties on the European Data Portal

The collection of data was performed on 10/10/2017 following the methodology described in section 2. The results of the queries show that the EDP publishes information about 80 catalogs. These catalogs contain all together 743.746 datasets and 982.771 distributions. Each of the following sub-sections gives an overview of the use of the measured occurrence of properties for the DCAT-AP classes: *dcat:Catalog, dcat:Dataset, dcat:Distribution* and *dcat:CatalogRecord.*

[24] For the camera ready version of this paper, the tools and data will be placed persistently on zenodo.

3.1.1. dcat:Catalog

The use of mandatory, recommended and optional properties for the *dcat:Catalog* class is presented respectively in figs. 5 to 7. The results are further discussed in section 4.

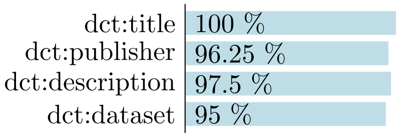

Figure 5: Use of mandatory properties of the class Catalog in the EDP

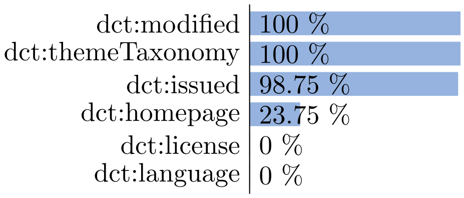

Figure 6: Use of recommended properties of the class Catalog in the EDP

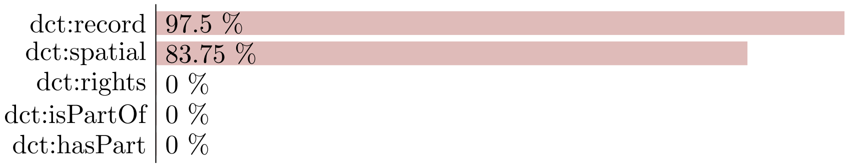

Figure 7: Use of optional properties of the class Catalog in the EDP

3.1.2. dcat:Dataset

Figures 8 to 10 present the use of mandatory, recommended and optional properties on the *dcat:Dataset* class. The results are further discussed in section 4.

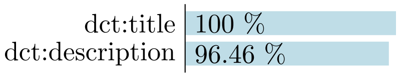

Figure 8: Use of mandatory properties of the class Dataset in the EDP

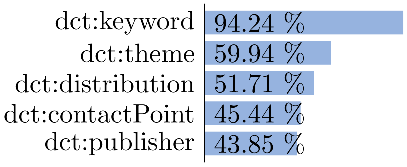

Figure 9: Use of recommended properties of the class Dataset in the EDP

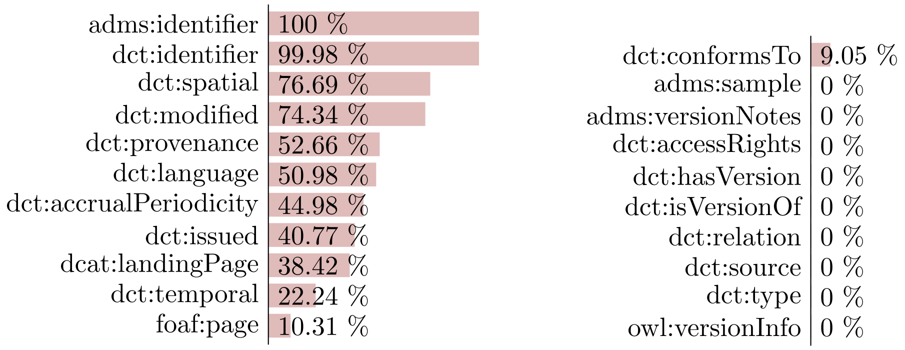

Figure 10: Use of optional properties of the class Dataset in the EDP

3.1.3. dcat:Distribution

For dcat:Distribution, the only mandatory property is dcat:accessURL. In our analysis of the EDP, we found out that 99.07% of the distribution indeed has this property. Figures 11 and 12 represent the use of optional and recommended properties for the *dcat:Distribution* class. The results are further discussed in section 4.

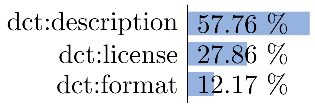

Figure 11: Use of recommended properties of the class Distribution in the EDP

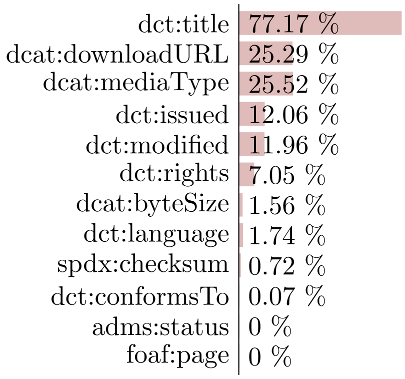

Figure 12: Use of optional properties of the class Distribution in the EDP

3.1.4. dcat:CatalogRecord

A Catalog Record is defined as *a description of a datasets entry in the catalog*. The analysis shows a 100% use of mandatory properties, as well as the recommended properties *dct:issued* and *adms:status*. Other properties are not used at the Catalog Record level.

3.2. Use of Controlled Vocabularies on the European Data Portal

In DCAT-AP version 1.1, 15 properties are expected to use controlled vocabularies as summarised in table 2.

Table 2: Comparison between the list of properties supposedly using Controlled Vocabularies and the lists of properties present on the EDP

Property	**Class**	**Used on the EDP?**	**Controlled Vocabulary**
dct:language	Dataset	Yes	MDR Languages Named Authority List[25]
dcat:accrualPeriodicity	Dataset	Yes	MDR Frequency Named Authority List [26]
dct:theme	Dataset	Yes	MDR Dataset Theme Vocabulary[27]
dct:publisher	Dataset	Yes	MDR Corporate bodies Named Authority List[28]
dct:publisher	Catalog	Yes	MDR Corporate bodies Named Authority List
dct:format	Distribution	Yes	MDR File Type Named Authority List[29]
dcat:mediaType	Distribution	Yes	IANA Media Types[30]
dct:spatial	Dataset	Yes	MDR Continents Named Authority List[31], MDR Countries Named Authority List[32], MDR Places Named Authority List[33], Geonames[34]
dct:spatial	Catalog	Yes	MDR Continents Named Authority List, MDR Countries Named Authority List, MDR Places Named Authority List, Geonames

[25] http://publications.europa.eu/mdr/authority/language/

[26] http://publications.europa.eu/mdr/authority/frequency/

[27] http://publications.europa.eu/mdr/authority/data-theme/

[28] http://publications.europa.eu/mdr/authority/corporate-body/

[29] http://publications.europa.eu/mdr/authority/file-type/

[30] https://www.iana.org/assignments/media-types/media-types.xhtml

[31] http://publications.europa.eu/mdr/authority/continent/

[32] http://publications.europa.eu/mdr/authority/country/

[33] http://publications.europa.eu/mdr/authority/place/

[34] http://sws.geonames.org/

adms:status	CatalogRecord	Yes	ADMS change type vocabulary[35]
dcat:themeTaxonomy	Catalog	Yes	MDR Dataset Theme Vocabulary
dct:language	Catalog	No	MDR Languages Named Authority List
adms:status	Distribution	No	ADMS status vocabulary[36]
dct:type	Agent	No	ADMS publisher type vocabulary[37]
dct:type	LicenseDocument	No	ADMS licence type vocabulary[38]

The 4 properties not used on the EDP are not analysed in this section but have dedicated descriptions in sections 3.3.3 and 4:

- dct:language for the class *dcat:Catalog*. This recommended property is not used for the class *Catalog* on the EDP.
- adms:status for the class *dcat:Distribution*. This is not an issue as the property is optional.
- dct:type for the *class foaf:Agent*. This mandatory class defines the range of the properties publisher (dct:publisher) on Dataset and Catalog. It has two properties, foaf:name and dct:type. The use of the property dct:type is recommended and is used to refer to a type of the agent that makes the Catalog or Dataset available. The class *Agent* is extensively used even though the type property is not provided for this class.
- dct:type for the class *dct:LicenseDocument*. The recommended class *LicenseDocument* defines the range of the property dct:license on classes *Catalog* and *Distribution*. As the class is not used by the EDP, the property dct:type for indicating the type of licence is not used either.

We analysed and summarised the results of the 11 properties which are present on the EDP (marked with *Yes* in table 2). Automated analyses were conducted for 9 out of the 11 properties with the help of the DCAT-AP validator[39]. The dct:publisher property for classes *dcat:Dataset* and *dcat:Catalog* was, however, analyzed manually. In table 3 we sum up the findings for the different properties supposedly using controlled vocabularies. Moreover, the column 'Errors' indicates the amount of errors registered. An error is recorded when an

[35] The MDR Change Type Vocabulary was never created. This issue has been reported to the editors of DCAT-AP (https://github.com/SEMICeu/DCAT-AP/issues/45) and will be addressed in the next version.

[36] http://purl.org/adms/status/1.0

[37] http://purl.org/adms/publishertype/1.0

[38] http://purl.org/adms/licencetype/1.0

[39] http://dcat-ap.semic.eu/dcat-ap_validator.html

instance that is supposed to refer to a controlled vocabulary, is not referring or is wrongly referring to its controlled vocabulary.

Table 3: Number of errors found per property using controlled vocabularies

Property		Instances in sample	Errors
Distribution			
dct:format	Recommended	1859	**1859**
dcat:mediaType	Optional	1365	**55**
Dataset			
dct:publisher	Recommended	See section 4.4.1	
dcat:theme	Recommended	5009	**0**
dct:accrualPeriodicity	Optional	992	**0**
dct:language	Optional	445	**0**
dct:spatial	Optional	3002	**3002**
CatalogRecord			
adms:status	Recommended	5536	**5536**
Catalog			
dct:publisher	Mandatory	See section 4.4.1	
dct:themeTaxonomy	Recommended	6	**6**
dct:spatial	Optional	6	**6**

3.3. Analysis of the National Extensions to the DCAT-AP

DCAT-AP v1.1 covers a basic set of properties and classes for online metadata exchange. Yet, to fulfill specific national needs that are not covered by the basic set of properties and classes, DCAT- AP can be extended without losing its compatibility to other DCAT-AP extensions. This extension can include changes, additions or removals in cardinality, classes and properties. Therefore, it is common for extensions to introduce new mandatory properties or move properties from the set of recommended to the set of optional ones. Along with that, extensions can include restrictions by adding cardinality, language and range constraints or even restricting semantics.

3.3.1. Analysed DCAT-AP Extensions

The analysis includes the national DCAT-AP extensions for Belgium, Germany, Ireland, Italy, the Netherlands, Norway, Spain, Sweden and Switzerland, as detailed in Table 1. Moreover, in Figure 13 is shown the map of analysed countries, where the darker colour indicates the existence of more changes for the national application profiles. A change means here either an addition or removal of a class or property or a change in the constraints (including cardinality) of a property.

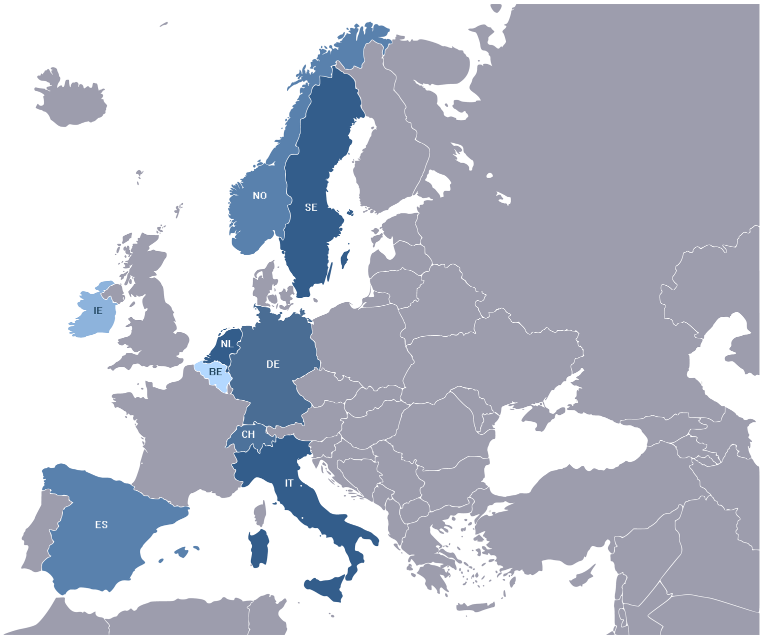

Figure 13: Countries with a national profile based on DCAT-AP. A darker colour indicates more extensive modifications and additions to DCAT-AP.

In order to better explain the changes made by the national profiles, we briefly describe and highlight below significant changes of each extension.

1. *DCAT-AP.be (Belgium):* Even though DCAT-AP.be does not include any additional properties, allows the data producers to add properties. In addition, each literal has a language tag, apparently to cater for the multilingual requirements in the Belgian context. Also, used keywords should map to the Data Theme Named Authority List of the Publication Office of the European Union[40]. Organisations in the metadata get an IRI which derives from the Belgian national company register. Moreover, the class *CatalogRecord* is not used in the Belgian extension.
2. *DCAT-AP.de (Germany):* The first version of the German extension of DCAT-AP focuses on copyright, licensing, and law restrictions corresponding to the German context. A second version was announced for 2018.
3. *DCAT-AP (Ireland):* The Irish extension is specified by a data exchange framework, which is based on the DCAT v1.0 profile. Here, the class *Dataset* is provided with some additional geospatial metadata properties. Some of them have been added to DCAT-AP v.1.1. However, the extension still includes some properties that are not part of DCAT-AP v.1.1.

[40] http://publications.europa.eu/mdr/authority/data-theme/

4. *DCAT-AP_IT (Italy):* The Italian metadata profile DCAT-AP_IT focuses on granulating the vCard:Kind class to enhance the information metadata for companies. To do so, the class *dcatapit:Organization* was introduced as a subclass of *vCard:Organization*, which is in turn is a subclass of *vCard:Kind*. Additionally to those changes, *dcatapit:Organization* provides new properties to describe the contact points name, email, telephone number, and URL (e.g., a homepage). Along with the changes to the class *vCard:Kind*, several other classes were provided with additional properties.
5. *DCAT-AP-NL (the Netherlands):* The Dutch extension of DCAT-AP is called DCAT-AP-NL and was updated in June 2017. The classes *Dataset*, *Agent*, *Catalog*, *Record* and *Distribution* were updated and include some additional properties like registration holder, language and identifier.
6. *DCAT-AP-NO (Norway):* In the Norwegian DCAT-AP extension, most of the changes took place in the class *Dataset* and *Distribution*. The extension provides additional subject, creator, and access right comment properties for the class *Dataset*. Additionally some properties have been added in order to describe relations between different datasets. DCAT-AP-NO also provides a new property dct:identifier for the class *Agent*. On the other hand, the property dcat:mediaType in the class *Distribution* was excluded, since it was considered to be the same as dct:format.
7. *DCAT-AP-Spain (Spain):* The Spanish DCAT profile was developed and created before DCAT-AP was published and therefore cannot be seen as real extension of DCAT-AP. Given that, the current Spanish standard includes some legacy features, which could cause some interoperability issues with other extensions. In the scope of this analysis we regard the Spanish profile as an extension, in fact, we describe the different violations against the DCAT-AP v1.1.
8. *DCAT-AP-SE (Sweden):* The DCAT-AP-SE extension of DCAT-AP was developed for ppnadata.se, the Swedish national portal for Open Data. One focus of the Swedish extension was licensing, while another focus was on restricting the *vCard:Kind* class to either *vCard:Individual* or *vCard:Organization*. In addition, further contact detail properties were added.
9. *CH-DCAT-AP (Switzerland):* The Swiss extension of DCAT-AP focuses on providing text elements in French, German, Italian and English. As a consequence, it is made mandatory that every text element is provided in each of these languages.

3.3.2. Property Changes

Adding, changing, or removing properties in extensions is common practice. In this overview we group these updates in mandatory, recommended/optional and exclusions of properties. A more detailed overview of how frequently properties are changed can be found in table A.6; the complete details can be found from the suplementary material (see section 2.3).

From the mandatory properties, *dct:identifier* was the one which changed the most. Indeed, four extensions changed this property. In addition to the changes listed in table A.6, there is a set of twenty-seven additional changes of mandatory properties. Since all of these were only updated in one profile, we do not discuss them in further detail. However, we review the *dct:license* property more carefully in part 4.3.2.

Next, we analyze also the most frequent changes on the recommended and optional properties. The difference between recommended and optional properties is not very strict. Either recommended or optional properties can be left unspecified, with the difference that recommended properties are suggested to be specified in order to fulfill the EU data interoperability standards. In contrast, optional properties can be left unspecified but could be used to provide additional information about the data and improve its value. In this respect, *dct:spatial* was the property which was changed most often. This includes three changes for the class *Dataset* and one in the class *Catalog*. Unlike most mandatory property changes, the modifications for *dct:spatial* vary between the different national extensions. These changes will be further discussed in Section 4. The most frequently changed properties are also included in table A.6. Besides these, there are additionally sixty-three other recommended and optional properties that were changed in only one of the analysed extensions. Hence, we will not discuss these any further.

In some cases, the exclusion of a recommended or optional property is also a valid option to improve the national extension of the DCAT-AP version. That is the case of DCAT-AP_IT, DCAT-AP-NL and DCAT-AP-NO, as shown in Table A.7, where all the listed properties where excluded. As a remark, the Italian extension DCAT-AP_IT has the most exclusions, while DCAT-AP-NL and DCAT-AP-NO only excluded properties for the class *Distribution*.

3.3.3. Analysis of the Use of Properties

In this section, we present the result from measuring the use of properties of the *Catalog*, *Dataset*, and *Distribution* classes in the selected national portals. As mentioned, we selected portals with a national profile and where a data dump in a standard RDF format was available. Using these criteria, we ended up with the national portals of Ireland, Norway, Spain, Sweden, and Switzerland. None of these countries use the *CatalogRecord* class. As mentioned, we show charts for properties which were used by at least one portal on at least 10% of the instances. However, all obligated and recommended properties matching these criteria are included in the charts. We give a short overview of the missing properties, since some portals rarely used them. All collected data, including the ones left out for sake of brevity in this paper, can be found in the supplementary material (see section 2.3).

Property use in the class Catalog

Here we present different charts in order to show the specific use of the properties in the class *Catalog*, this for every analyzed extension. As a reference, the color code used for the countries can be found in fig. 14.

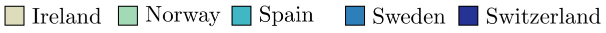

Figure 14: Colour code for the countries

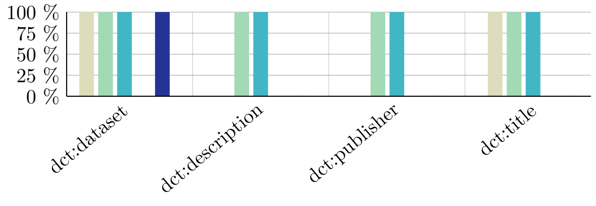

Figure 15: Mandatory properties in the Catalog class

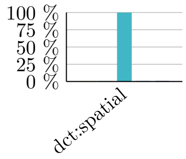

Figure 16: Optional properties in the Catalog class

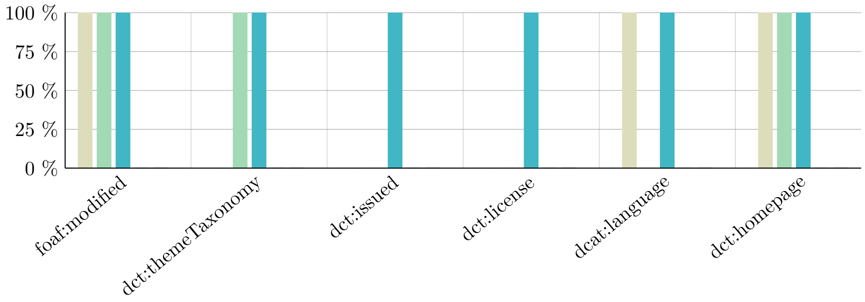

Figure 17: Recommended properties in the Catalog class

We observed that for the class *Catalog* all mandatory and recommended properties were frequently used in at least one of the portals, as shown in Figure 15 and 17 respectively. However, *dct:hasPart,dct:isPartOf, dcat:record, dct:rights* and *dct:spatial* were not used at all, and are correspondingly excluded from fig. 16.

Property use on the class Dataset

Figures 18 to 20 show the use of properties on the class Dataset. Besides the ones in the charts, there were also properties that were not used very often (i.e., in not more than 10% of the datasets in any given portal). In particular, the property *dct:conformsTo* rarely occurred in the Spanish and Swiss data dumps. Moreover, the Swedish data dump was the only one, that used *dct:isVersionOf, dct:provenance, owl:versionInfo* and *adms:versionNotes*, but also very infrequently. The properties *dct:hasVersion, adms:sample, dct:type* were not used at all.

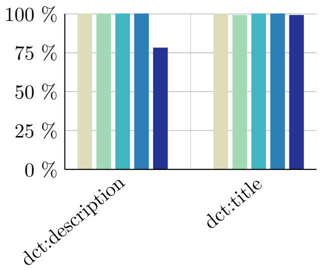

Figure 18: Mandatory properties in the Dataset class

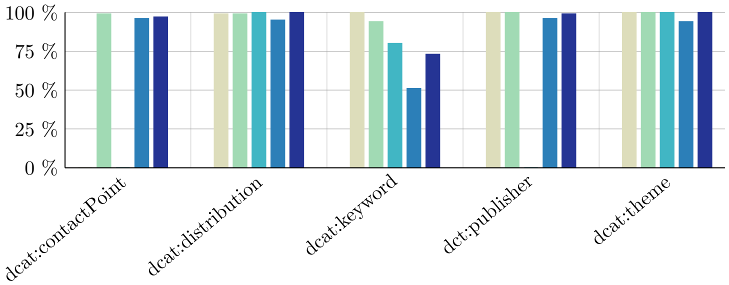

Figure 19: Recommended properties in the Dataset class

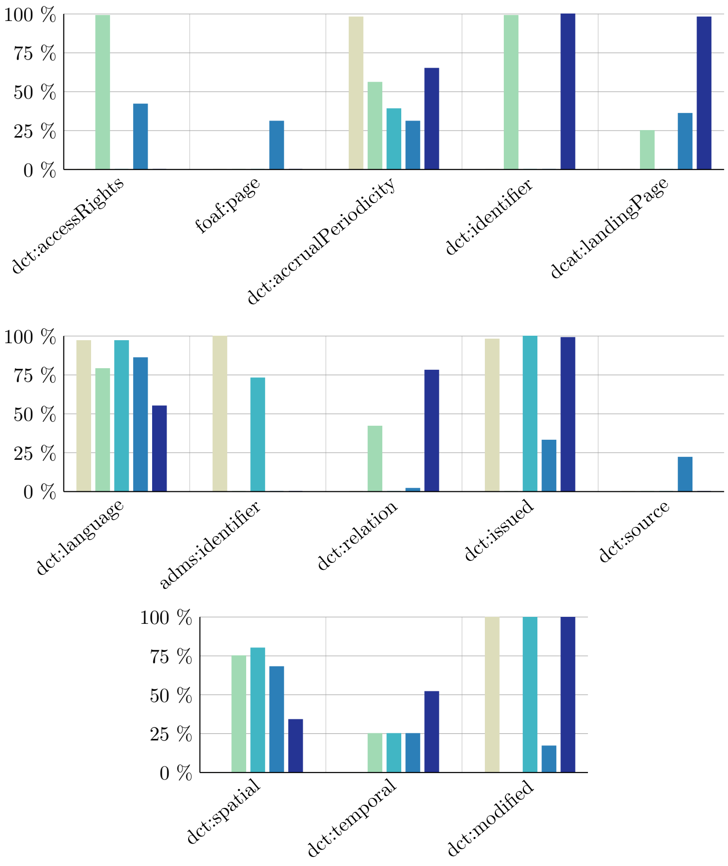

Figure 20: Optional properties in the Dataset class

Property Use on the Class Distribution

For the class *Distribution*, the properties *spdx:checksum*, *adms:status* were used by none of the portals. Furthermore, the property *foaf:page* occurred only in the Swedish data dump and in not more than on out of ten Distributions. The findings for other properties, which had more than 10% usage in any given portal, are summarized in figs. 21 to 23.

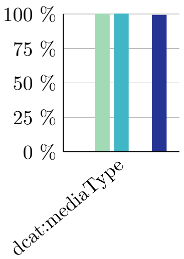

Figure 21: Mandatory properties in the Distribution class

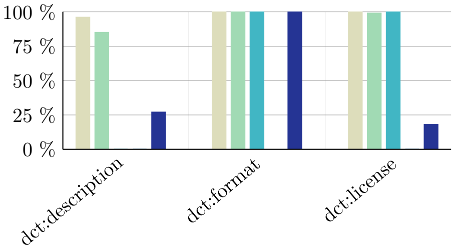

Figure 22: Recommended properties in the Distribution class

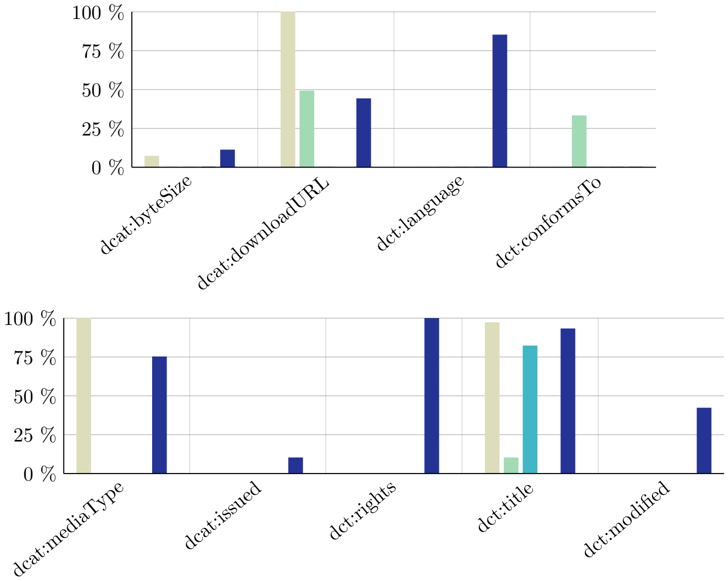

Figure 23: Optional properties in the Distribution class

4. Discussion

There are many small observations to be made from our analysis. They can be presented in many ways and the most logical order depends on the viewpoint of the reader. Hence, we chose to present them one DCAT class at a time. Then, we elaborate on issues which deal with different classes at the same time. To make it possible for the reader to choose their own path trough the observations, we annotate the subsection with the following tags:

[EDP] if the section discusses observations from the European Data Portal,

[NAT] if the section deals with observations from national portals, and

[VOC] if the section is specifically about the use of controlled vocabularies.

These tags can either be used to guide the reader trough the section or as a reference point to make sure which part of the analysis the subsection refers to.

4.1. dcat:Catalog

In this section we discuss several aspects about the dcat:Catalog class. We start by looking at several general observations and then look into issues related to the use of the controlled vocabularies.

4.1.1. **[EDP]** *General*

As shown in Figure 5, the mandatory properties for the *dcat:Catalog* class are respected in almost all cases. From the 3 catalogs with no dct:publisher attributed, 2 have in practice an organisation specified as publisher.

The recommended properties, dcat:themeTaxonomy and dct:modified were always present, meaning that all the catalogs queried specify at least one theme and a date. The EDP populates these two properties by default:

- for the theme taxonomy, by using the MDR data themes NAL or a mapping of the data themes harvested to the MDR data themes NAL[41], as recommended in *'How to use the MDR data themes vocabulary'*[42]; and
- for the modified date, as the latest harvested date of each catalog by the EDP.

Note that this means that even in case national portals do not specify these properties, which happens as we observed in the analysis reported in section 3.3, the EDP will present a value. Almost all the catalogs (98.75%) provide the first harvested date with the property *dct:issued*, and 23.75% of the catalogs include a homepage. An important observation concerning the use of the recommended properties *dct:license* and *dct:language* is that no catalog specified any of the two properties. The observation on the licence is not in contradiction with the usage note of the property *dct:license* on *dcat:Catalog.* The licence attribute on a catalog refers to the licence of the catalog itself, not of the individual datasets nor distributions.[43]

4.1.2. **[NAT]** *General*

We notice from the RDF dumps that the Norwegian and Spanish portal do publish their data with a Catalog which is annotated with all mandatory properties. The Swedish dump did not include a catalog at all, while the Swiss one only used it to provide links to the Datasets. The Irish one was similar to

[41] http://publications.europa.eu/mdr/authority/data-theme/index.html

[42] https://joinup.ec.europa.eu/release/dcat-ap-how-use-mdr-data-themes-vocabulary

[43] https://www.w3.org/TR/vocab-dcat/#class-catalog

the Swiss one, but also included a title. Also for optional properties, the Spanish portal gives the most complete information, followed by the Norwegian and Irish one. These issues seem like they are easy to solve for the maintainers of these portals as they usually only have one catalog and hence only on change would be needed to include the information. Possibly, this sloppiness is due to the mindset that anyone downloading their data dump knows where it comes from and is hence able to infer much of this information without it being specified. For example, one knows that all Datasets mentioned in the dump downloaded from the Swedish portal belong to their (implicit) Catalog.

4.1.3. **[NAT]** *Licensing*

The recommended licence property associated with Catalog, describes how the Catalog itself is licensed. Making license a mandatory property would make legal issues around reuse of Catalogs clearer, but this is only done in the Spanish application profile. That is also the only catalog which did specify a license in our further analysis. As this is a change at the Catalog level, providing it does not incur much of any overhead for the publisher, as the number of Catalog compared to the Dataset and Distribution are very small. Hence, we suggest making this a mandatory property.

4.1.4. **[EDP][VOC]** *adms:status on dcat:CatalogRecord*

The property *adms:status* for the class *dcat:CatalogRecord* refers to the type of the latest revision of a Dataset's entry in the Catalog. All the catalogs analysed use it and all the instances of this property, or 5.536, have the value *:modified* in place of one from the list provided by DCAT-AP: *:created, :updated, :deleted*. The ratio and the similar value among all the catalogs selected seem to demonstrate the use of this property by the EDP itself and not by the catalog owners.

4.1.5. **[EDP][VOC]** *dcat:themeTaxonomy on dcat:Catalog*

The property *dcat:themeTaxonomy* for the class *dcat:Catalog* follows the same Controlled Vocabulary than *dcat:theme* for the class *Dataset*. All catalogs in the sample do not appropriately use the property. All of them indicate the correct first part of the URI,
<*http://publications.europa.eu/resource/authority/data-theme*>, but not the full URI expected with the authoritative code, such as
<*http://publications.europa.eu/resource/ authority/data-theme/**AGRI***>. As described in Section 3.3.1, *dcat:themeTaxonomy* is populated by the EDP for all catalogs harvested.

4.2. dcat:Dataset

4.2.1. **[EDP][NAT]** *Use of dct:description*

dct:description is a mandatory property, yet it is not used on all datasets. Portal owners are encouraged to complete the use of descriptions to improve the understanding and the discoverability of the datasets, as titles for datasets are often not self-explanatory, which is the case for many of these 26,328 data sets on the EDP sample without description.

From the sample used in the national profiles, all did contain a title and most did in fact contain a description. In fact, only for the Swiss portal did we notice a significant (about 25%) amount of Datasets without description.

4.2.2. **[EDP]** *Use of dct:identifier and adms:identifier*

DCAT-AP foresees two optional properties for identifying a dataset, *dct:identifier* and *adms:identifier*. While the first one is described as the main unique identifier, e.g. a URI, the second property is provided as a secondary identifier. In practice, the second is used in 100% of the cases by the EDP itself while the first is in some cases provided by the catalog of datasets harvested. In most of the cases assessed, the *dct:identifier* follows the same format as the *adms:identifier*. This tends to confirm that the EDP provides a *dct:identifier* and a dataset URI for all the datasets which do not have one already.

4.2.3. **[EDP]** *Use of dct:spatial*

DCAT-AP mandates to use controlled vocabularies for the optional property *dct:spatial*. The list of controlled vocabularies is:

- MDR Continents Named Authority List[44] ;
- MDR Countries Named Authority List[45] ; and
- MDR Places Named Authority List[46] .
- Otherwise, Geonames[47] URIs.

The guideline *How should dct:spatial and dct:Location be used?*[48] reminds also that geographic coordinates can be used to represent a spatial region following the approach described in GeoDCAT-AP. We looked at the use of *dct:spatial*. As shown in table 4, the most used URIs are from Spain (i.e. datos.gob.es) with a local vocabulary, from Flanders (Belgium) with geonames, and from multiple countries with the MDR Countries Named Authority List. The correct use of *dct:spatial* following the controlled vocabularies is detailed in section 4.4.6.

Table 4: Use of dct:spatial for dcat:Dataset

spatial URI	**Count**
`http://datos.gob.es/recurso/sector-publico/territorio/Autonomia/Pais-Vasco`	4524

[44] http://publications.europa.eu/mdr/authority/continent/

[45] http://publications.europa.eu/mdr/authority/country/

[46] http://publications.europa.eu/mdr/authority/place/

[47] http://www.geonames.org/

[48] https://joinup.ec.europa.eu/release/how-should-dctspatial-and-dctlocation-be-used

http://sws.geonames.org/3337388	3824
http://datos.gob.es/recurso/sector-publico/territorio/Pais/Espa%C3%B1a	3272
http://sws.geonames.org/3337388/	3260
http://datos.gob.es/recurso/sector-publico/territorio/Autonomia/Aragon	3108
http://datos.gob.es/recurso/sector-publico/territorio/Provincia/Madrid	885
http://datos.gob.es/recurso/sector-publico/territorio/Provincia/Malaga	837
http://sws.geonames.org/2802361/	825
https://ruian.linked.opendata.cz/zdroj/st%C3%A1ty/1	653
http://datos.gob.es/recurso/sector-publico/territorio/Autonomia/Galicia	541
http://datos.gob.es/recurso/sector-publico/territorio/Autonomia/Castilla-Leon	526
http://sws.geonames.org/2797656/	422
http://publications.europa.eu/resource/authority/country/FRA	420
http://publications.europa.eu/resource/authority/country/DNK	415
http://publications.europa.eu/resource/authority/country/EST	414
http://publications.europa.eu/resource/authority/country/SVN	410
http://publications.europa.eu/resource/authority/country/LVA	409

A similar query was run for the class *dcat:Catalog*. The optional *dct:spatial* property is used with 33 countries as shown in Table 5.

Table 5: Use of dct:spatial for the class dcat:Catalog

spatial	**Count**	**spatial**	**Count**	**spatial**	
IT	6	LV	2	MT	1
ES	5	NL	2	LI	1
CZ	4	FI	2	RS	1
DE	3	SI	2	CY	1
AT	3	PT	2	IE	1
HR	3	SE	2	CH	1
PL	3	SK	2	HU	1
EE	2	NO	2	LT	1
IS	2	FR	2	BG	1
RO	2	GB	2	MD	1

LU	2	BE	2	DK	1

4.2.4. **[NAT]** *Geospatial Properties*

Additional geospatial metadata elements are found in few of the national extensions. In the Irish AP, there are three additional spatial attributes associated with the class *Dataset*. These additional attributes are *GeographicBoundingBox*, *SpatialReference System*, and *Spatial Resolution*. Similarly *politicalGeocodingLevelURI*, *politicalGeocodingURI*, and *geocodingText* are added to the German profile. Geometry, *geographicalIdentifier*, and *geographicalName* are added the Italian application profile. It has further added some restrictions on the loc:geometry attribute, by associating *CRS*, *Coordinates*, and geometry type as mandatory properties.

The property dct:spatial in DCAT-AP1.1 is used to define the geographic coordinates system of a Catalog. Norway, Spain, Sweden, Switzerland, and the Netherlands have introduced spatial as an attribute for the class *Dataset*. Sweden has also made it more restricted for class *Catalog*. There are many updates in national extension for this property, which makes it a suitable candidate for the new versions of DCAT and DCAT-AP1.1.

4.2.5. **[EDP]** *Use of dct:accrualPeriodicity*

The optional property *dct:accrualPeriodicity* indicates the frequency at which a dataset updated by its owner. From the different periodicity with their use, the most used frequencies are visualised in Figure 24. One important point to notice concerns the quantity of duplicated frequencies (e.g., IRREG and IRREGULAR). For example, some datasets are described using the authority code (i.e., IRREG) provided by the frequency authority list from the Publications Office[49] while other datasets use the label from the same authority list (i.e., IRREGULAR).

[49] http://publications.europa.eu/mdr/resource/authority/frequency/html/frequencies-eng.html

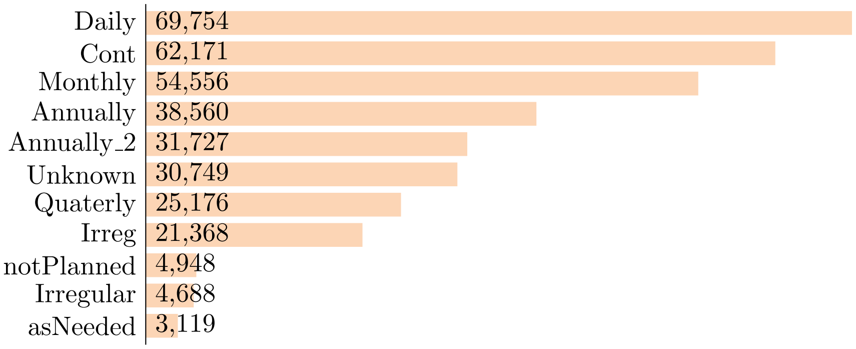

Figure 24: Use of the property dct:accrualPeriodicity

4.2.6. **[EDP]** *Use of dct:provenance*

The optional *dct:provenance* is supposed to be used for providing a statement about the lineage of a dataset. In a guideline on how to model and express provenance[50], the following recommendation was made: *As the provision of provenance information is not wide-spread and information in free text does not allow further processing, the usefulness of such information in (international) harvesting is questionable and the information may be ignored. Local implementations are of course free to provide provenance information satisfying local requirements.* However, the guideline recognises the potential of such information: *It could support credibility of a dataset to know which organisation created the metadata for it in the first place and how the description was modified along a chain of exchanges.*

In practice, different uses of provenance as described above were observed, which demonstrates that the use of this property is not standardised. For example, a dataset[51] describes the provenance by mentioning the context and the organisation responsible for the creation of the dataset (translation from Dutch): *Label: This map was established on the basis of an inventory during the establishment of the provincial policy for recreational and professional use of waterways.* Moreover, the provenance is also used to keep track of the modifications applied to a dataset. One example is `https://www.europeandataportal.eu/data/dataset/de-pangaea-dataset864016` with: *Label: The data set was checked for completeness, correctness, and consistency of metainformation. Validity of used methods was checked and - if applicable - precision and range of data.* Consequently, we would recommend that the owners and users of DCAT-AP further implement the existing guideline, as it is still relevant.

[50] https://joinup.ec.europa.eu/release/dcat-ap-how-model-and-express-provenance

[51] https://www.europeandataportal.eu/data/dataset/00dfaddf-f2f0-487a-b28e-aad53a318521

4.2.7. **[EDP][VOC]** *dct:language on dcat:Dataset*

The property *dct:language* for the class *dcat:Dataset* is used to indicate a language for a dataset. DCAT-AP requests implementers to provide a value from the MDR Languages Named Authority List[52] such as *<http://publications.europa.eu/resource/ authority/language/ENG>* for English. In our sample, only one catalog uses *dct:language* for the class Dataset pointing correctly to the MDR Languages Named Authority List. The other catalogs are not specifying the language for the class *dcat:Dataset*. See also some further observations on language in section 4.4.4.

4.2.8. **[EDP][VOC]** *dct:accrualPeriodicity on dcat:Dataset*

Similarly, the property *dct:accrualPeriodicity* is used to refer to the frequency at which the Dataset is updated. DCAT-AP requests to use the MDR Frequency Named Authority List[53] , for example with a URI value as follows: *<http://publications.europa.eu/resource/authority/frequency/ANNUAL>*. Among the 6 catalogs analysed, a single one provided information about the frequency of update, providing the correct value from the MDR Frequency Named Authority List.

4.2.9. **[EDP][VOC]** *dcat:theme on dcat:Dataset*

The property *dcat:theme* refers to a category of the Dataset. A Dataset may be associated with multiple themes. As for the previous properties, a Controlled Vocabulary must be followed, the MDR data theme Named Authority List[54] with a URI pointing to one authority code from the list. From our sample analysis, the Controlled Vocabulary is perfectly used by all 6 catalogs.

4.3. dcat:Distribution

In this subsection we discuss various aspects related to the Distribution class.

4.3.1. **[EDP]** *Relationship Between accessURL, downloadURL and Distributions*

As mentioned in section 3.1.3, almost all distributions respect the mandatory property *dcat:accessURL*. There was no case for which a distribution lacked both the *accessURL* and *downloadURL*.
We also looked at the use of the two properties combined, to confirm if this use is differentiated or not. In practice, 16.866 distributions queried have different URLs for access and download while 253.006 distributions[55] have the same URL. This shows that the high majority of the distributions using the two

[52] http://publications.europa.eu/mdr/authority/language/

[53] http://publications.europa.eu/mdr/authority/frequency

[54] http://publications.europa.eu/mdr/resource/authority/data-theme/html/data-theme-eng.html

[55] The number of distributions using downloadURL is slightly different than in the percentage reported in Section 3.1 (optional property dcat:downloadURL) due of the time gap between queries.

properties do not provide different information for both properties but simply copy twice a downloadURL, as explained in the guideline *How to use accessURL and downloadURL?'*[56].

4.3.2. **[EDP]** *Use of dct:licence*

The low amount of licences defined at the level of distributions, where the use of *dct:license* is recommended, can be a barrier for the reuse of the Open Data, as many potential users might not take the risk of using distributions without knowing under which conditions they can do it. The EDP also adds that 90% of the licences of the datasets on the portal are unknown[57]. The portal considers a licence as unknown if it is not part of the list of licences provided by CKAN[58]. Using known licences for the datasets would greatly simplify the work required for potential users before deciding if they can use specific datasets or not.

In general, we also found that the distributions providing a known licence are compliant with the guideline on *How to refer to licence documents and licence URIs?*[59]. The guideline specifies that licences should always be identified with URIs which should resolve to the description of the licence.

4.3.3. **[NAT]** *Licence*

Licence and copyrights information has been updated in many national application profiles. The German extension has suggested the addition of attribution text for licence, till then it will use dcatde:licenseAttributionByText. The attribute dct:license for class *Distribution* has been made mandatory in the German, Italian, and Swiss application profiles. The Dutch extension consider it unnecessary for class *Distribution*, therefore they removed it from this class. But added it to the class *Dataset*, but a small set of possible values. The reason behind this change is the fact that multiple distribution of the same Dataset share the same licence. A Distribution with a different licence, will be handled as a different Dataset by the Dutch application profile as detailed in a Joinup issue[60]. Although the Dutch profile can support distributions with different licences this way, the approach is not consistent with other national application profiles. Therefore, any other catalog harvesting data from Dutch profiles needs special work around to capture this information. This change is clearly nonconformant.

4.3.4. **[NAT]** *dcat:mediaType and dct:format*

dcat:mediatype is defined as an optional sub property of the recommended property dct:format for class *Distribution*, both in DCAT and DCAT-AP1.1.

[56] https://joinup.ec.europa.eu/release/how-use-accessurl-and-downloadurl

[57] https://www.europeandataportal.eu/mqa-service/en

[58] https://www.europeandataportal.eu/en/licence-assistant

[59] https://joinup.ec.europa.eu/release/dcat-ap-how-refer-licence-documents-and-licence-uris

[60] `https://joinup.ec.europa.eu/asset/ogd2_0/issue/granularity-conflicts-license-and-status`

This property helps in choosing the type of software that could be used to process data. The Norwegian application profile has excluded dcat:mediaType considering dct:format sufficient to capture the type as well. Similarly, the Swedish DCAT-AP extension has discouraged the use of dcat:mediaType. The Swiss application profile has, conversely, made dcat:mediaType a recommended property of Distribution. Concluding, we believe the changes made by local implementers regarding the use of dcat:mediaType and dct:format are contradictory. Looking at the different types of update made to mediaType property, it seems more appropriate to keep dcat:mediaType optional, as defined in DCAT-AP v1.1.

4.3.5. **[EDP]** *Relationship Between dct:title, dct:format and dcat:mediaType*

DCAT-AP advises to use:

- *dct:format* (recommended) to give information about the file format of the distribution;
- *dcat:mediaType* (optional), as a subproperty of dct:format, to follow the official register of media types managed by IANA[61]; and
- *dct:title* (optional at the level of distribution, mandatory at the level of dataset) to give a name to the distribution.

However, our analysis proved that many distributions do the opposite: *dct:title* is in many cases used to inform about the format of the file instead of *dct:format*. Despite this misuse, when looking at the combined use of *dct:title* and *dcat:mediaType*, most distributions use appropriately the two properties: *dct:title* as a name for the distribution and *dcat:mediaType* with a value from IANA. Some members of the DCAT-AP community have expressed a preference to use only *dct:format* with IANA media type. One reason for using IANA media types, is that you can express the *innerMimeType* for ZIP-Files. On the other hand, *dct:format* is more flexible. Even though IANA does not include all geospatial values, they can be added to the MDR list (*dct:format*). Sections 4.3.6 and 4.3.7 go deeper in the analysis of the controlled vocabularies for *dct:format* and *dcat:mediaType*.

4.3.6. **[EDP][VOC]** *dct:format on dcat:Distribution*

The *dct:format* property must refer to the MDR File Type Named Authority List[62] for describing the file format of a distribution. All 6 catalogs have datasets with the property *dct:format*. For all instances identified using *dct:format*, the MDR controlled vocabulary is not followed. Instead, the value gives a URI pointing to an instance of *dct:MediaTypeOrExtent* with the format described as a literal in *rdf:label*, e.g. *WMS*. *WMS* means Web Map Service, a standard protocol for geospatial data . The fact that such web services are not included in the MDR File Type NAL partly explains why the code list is not used.

[61] https://www.iana.org/assignments/media-types/media-types.xhtml

[62] http://publications.europa.eu/mdr/authority/file-type/

4.3.7. **[EDP][VOC]** *dcat:mediaType on dcat:Distribution*

dcat:mediaType is a sub-property of *dct:format* used to express the media type of the Distribution as defined in the official register of media types managed by IANA[63]. Only one catalog out of 6 uses *dcat:mediaType.* The values provided for the sub-property are almost always correct, with a percentage error of 4% (55/1365). For the errors identified, a media type not referenced in IANA was used, such as xml/soap.

4.4. Observations Involving Multiple Classes

Some properties and observations involve more than one DCAT class. These are either properties which exist on more than one class, or those which discuss a relation between two classes. These are discussed in this subsection.

4.4.1. **[EDP]** *Relationship Between dct:publisher and dcat:contactPoint*

In the guideline *How are publisher and contact point modelled?*[64], the following recommendation is made: *The way that DCAT and DCAT-AP distinguish between the publisher (the organisation that makes the catalog or dataset available) and contact information (address where more information can be requested, or feedback can be given) is a continuing source of confusion. It is important to differentiate and provide the two types of information: Publisher is necessary to identify the entity and Contact point allows any person/organisation to communicate and provide feedback.*

As visualised in Figure 9, the two recommended properties *dct:publisher* and *dcat:contactPoint* have approximately the same frequency of use. When analysing more in details if the use of the two properties is aligned with the guideline, we found that among the 38.672 datasets which have both properties defined: 34% (13.133) have a different publisher than the contact point and 66% (25.539) duplicate the information from *dct:publisher* to *dcat:contactPoint*, as specified in the guideline.

4.4.2. **[EDP]** *Relationship Between Dataset and Distribution*

Since a dataset can have several distributions, for example a distribution in CSV format and one in XML, it was expected that the number of distributions (932.771) is higher than the number of datasets (743.000). The analysis however shows that not all datasets have a distribution. This can be explained by the fact that there are other ways to access the data, for example via the optional properties *dcat:landingPage* (38.42%) or *foaf:page* (10.31%), which is often used for geospatial datasets that are accessed through web APIs, such as map services. We also verified that data publishers were not using multiple properties (distribution, landing page and/or page) for the same dataset. The only overlapping use is really minor as it only concerns 17 datasets having at the same time a *dct:landingPage* and a *foaf:page.*

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

[64] https://joinup.ec.europa.eu/release/how-are-publisher-and-contact-point-modelled

4.4.3. **[NAT]** *Relationships between Catalogs, Datasets, and Distributions*

DCAT-AP1.1 has few relationship attributes for class *Catalog* and *Dataset*. Relationship properties are used to represent association between catalogs. It could be used to represent inheritance or containment among different *Catalog* instances. If a *Catalog* is a physical or logical part of another *Catalog* that can be indicated using *dct:isPartOf*. A Catalog containing another Catalog as a physical or logical part could be represented using the inverse property *dct:hasPart*. Similarly *dct:isVersionOf* is used to describe adaptation, version of or edition association among two datasets. More properties could be added to show different possible association among *Dataset*, *Catalog* and *Distribution*. The analysis shows, relationship properties are mostly added in the application profiles of Norway and Spain.

However, in the further analysis we could see that these properties were hardly ever used. We could, over all 5 portals for which the actual property use was analyzed, only find one pair of Datasets that specified the dct:isVersionOf and dct:hasVersion properties. We could find more usage of the property dct:relation, but that property does not specify the precise relation nor the type of the other resource.

4.4.4. **[NAT]** *Language*

Some national extensions provide or even require multi language support. In particular, the Swiss extension has added xml:lang attribute, to support multilingual elements. The example shown in 13 shows the use aforementioned attribute. This example is taken from the Swiss National DCAT-AP website.

```
<dct:title xml:lang="fr">FR Titre</dct:title>
<dct:title xml:lang="de">DE Titel</dct:title>
<dct:title xml:lang="it">IT Titolo</dct:title>
<dct:title xml:lang="en">EN Title</dct:title>
```

The Spanish DCAT-AP is using dc:language instead of dct:language. The range of the language attribute is rdfs:Literal, unlike dct:LingusiticSystem in DCAT-AP. The Spanish profile is older than DCAT-AP, and it might be the reason of this inconsistency.

4.4.5. **[EDP][VOC]** *dct:publisher on dcat:Dataset and dcat:Catalog*

The values of dct:publisher can not automatically be checked by the validator since the rules for this property depend on the nature of the publisher. The property *dct:publisher* must follow the MDR Corporate bodies Named Authority List[65] for which DCAT-AP specifies that it *must be used for European institutions and a small set of international organisations. In case of other*

[65] http://publications.europa.eu/mdr/authority/corporate-body/

types of organisations, national, regional or local vocabularies should be used[66]. As data portals can indicate national or sub-national publishers, this property could not be verified automatically with the DCAT-AP validator. However, from manual processing, the following conclusions were found:

- 5 of 6 catalogs give information about *dct:publisher* for the class *Dataset*;
- None of the 5 catalogs is using the MDR Corporate bodies Named Authority List. Instead, for 4 catalogs, a URI pointing to the properties *foaf:name*, and *rdf:type* is given. For the URIs of one of those 4 catalogs, the property *foaf:mbox* (i.e. mailbox) is also provided. The fifth catalog gives the URI of a separate code list maintained independently from the MDR Corporate bodies.
- For the class *dcat:Catalog*, all catalogs analysed also provide information about *dct:publisher* using a URI pointing to *foaf:name* and *rdf:type*, therefore not using a Controlled Vocabulary as specified by DCAT-AP.

4.4.6. **[EDP][VOC]** *dct:spatial on dcat:Dataset and dcat:Catalog*

For *dct:spatial* under the class *dcat:Dataset*, multiple Controlled Vocabularies are requested by DCAT-AP, respectively:

- MDR Continents Named Authority List[67] for continents;
- MDR Countries Named Authority List[68] for countries;
- MDR Places Named Authority List[69] for places; and
- Geonames URIs if a particular location is not in one of the mentioned Named Authority Lists.

Among the 4 catalogs expressing *dct:spatial*, none comply with the code lists of DCAT-AP. Two types of inconsistencies were identified:

- 1.719 instances out of 3.002 contain a URI pointing to a polygon (*locn:geometry*); and
- For 1.283 instances out of 3.002, *dct:spatial* points to a code list of territories specific to the country of the catalog and not following the Controlled Vocabularies provided by DCAT-AP.

The same use is expected from the catalogs for the class *dcat:Catalog*, however, the 6 catalogs analysed use the ISO code (e.g. *IS* for Iceland) of the country. Since the countries in our sample are listed in the MDR Countries Named Authority List, the NAL URIs should have been used instead.

[66] https://joinup.ec.europa.eu/release/dcat-ap-v11

[67] Publications Office of the European Union. Metadata Registry. Authorities. Continents.

[68] Publications Office of the European Union. Metadata Registry. Authorities. Countries.

[69] Publications Office of the European Union. Metadata Registry. Authorities. Places.

4.4.7. **[NAT]** *Non-Conformant Changes in Specifications*

DCAT-AP has defined a clear set of rules for extension. According to the extension rules updates that generalize an existing property or a class might cause inconsistencies and therefore should be avoided. Similarly, changing a mandatory property to an optional one, by relaxing cardinality constraints might result in data integration problem, therefore should be avoided. Any update that is not in alliance with extension rule is identified as a non-conformant update in this analysis. Three cases of such updates are presented below:

- DCAT-AP-NL: The application profile has added a new property version, while there is already a property with same name, but a different property URI. Introducing similar properties and classes is discouraged in DCAT-AP extension rules.
- CH-DCAT-AP: This national profile has introduced an attribute dcat:coverage, similar to the existing spatial property dct:coverage in DCAT. It seems more like an error, which is preserved for backward compatibility, rather than an intentional update, which seems supported by the suggestion to avoid the usage of this property altogether.
- DCAT-AP-Spain: The national application profile of Spain is older than DCAT-AP1.1, therefore there are some inconsistencies among the two. In particular the difference in adding support for multiple language might cause interoperability issues with other application profiles.

5. Future Work and Conclusion

In this paper, the analysis of national profile specifications implementing DCAT-AP v1.1 has been presented together with the actual use of the specification both from several national Open Data portals and the European Data Portal. Based on this work, we could see that much effort has been done to create interoperable open data platforms, but that more effort is still needed. In particular, we found that the practical use of mandatory properties is not always as it should be. Moreover, when the range of a property is fixed to a specific vocabulary, this is not always respected. While this could indicate that there the properties as defined does not reflect actual needs for these portals, incorrect use will hamper reuse and hence interoperability.

Starting from this work, several further investigations are possible.

A related investigation would look into differences between the use of standards on the same platform. This would highlight cases where information is available in a platform in one format, but incompletely represented in another format (either due to technical limitations or missing implementation). This would then in turn lead to recommendations towards standardization efforts and perhaps illustrate elementary differences between the standards. Moreover, this investigation could indicate reasons why there could be differences between what exists at the national level and what becomes available at the European level.

One aspect noticed while investigating the changes made is that they often do not really reflect issues which one might associate with changes needed to serve the needs of a nation or country better. Expected aspects would be, for example, the addition of nationally used identifiers, coordinates in a nationally used coordinate system, or making the translation of labels into national language(s) mandatory. However, the changes made are usually not of that kind. We do suspect that some of these changes have been made due to compatibility with, or limitation within, preexisting national systems. We do believe that finding out the precise reasons for these changes is worth further investigation.

Through our analysis, we have indicated several properties which could be discussed for inclusion in the next iteration of DCAT-AP or the W3C DCAT recommendation. New properties to be considered for future revisions of DCAT-AP include those related to spatial properties and relationships between the class *Dataset* and *Distribution*.

We also found examples of already existing properties which have been modified frequently including, *dct:identifier*, *dct:publisher*, *dcat:theme*, and the way to use the *vCard* class. Moreover, we identified a need to standardise, or rather document more clearly, how *license* and *mediaTypes formats* are to be used.

It appears to us that especially on the Catalog level it would be useful, and rather easy for implementers, to have some more mandatory properties. Especially licensing of the Catalog is essential for reuse.

We also found several changes made by national profiles which limit interoperability or which only help implementations capable of dealing with these specific requirements, while other implementations ignore the information as they are unable to interpret it.

In the future, the ISA^2 Programme could help DCAT-AP implementers overcome these interoperability challenges by, for example, creating additional guidelines that ensure the compatibility of extensions with DCAT-AP and the interoperability of extensions among each other, or by checking the compliance of national extensions with DCAT-AP.

Last, a possible association would be to show succession. An example would be the yearly water consumption report of a country. Representing it using *dct:isVersionOf* deludes the real semantics. Because a new yearly report is just sequence of the dataset. It is not really a version of the last years report, nor does it replace the previous one. This type of temporal sequencing doesnt exist in DCAT-AP1.1, and we didnt find it in any of the national extensions.

Establishing a standard format for the representation of national DCAT-AP profiles would make the identification and documentation of changes easier. Currently, some countries publish their specification in a pdf file, others in an html web page, both of which do not have an agreed upon structure. Some specifications explicitly document updates and additions, while others maintain one huge list of properties, and finding out what has changed is left as a tedious job to the reader.

6. Acknowledgments

We would like to thank Makx Dekkers for reviewing the section where the European Data Portal analysis is presented.

7. Bibliography

[1] F. Ahmadi Zeleti, A. Ojo, E. Curry, Exploring the economic value of open government data, Government Information Quarterly 33 (3) (2016) 535–551. doi:10.1016/j.giq.2016.01.008.

[2] A. Zuiderwijk, M. Janssen, G. van de Kaa, K. Poulis, The wicked problem of commercial value creation in open data ecosystems: Policy guidelines for governments, Information Polity (2016) 223–236doi:10.3233/IP-160391.

[3] E. Ruijer, S. Grimmelikhuijsen, A. Meijer, Open data for democracy: Developing a theoretical framework for open data use, Government Information Quarterly (2017) 45 – 52doi:https://doi.org/10.1016/j.giq.2017.01.001.

[4] E. Ruijer, E. Martinius, Researching the democratic impact of open government data: A systematic literature review, Information Polity (2017) 233–250doi:10.3233/IP-170413.

[5] J. Attard, F. Orlandi, S. Auer, Data driven governments: Creating value through open government data, in: Transactions on Large-Scale Data- and Knowledge-Centered Systems XXVII, Springer-Verlag New York, Inc., 2016, pp. 84–110. doi:10.1007/978-3-662-53416-8_6.

[6] A. Zuiderwijk, M. Janssen, Open data policies, their implementation and impact: A framework for comparison, Government Information Quarterly (2014) 17 – 29doi:https://doi.org/10.1016/j.giq.2013.04.003.

[7] K. Janssen, The influence of the psi directive on open government data: An overview of recent developments, Government Information Quarterly 28 (4) (2011) 446 – 456. doi:https://doi.org/10.1016/j.giq.2011.01.004. URL http://www.sciencedirect.com/science/article/pii/S0740624X11000517

[8] K. Janssen, S. Hugelier, Open data as the standard for europe? a critical analysis of the european commissions proposal to amend the psi directive, European Journal of Law and Technology.

[9] L. Ding, V. Peristeras, M. Hausenblas, Linked open government data, IEEE Intelligent Systems (2012) 11–15doi:10.1109/MIS.2012.56.

[10] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. Zheng, Z. Shangguan, J. Flores, D. L. McGuinness, J. A. Hendler, Twc logd: A portal for linked open government data

ecosystems, J. Web Sem. (2011) 325–333doi:https://doi.org/10.1016/j.websem.2011.06.002.

[11] S. van der Waal, K. Wecel, I. Ermilov, V. Janev, U. Milosevic, M. Wainwright, Lifting open data portals to the data web, in: S. Auer, V. Bryl, S. Tramp (Eds.), Linked Open Data - Creating Knowledge Out of Interlinked Data - Results of the LOD2 Project, Vol. 8661 of Lecture Notes in Computer Science, Springer, 2014, pp. 175–195. doi:10.1007/978-3-319-09846-3_9.
URL https://doi.org/10.1007/978-3-319-09846-3_9

[12] J. Attard, F. Orlandi, S. Scerri, S. Auer, A systematic review of open government data initiatives, Government Information Quarterly (2015) 399 – 418doi:https://doi.org/10.1016/j.giq.2015.07.006.

[13] A. Zuiderwijk, M. Janssen, The negative effects of open government data - investigating the dark side of open data, in: DG.O, 2014, pp. 147 – 152. doi:https://doi.org/10.1145/2612733.2612761.

[14] S. Neumaier, J. Umbrich, A. Polleres, Automated quality assessment of metadata across open data portals, J. Data and Information Quality (2016) 2:1–2:29doi:10.1145/2964909.

[15] M. Janssen, E. Estevez, T. Janowski, Interoperability in big, open, and linked data–organizational maturity, capabilities, and data portfolios, Computer (2014) 44–49doi:10.1109/MC.2014.290.

[16] F. Maali, R. Cyganiak, V. Peristeras, Enabling Interoperability of Government Data Catalogues, in: M. A. W. J.-L. C. M. J. H. J. Scholl (Ed.), 9th IFIP WG 8.5 International Conference on Electronic Government (EGOV), Electronic Government, Springer, Lausanne, Switzerland, 2010, pp. 339–350. doi:10.1007/978-3-642-14799-9_29.

[17] R. Cyganiak, F. Maali, V. Peristeras, Self-service linked government data with dcat and gridworks, in: Proceedings of the 6th International Conference on Semantic Systems, I-SEMANTICS '10, ACM, 2010, pp. 37:1–37:3. doi:10.1145/1839707.1839754.

[18] S. Neumaier, J. Umbrich, A. Polleres, Lifting data portals to the web of data, in: 10th Workshop on Linked Data on the Web (LDOW2017), Perth, Austrialia, 2017, pp. 1–10.

[19] M. Pellegrino, Representing statistical metadata by using the dcat application profile for data portals in europe, in: Workshop on Integrating Geospatial and Statistical Standards 2017, 2017, pp. 339–350.

[20] M. Dekkers, S. Kotoglou, C. Nelson, M. Pellegrino, N. Hohn, V. Peristeras, Statdcat-ap, a common layer for the exchange of statistical metadata in open data portals, in: 4th International Workshop on Semantic Statistics,

Co-Located with the 15th International Semantic Web Conference, 2016, pp. 1–12.

[21] W. Carrara, M. Dekkers, B. Dittwald, S. Dutkowski, Y. Glikman, N. Loutas, V. Peristeras, B. Wyns, Towards an open government data ecosystem in europe using common standards [www document], Eur.Comm.

[22] J. Klmek, J. Kuera, M. Neask, D. Chlapek, Publication and usage of official czech pension statistics linked open data, Journal of Web Semantics (2018) 1 – 21doi:https://doi.org/10.1016/j.websem.2017.09.002.

[23] J. Winn, Open data and the academy: An evaluation of ckan for research data management, in: IASSIST 2013, 2013, pp. 1–21.

[24] R. V. Guha, D. Brickley, S. Macbeth, Schema.org: Evolution of structured data on the web, Commun. ACM 59 (2) (2016) 44–51. doi:10.1145/2844544.
URL http://doi.acm.org/10.1145/2844544

[25] W. Kresse, K. Fadaie, ISO Standards for Geographic Information, Springer Science+Business Media, 2004.

[26] J. Nogueras-Iso, J. Barrera, A. F Rodrguez, R. Recio, C. Laborda, F. Zarazaga, Development and deployment of a services catalog in compliance with the inspire metadata implementing rules, SDI Convergence: Research, Emerging Trends, and Critical Assessment.The Netherlands Geodetic Commission (NGC) (2009) 21–34.

[27] S. Capadisli, S. Auer, A. N. Ngomo, Linked SDMX data: Path to high fidelity statistical linked data, Semantic Web 6 (2) (2015) 105–112. doi:10.3233/SW-130123.
URL https://doi.org/10.3233/SW-130123

[28] K. Jeffery, N. Houssos, B. Jrg, A. Asserson, Research information management: the cerif approach, International Journal of Metadata, Semantics and Ontologies 9 (1) (2014) 5–14. arXiv:https://www.inderscienceonline.com/doi/pdf/10.1504/IJMSO.2014.059142, doi:10.1504/IJMSO.2014.059142.
URL https://www.inderscienceonline.com/doi/abs/10.1504/IJMSO.2014.059142

[29] K. Alexander, M. Hausenblas, Describing linked datasets - on the design and usage of void, the vocabulary of interlinked datasets, in: In Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW 09, 2009.

[30] S.-A. Sansone, A. Gonzalez-Beltran, P. Rocca-Serra, G. Alter, J. Grethe, H. Xu, I. Fore, J. Lyle, A. Gururaj, X. Chen, H.-e. Kim, N. Zong, Y. Li,

R. Liu, B. Ozyurt, L. Ohno-Machado, bioCADDIE Working Groups, Dats: the data tag suite to enable discoverability of datasets, Scientific Data.

[31] M. Freudenberg, M. Brümmer, J. Rücknagel, R. Ulrich, T. Eckart, D. Kontokostas, S. Hellmann, The metadata ecosystem of dataid, in: MTSR, 2016, pp. 1–15.

[32] S. Neumaier, J. Umbrich, A. Polleres, Challenges of mapping current CKAN metadata to DCAT, in: W3C Workshop on Data and Services Integration, Amsterdam, the Netherlands, 2016, pp. 1–5.
URL https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_16

[33] C. Bizer, K. Eckert, R. Meusel, H. Mühleisen, M. Schuhmacher, J. Völker, Deployment of rdfa, microdata, and microformats on the web &#151; a quantitative analysis, in: Proceedings of the 12th International Semantic Web Conference - Part II, ISWC '13, Springer-Verlag New York, Inc., New York, NY, USA, 2013, pp. 17–32. doi:10.1007/978-3-642-41338-4_2.
URL http://dx.doi.org/10.1007/978-3-642-41338-4_2

[34] B. T. Balci, U. Simsek, E. Kärle, D. Fensel, Analysis of schema.org usage in the tourism domain, CoRR abs/1802.05948. arXiv:1802.05948.
URL http://arxiv.org/abs/1802.05948

[35] R. B. Johnson, A. J. Onwuegbuzie, L. A. Turner, Toward a definition of mixed methods research, Journal of Mixed Methods Research 1 (2) (2007) 112–133. arXiv:https://doi.org/10.1177/1558689806298224, doi:10.1177/1558689806298224.
URL https://doi.org/10.1177/1558689806298224

[36] V. Cox, Exploratory data analysis, in: Translating Statistics to Make Decisions : A Guide for the Non-Statistician, Apress, Berkeley, CA, 2017, pp. 47–74. doi:10.1007/978-1-4842-2256-0_3.
URL https://doi.org/10.1007/978-1-4842-2256-0_3

[37] S. J. Barnes, R. T. Vidgen, Data triangulation and web quality metrics: A case study in e-government, Inf. Manage. 43 (6) (2006) 767–777. doi:10.1016/j.im.2006.06.001.
URL http://dx.doi.org/10.1016/j.im.2006.06.001

Appendix A. Detailed of Changes in National Profiles

The following tables summarize the changes made in national profiles. In table A.6 we introduce the type of property (i.e. mandatory [M], recommended [R], optional [O]), the name of the property and the corresponding class, the number of changes and the specific change. In table A.7 we list the properties which were excluded from the given profiles.

Table A.6: Detail of property changes

Type	Property	Class	No. of Changes	Specific Change
M	dct:identifier	Dataset	4	From Optional to Mandatory
M	dct:identifier	Agent	2	From Optional to Mandatory
M	dct:identifier	Standard	1	From Optional to Mandatory
M	dct:publisher	Dataset	4	From Recommended to Mandatory
M	dct:license	Distribution	3	From Recommended to Mandatory
M	dct:license	Distribution	1	The extension excluded this property and added a new licence attribute with a limited number of possible values
M	dcat:theme	Dataset	4	From Optional to Mandatory
M	vcard:fn	Organization	2	
M	vcard:fn	Contact	1	
M	vCard:hasEmail	Organization	2	Restricted the allowed values for vCard properties. One change was restricting it to organization and the other also allows the use of vCards describing individuals
M	vCard:hasEmail	Contact	1	Restricted the allowed values for vCard properties. One change was restricting it to organization and the other also allows the use of vCards describing individuals
M	dct:modified	Dataset	1	
M	dct:modified	Catalog	1	
M	dct:issued	Catalog	2	
M	dcat:mediaType	Distribution	2	
M	rdf:type		2	
M	dct:format		2	
M	dcat:accessURL	Distribution	2	
R	dct:license	Dataset	2	A recommended property was changed to an optional and vice versa
R	dct:license	Distribution	1	A recommended property was changed to an optional and vice versa
R	dct:license	Catalog	1	A recommended property was changed to an optional and vice versa

R	dct:subject	Dataset	3	A recommended property was changed to an optional and vice versa
O	dct:spatial	Dataset	3	The changes vary between the different national extensions
O	dct:spatial	Catalog	1	The changes vary between the different national extensions
O	dct:identifier	Agent	1	A recommended property was changed to an optional and vice versa
O	dct:identifier	Catalog	1	A recommended property was changed to an optional and vice versa
O	vCard:hasTelephone	Organization	2	A recommended property was changed to an optional and vice versa
O	vCard:hasTelephone	Contact	1	A recommended property was changed to an optional and vice versa
O	dct:creator	Dataset	2	A recommended property was changed to an optional and vice versa
O	dct:creator	Distribution	3	A recommended property was changed to an optional and vice versa
O	dct:description	Standard	1	A recommended property was changed to an optional and vice versa

Table A.7: Detail of excluded properties

Extension	**Class**	**Property**
DCAT-AP_IT	Distribution	spdx:checksum
DCAT-AP_IT	Distribution	foaf:page
DCAT-AP_IT	Distribution	dct:language
DCAT-AP_IT	Distribution	dct:conformsTo
DCAT-AP_IT	Distribution	dct:rights
DCAT-AP_IT	Distribution	adms:status
DCAT-AP_IT	Distribution	dct:issued
DCAT-AP_IT	Distribution	dcat:mediaType
DCAT-AP_IT	Dataset	dct:relation
DCAT-AP_IT	Dataset	dct:source
DCAT-AP_IT	Dataset	dct:accessRights

DCAT-AP_IT	Dataset	dct:provenance
DCAT-AP_IT	Dataset	foaf:page
DCAT-AP_IT	Dataset	dct:hasVersion
DCAT-AP_IT	Dataset	adms:sample
DCAT-AP_IT	Dataset	dct:type
DCAT-AP_IT	Dataset	adms:versionNotes
DCAT-AP_IT	Catalog	hasPart
DCAT-AP_IT	Catalog	isPartOf
DCAT-AP_IT	Catalog	dcat:record
DCAT-AP_IT	Catalog	dct:spatial
DCAT-AP_IT	Catalog	dct:license
DCAT-AP_IT	Catalog	dct:rights
DCAT-AP_IT	Agent	dct:type
DCAT-AP-NL	Distribution	license
DCAT-AP-NL	Distribution	conformsTo
DCAT-AP-NL	Distribution	language
DCAT-AP-NL	Distribution	page
DCAT-AP-NL	Distribution	checksum
DCAT-AP-NL	Distribution	type
DCAT-AP-NL	Distribution	rights
DCAT-AP-NO	Distribution	mediaType