European Public Sector Information Platform

Topic Report No. 2013 / 09

Impact of Standards in European Open Data Catalogues. A Multilingual perspective of DCAT

Authors: Elena Montiel-Ponsoda, Boris Villazón-Terrazas

Published: September 2013

Table of Contents

Keywords:

Data catalogs, DCAT, Public Sector Information, Multilinguality

Abstract/ Executive Summary:

Several countries from around the world are establishing data platforms. Within the European Union member states are setting up official data catalogues as entry points to simplify public access to PSI (Public Sector Information) of the country. In this context it is important to describe the metadata of these data portals, i.e., data catalogs and allow the interoperability among those. To tackle these issues, the Government Linked Data Working Group is developing DCAT (Data Catalog Vocabulary), an RDF vocabulary for describing the metadata of data catalogs. This topic report describes the advantages of having a standard for data catalogs and analyses the multilingual perspective of DCAT.

1 Introduction

In recent years data has become the new oil. Indeed, just like oil, it needs to be discovered, extracted from its sources, and refined from the raw material into products with a high added value. Following this trend, many national, regional and local governments, as well as other organizations inside and outside the public sector, are operating data catalogs – web portals - that provide access to machine-readable public data published by these organizations. The need for a standard format to represent the metadata contained in these catalogs has been recognized (Maali et al., 2010), as a way to improve interoperability and exchange of data and in order to avoid catalogs ending up being data silos.

In this line, the W3C Government Linked Data Working Group[1] is developing DCAT (Data Catalog Vocabulary), an RDF vocabulary designed to facilitate interoperability between data catalogs published on the Web (Maali et al., 2013). DCAT was first developed and published by DERI[2] and has seen widespread adoption at the time of this publication. The original vocabulary was further developed by the eGov Interest Group [3], before being brought onto the Recommendation Track by the Government Linked Data (GLD) Working Group.

By using DCAT to describe datasets in data catalogs, publishers increase discoverability and enable applications easily to consume metadata from multiple catalogs. It further enables decentralized publishing of catalogs and facilitates federated dataset search across sites. Aggregated DCAT metadata can also serve as a manifest file to facilitate digital preservation.
Within the European Context, there are more than 70 official data catalogues[4] and there are 23 officially recognized languages. In this sense, our claim is that multilingualism is an important feature vocabularies have to take into account. This report analyzes DCAT from a multilingual perspective and proposes two options for including multilingualism in a given vocabulary.

[1] http://www.w3.org/2011/gld/charter

[2] http://deri.ie/

[3] http://www.w3.org/2009/06/eGov/ig-charter.html

[4] A list of official data catalogues in EU27 member states can be found at http://datacatalogs.org/group/eu-official

2 DCAT Overview

DCAT is an RDF vocabulary well-suited for representing data catalogs. It defines three main classes (see Figure 1):

- `dcat:Catalog`, a class that defines a curated collection of metadata about datasets. This class is described by the following properties: title, description, date of formal publication of the catalog, most recent date of modification of the catalog, language of the catalog, a link to the license document under which the catalog is made available, the rights under which the catalog can be used/reused, the spatial coverage of the catalog, and a link to the homepage of the catalog (see Figure 1).

- `dcat:DataSet`, is defined as a collection of data, published or curated by a single agent, which is available for access or download in one or more formats. As can be seen in Figure 1, there are numerous properties that describe this class: title, description, date of formal issuance and most recent date of modification, a unique identifier of the dataset, a keyword or tag describing the dataset, the language of the dataset, the temporal period that the dataset covers, the spatial coverage of the dataset, the frequency at which the dataset is published, and, finally, the landing page, i.e., a Web page that can be navigated to in a Web browser to gain access to the dataset, its distributions and/or additional information.

- `dcat:Distribution`, a class which connects a dataset to its available distributions. The latter are defined by properties such as title, description, date of formal publication and most recent date of modification, links to the license document under which the distribution is made available, a URL that gives access to a distribution of the dataset, a direct link to a downloadable file in a given format, media type of the distribution, the file format of the distribution, and the size of a distribution in bytes.

Moreover, through the `dcat:theme` property, DCAT relies on the class `skos:Concept` for classifying its datasets according to a set of domains or categories, which in its turn are categorized or organized in a taxonomy used to represent themes/categories of datasets in the catalog (`dcat:themeTaxonomy`).

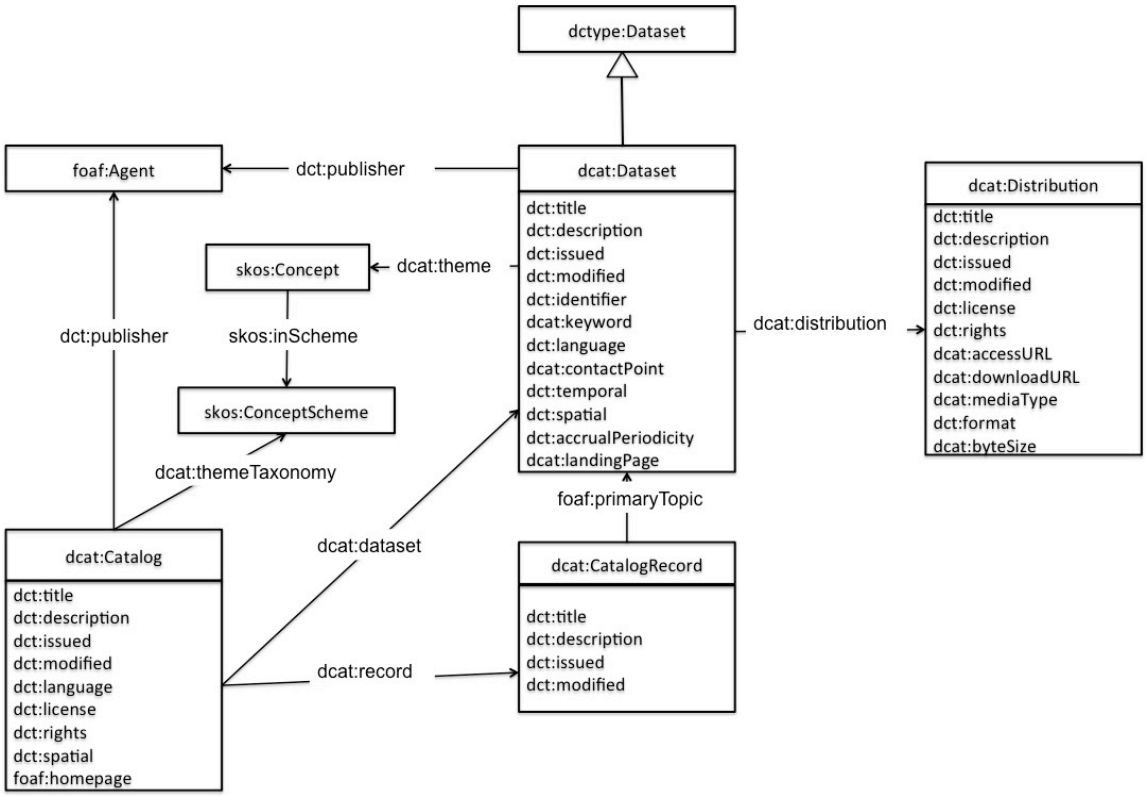

Figure 1. DCAT main classes (Maali et al., 2013)

The rest of the document is organized as follows. In section 3 we provide an analysis of some data catalogs which are annotated with the DCAT vocabulary, and which contain data in several languages. The analysis provides an interesting insight into the actual use of the vocabulary and highlights some issues related with the use of the vocabulary in multilingual settings or by publishers whose main language is not English. These issues are summarized, exemplified and discussed in section 4. Then, in section 5 we present several modeling options for the enrichment of RDF vocabularies, such as the DCAT vocabulary, with multilingual information, what would come to solve some of the problems of its current use in multilingual settings. Finally, we end the report in section 6, by presenting approaches for the representation of culturally influenced elements in vocabularies.

3 DCAT compliant data catalogs

In order to assess the current use of the DCAT vocabulary in European public data catalogs, firstly we analyzed in detail several catalogs that make use of this vocabulary. Specifically, the catalogs used in our analysis are:

- PublicData.eu Europe's public data[5]
- The data catalog of the Local Government of Gijón[6], in Spain.
- Gencat, the data catalog of the Regional Government of Catalonia[7], in Spain.
- The data catalog of the Local Government of Zaragoza[8], in Spain.

The selection of these catalogs was motivated by a previous study carried out by DERI –first developers and publishers of the DCAT vocabulary- which resulted in a list of data catalogs currently making use of this vocabulary. The list, which has been included below for convenience (see Table 1), was made available to us and contained nine catalogs in total.

[5] http://publicdata.eu/

[6] http://datos.gijon.es/

[7] http://www20.gencat.cat/portal/site/dadesobertes

[8] http://www.zaragoza.es/ciudad/risp/

Table 1: Data catalogs DCAT compliant

Catalog	Website	Type (national, local, regional)	dcat availabe at/as	#datasets	theme?	keywords?	publisher	extent (spatial)	distribution format	distribution size
Semantic CKAN	http://semantic.ckan.net	aggregate	SPARQL endpoint http://semantic.ckan.net/sparql	27592	nope	yes... represented using moat:taggedWithTag	nope	nope	yes	nope
Publicdata.eu	http://publicdata.eu	aggregate	A list of RDF files http://rdf.opendatasearch.org/	11655	yes, but each aggregated catalog has its own hierarchy i.e. themes are not reconciled	yes	yes (though uses dc:creator)	yes	yes	nope
Barcelona	http://w20.bcn.cat/opendata/	City	RDF dump http://w20.bcn.cat/opendata/CatalegRDF.aspx	501	nope	yes	yes	nope	yes	nope
Gijón	http://datos.gijon.es	City	an RDF file per dataset	29	nope	yes (though wrongly uses dc:keyword)	yes	nope	yes	yes
Catalonia	http://dadesobertes.gencat.cat	Regional	RDF dump http://dadesobertes.gencat.cat/recursos/datasets/cataleg.rdf	127	yes (though uses dc:subject)	yes	yes	yes	yes	yes
Balearic Islands	http://www.caib.es/caibdatafront	Regional	RDF dump http://dadesobertes.caib.es/recursos/datasets/catalog/rdf	28	yes (though uses dc:subject)	nope	yes (though wrongly uses dc:editor)	nope	yes (though in a wrong way)	yes (in a wrong way though)
Saragossa	http://datos.zaragoza.es	City	SPARQL endpoint at http://www.zaragoza.es/datosabiertos/sparql	271	yes (though uses dc:subject)	yes	yes(wrongly uses dc:publiser with no label or any other property. also uses the non-existing dc:editor)	yes	yes	nope
Fingal	http://data.fingal.ie	Regional	by third party (DERI)	73	yes	nope	yes	yes	yes	nope
lab.linkeddata.deri.ie	http://lab.linkeddata.deri.ie/govcat	Aggregate	SPARQL endpoint at http://lab.linkeddata.deri.ie/govcat/sparql	1710	yes	yes	yes	yes	yes	yes

In the following, we provide a brief description of the selected catalogs focusing on the purpose of the catalog, language of the portal, the language of the datasets, and the search options.

The **PublicData.eu portal** (Figure 2) aims to provide access to open datasets from local, regional, and national public bodies across Europe. Apart from being a single point of access to scattered datasets in Europe, it enables users to browse datasets by region, subject matter and format. Although the portal language is English, we can say that the PublicData.eu catalog is multilingual in the sense that it contains datasets originating from several European countries and which contain information in various natural languages. The datasets come from UK, France, Spain,

Austria, Denmark, Italy, Czech Republic, The Netherlands, Germany, Russia, Sweden, and Ireland. In fact, an additional browse option of the portal is by country of origin (see Figure 3).

Figure 2. Snapshop of the PublicData.eu portal

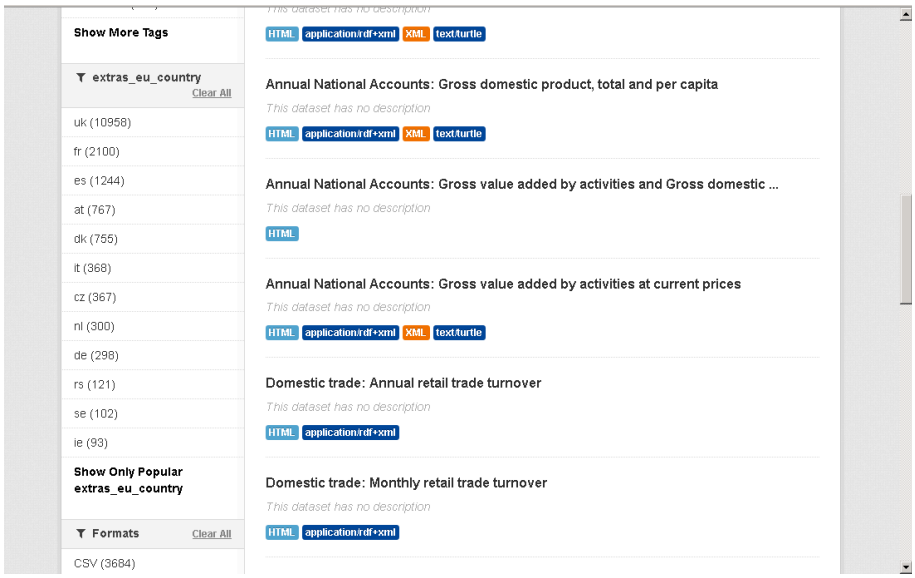

Figure 3. PublicData.eu browse option by dataset language

The next catalog, the catalog of the Local Government of Gijón, a city in the north of Spain, contains around 400 datasets in Spanish managed by the local government. The main purpose of the publication of datasets in RDF is to contribute to improve citizen participation, promote

innovation and give companies the opportunity to create added value, so that both people and market players can benefit from public information and offer new products and services. The portal is available only in Spanish and one can filter and customize the search results by using the options: keyword, category and format, as you can see in Figure 4.

Figure 4. Browse options in the portal of the Local Government of Gijón

Gencat is the third catalog we have studied (see Figure 5). Most of the datasets in this portal of the Government of Catalonia, in Spain, are in Catalan, the official language in this Spanish region together with the Spanish language. The portal is available in English, Spanish and Catalan. According to the information in the portal, this catalog comprises a database of the "26,000 official facilities of Catalonia, the 1,400 procedures handled in the Government's offices and some of its multimedia archives". The browse options are by categories (called tags in the English version of the portal), format, and data sources, i.e., the organisms that have created the corresponding datasets.

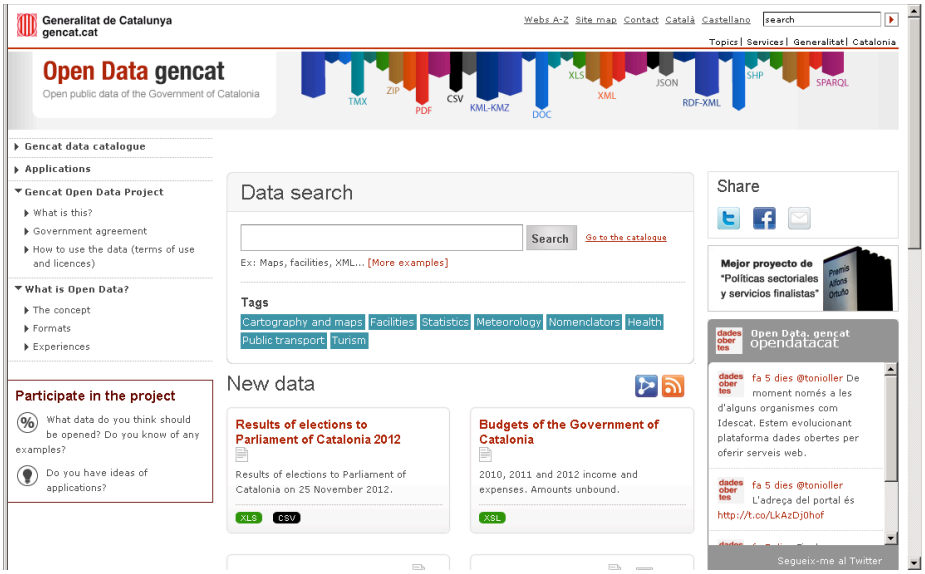

Figure 5. Snapshot of the search options in the English version of Gencat

The fourth and last catalog that we have considered for this study is the data catalog of the Local Government of Zaragoza, also a Spanish region (Figure 6). The portal is only available in Spanish, and the information in the datasets it contains is also in Spanish. As in the preceding catalogs, the purpose of this portal is to provide access to citizens to public data as well as to foster the reuse of that information. The browser or search options are categories, type of update proceeding of the dataset (yearly, quarterly, monthly, daily, instantly), format and keywords.

Figure 6. Portal of the data catalog of the Local Government of Zaragoza

4 Some issues related with the current use of the DCAT vocabulary: a language perspective

The next step in our analysis was to access some of the datasets contained in the different catalogs, available in the RDF format and annotated with the DCAT vocabulary, and look into the use they made of the DCAT classes and properties. The main conclusions of this study are discussed below.

1. Some datasets are not using the last version of the DCAT vocabulary. For example, the dataset *List des IFSI en Ile de France* contained in the PublicData.eu catalog makes use of the properties `dct:creator` and `foaf:name` to refer to the publisher of the dataset, instead of the `dct:publisher` property and `foaf:Agent` class defined by the current version of the DCAT vocabulary.

   A similar example is found in the catalog of the Local Government of Gijón. In the case of a dataset of hostels, we find the `foaf:Organization` class instead of `foaf:Agent` when defining the publisher of the datasets; or the `dc:mediaTypeorExtent` instead of the `dcat:mediaType` defined in the current version of the vocabulary.

2. Some datasets make a "free use" of the DCAT vocabulary, i.e., they are not fully compliant with DCAT. By this we mean that they use properties of a certain class in the description of another class. For instance, in the same dataset mentioned above from the PublicData.eu catalog, *List des IFSI en Ile de France*, the property `foaf:homepage` is a property of the class `dcat:Dataset`, i.e., it is describing the dataset, whereas it should be a property of the class `dcat:Catalog`, as established by the DCAT vocabulary.

3. Another remarkable aspect of the analyzed datasets is that they do not make use of the same amount or type of metadata. This may be, to some extent, reasonable, since each publisher might decide which elements of the vocabulary cover the needs of his or her catalog. Most catalogs make use of the descriptive information relative to the dataset, such as, title, description, date of issue or date of modification, and also information related to the distribution of the dataset (see the descriptive properties of the

`dcat:Distribution` class in section 2). However, very few contain information of the Catalog itself, of the Record, or of the theme and theme taxonomy used by the catalog or dataset in question.

4. When accessing the code of the dataset in RDF, we realized that ALL catalogs reused the DCAT vocabulary as it is, i.e., with the labels for classes and properties in English, as defined by the authors of the vocabulary. None of the publishers translated the DCAT vocabulary itself into its own language, even when the real data o information in the datasets was in a language different from English.

   This is the common choice when the ontology or vocabulary is *shareable* and *valid* for different cultures. By this we mean that a certain conceptual organization (i.e., the classes and properties that make up an ontology or vocabulary and the way in which they have been organized) is "universal", in the sense that it does not solely reflect the needs of a certain culture or how a certain culture approaches a particularly area of knowledge, but it is valid or translatable to other cultures. In fact, the set of classes and properties proposed in the DCAT vocabulary are general enough so as to be accepted by any publisher. Obviously, it may happen that some properties that are relevant for one publisher are not relevant for another one. For example, the `dcat:language` property defining the `dcat:Dataset` class is highly relevant for the PublicData.eu catalog, since it contains datasets in several natural languages. In the case of the catalog of the Local Government of Zaragoza this is not the case, since all datasets contain information in the same language, Spanish, and this is not a property that deserves a special mention.

   However, what we observe in the portals analyzed is that those publishers in countries where English is not the native language have provided different translations for the classes or properties of the DCAT vocabulary that are shown to the user in the web page. For example, in the catalog of the Local Government of Zaragoza, one of the search options is by **Materias** (Subjects) or **Temas** (Topics), corresponding to the `dcat:theme` property, whereas the Local Government of Gijón has translated this as **Sectores** (Sectors) o **Sectores temáticos** (Domain areas). In the case of Gencat, the catalog of the Government of Catalonia, these are **Categories** (categories), and in the PublicData.eu, they are dubbed **Groups**.

It could be said that *materia, tema, sector, sector temático, categoria* or *group* are synonyms or term variants that can be associated to the same concept. One could also argue that each publisher can translate the names of these metadata as he or she wishes, as long as the type of information provided by the different metadata is the appropriate one. And we agree with this. Nevertheless, when searching different catalogs in the same language, let us say Spanish, one would expect to find "the same information labeled in the same way". We believe this would avoid confusion and simplify the search.

For this reason, and also for tasks such as the automatic generation of web pages from ontologies or vocabularies, we would be in favor of proposing "official translations" of the DCAT vocabulary, which could be enriched with as many variants as wished in order to cover the needs of all publishers (see section 5 for more details on this).

5. The last issue which we came across regarding the use of DCAT by different publishers in Europeis that the categorization they make of the datasets is also different. The authors of the DCAT do not prescribe the topics or categories schema that should be followed when using this vocabulary. They only determine that the property `dcat:theme` be linked to a `skos:Concept`, which in its turn be included in a `skos:ConceptScheme`. Because of this, each publisher has adopted a different categorization or taxonomy of categories to classify datasets. We had the chance to ask some of the publishers and they said that they had simply adopted the categorization that the different public bodies already made in their respective web pages.

   By way of example we include below the classification followed in the PublicData.eu portal (Figure 7), and the taxonomy used in the Gencat portal from the Government of Catalonia (see Table 2).

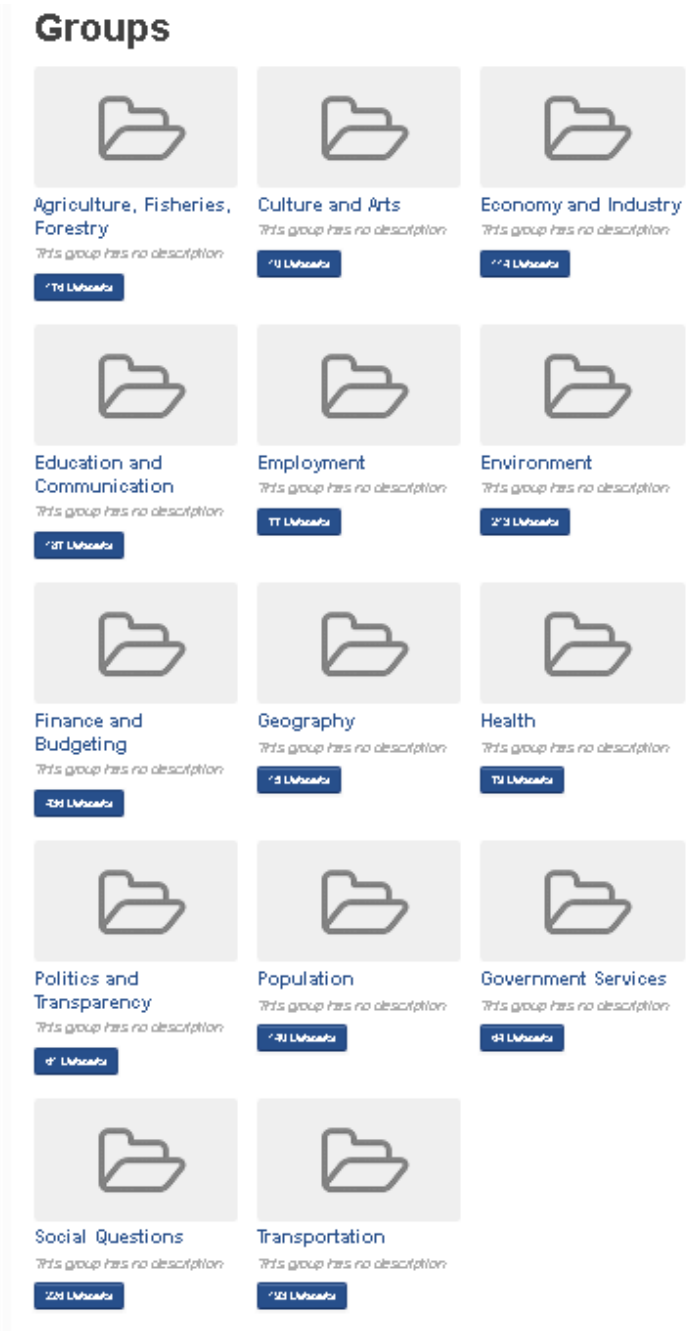

Figure 7. Classification of datasets on the PublicData.eu portal

Table 2. Taxonomy used in the Gencat portal from the Government of Catalonia

Theme	Includes	Sector ID
Administració pública	Oposicions, Publicacions, Recerca, estudis i anàlisis, Altres	administracio
Agricultura, ramaderia i pesca	Agricultura, Foment de la producció, Infraestructures rurals, Pesca. Aqüicultura, Recerca, estudis i anàlisis	agr-ram-pesca
Associacionisme i participació	Civisme, Cooperació al desenvolupament, Entitats, Equipaments, Participació ciutadana, Pau i drets humans, Voluntariat	participacio
Comerç i consum	Consum	comerc
Economia	Assegurances, Defensa competència, Deutes i sancions, Pagaments, Tributs	economia
Educació i formació	AMPA, Educació en el lleure, Educació infantil, Educació secundària obligatòria, Estudiar a l'estranger, Formació adults, Formació per a docents, Formació professional, Idiomes, Material didàctic, Mobilitat educativa, Preinscripció i matriculació, Proves d'accés, Suport a l'alumnat, Altres, Oposicions	educacio
Cultura	Arts escèniques, Arts visuals, Arxius i biblioteques, Cinema i audiovisuals, Cultura popular, Lletres, Memòria històrica, Museus, Música, Patrimoni, Recerca, estudis i anàlisis, Altres	cultura
Esports, lleure i oci	Caça, Esports, Jocs i espectacles, Nàutica i busseig, Pesca, Proves esportives, Vacances i estades	esports-oci
Indústria i energia	Estalvi energètic	industria-energia
Habitatge	Compra, Lloguer, Protecció oficial, Rehabilitació, Altres	habitatge
Justícia	Acadèmies, Associacions, Col·legis professionals, Federacions, Fundacions, Gestors administratius, Oposicions, Recerca, estudis i anàlisis	justicia
Llengua i comunicació	Occità, Llengua catalana, Mitjans de comunicació	comunicacio
Medi Ambient	Aigua, Ecologia, Estalvi energètic, Flora i fauna, Sostenibilitat, Altres	medi-ambient
Mobilitat i transport	Trànsit, Transports	transport
Salut	Assistència sanitària, Foment salut, Higiene, Recerca, estudis i anàlisis	salut
Serveis Socials	Cohesió social, Dependència, Discapacitat, Altres	serveis-socials
Societat, ciutadania i famílies	Adopció i acolliment, Afers exteriors, Afers religiosos, Dones, Gais, lesbianes i transsexuals, Gent gran, Igualtat, Immigració, Infants i famílies, Joves, Recerca, estudis i anàlisis, Altres	societat
Tecnologia. Recerca i Innovació	Foment, Noves tecnologies, Recerca, Societat de la Informació, Telecomunicacions, Innovació	tecnologia
Territori i paisatge. Urbanisme	Boscos, Costes, Paisatge, Ports	territori
Treball	Borsa de treball, Cooperatives, Emprenedoria, Formació, Igualtat d'oportunitats, Ocupació, Oposicions, Relacions laborals, Seguretat i salut laboral	treball
Turisme	Establiments turístics, Foment turisme, Altres	turisme

Theme	Includes	Sector ID
Universitats	Formació en empreses, Mobilitat educativa, Postgraus, doctorats i màsters, Preinscripció i matriculació, Proves d'Accés a la Universitat, Recerca, Altres	universitats

In the Gencat taxonomy, the first row identifies the main topics; the second, the sub-topics related to the main ones, and the third row, the term used to identify the topics in the web page. At a first sight, one realizes that the categorizations of the PublicData.eu portal and the ones from Gencat do not match in coverage or granularity.

In the case of categorizations of datasets, we observe some general concepts or topics which are present in any categorization, whereas others, maybe more culturally bound, are only present in certain categorizations. In the case of PublicData.eu, they propose 14 groups, such as “Agriculture, Fisheries and Forestry” or “Economy and Industry”, whereas the Gencat catalog proposes 22 broad categories. When comparing the PublicData.eu catalog to the Gencat one, we see that there is some overlap. In both catalogs we find datasets grouped under the categories of “Agriculture, Fisheries and Forestry” – “Agricultura, ramaderia i pesca”; “Culture and Arts” – “Cultura”; “Environment” – “Medi Ambient”; or “Health” – “Salut”. In other cases, the PublicData.eu catalog merges related categories, and the Gencat catalog keeps them apart. It is the case of “Education and Communication” in the PublicData.eu catalog vs. “Universitats” (Universities), “Tecnologia, recerca i innovació” (Technology, research and innovation), “Llengua i comunicació” (Language and communication) and “Esports, lleure i oci” (Sports, free time and hobbies). In this sense, we can say that the taxonomy of categories used by the Gencat catalog is far more specific than the Public.Data.eu catalog.

Due to the fact that the Gencat catalog has been created by the Government of a region, in this case, Catalonia, it contains categories that comply with the purpose of the catalog, such as “Administració pública” (Public administration). Also highly related with the distinctive features of this region, we find the category of “Llengua i comunicació” (Language and communication), which contains the sub-categories “Occità, Llengua catalana, Mitjans de comunicació” (Occitane, Catalan language, Media). Such distinctions may represent an issue for portals such as the PublicData.eu one, which aim

at providing a unified access to datasets across Europe, because they will have to previously analyze the categorizations made by the different portals to aggregate datasets under meaningful categories. Since no categorization or taxonomy is prescribed by the DCAT vocabulary, after analyzing several categorizations of this sort, a tentative categorization could be proposed, which could be extended or adapted to cover specific cultural needs. For this purpose, several representational approaches could be followed. We provide a summary of this in section 6.

5 Enriching RDF vocabularies with multilingual information

As mentioned in point 4 of the above section, with the aim of enhancing the use of the DCAT vocabulary at an international level, it would be recommendable to provide translations of the labels that describe the DCAT classes and properties to languages other than English. Some of the advantages of having multilingual versions of this vocabulary would be that publishers in countries where English is not the official language could make use of these descriptions in their own language, and they could also directly reuse these terms or labels in their portals or final applications. This would also result in all portals making use of the same terms or labels, contributing in this way to interoperability.

The idea of enriching ontologies and RDF vocabularies with multilingual linguistic information is not new and has been the object of research and study for a decade now. To the best of our knowledge, some of the first approaches to enrich ontologies with linguistic descriptions are LingInfo (Buitelaar et al., 2006), LexOnto (Cimiano et al. 2007), LIR-Linguistic Information Repository (Peters et al., 2007; Montiel-Ponsoda et al., 2010) or LexInfo (Buitelaar et al., 2009; Cimiano et al., 2010). These models mainly differ in the type of linguistic descriptions they aim at accounting for. For instance, whereas the LingInfo model focused on the representation of the morphological and syntactic structures of those labels or terms describing ontology classes and properties, the LIR model focused on the representation of term variants and translations. Currently, researchers in this domain have joined forces and are working towards the standardization of a model that will intend to capture a wide range of linguistic descriptions relative to ontologies or RDF vocabularies. We are referring to the W3C Ontology-Lexica Community Group[9] . This standardization initiative has taken the *lemon* (LExicon Model for ONtologies) model (McCrae et al., 2011; http://lemon-model.net/) as basis for its work, and it is evolving it into a model which, in combination with the semantic information captured in the ontology, is aimed at improving the performance of NLP (Natural Language Processing) tools, amongst other objectives. See Figure 8 to gain an impression of the kind of linguistic descriptions that can be associated to ontology elements in the *lemon* model.

[9] http://www.w3.org/community/ontolex/

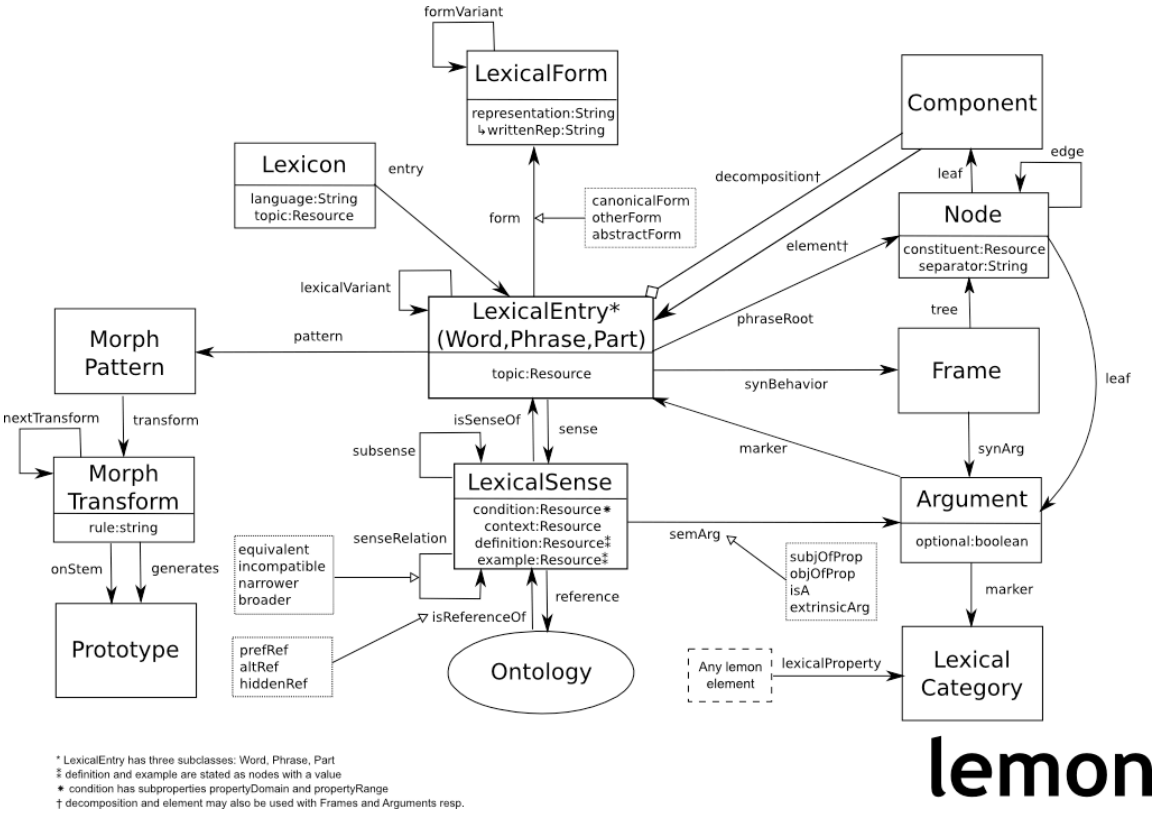

Figure 8. Overview of the lemon model

As for the specific case of the DCAT vocabulary, a model such as *lemon* would allow for the inclusion of term variants in different languages for the classes and properties of the vocabulary. Coming back to the previously mentioned example of the several translations in Spanish of the `dcat:theme` property (*materia*, *tema*, *sector*, *sector temático*, *categoria* or group), they could all be accounted for as variants or linguistic realizations of the property `dcat:theme`. For a property such as `dct:modified`, we could have "more readable" terms or labels (last update, change date, *fecha de modificación, fecha de actualización, Änderungsdatum, Datum der letzten Aktualisierung*, etc.), which could then be used for the automatic generation of web pages.

6 Approaches for the representation of culturally influenced elements in ontologies

Closely related with the approaches proposed to enrich ontologies and RDF vocabularies with multilingual linguistic information is the issue of capturing culturally-bounded classes and properties in ontologies. As mentioned in issue number 5, section 3, the DCAT vocabulary does not prescribe any categorization or taxonomy of categories or themes into which datasets can be classified. In the catalogs analyzed we found out that the categorizations of datasets showed some differences, mainly motivated by the idiosyncrasy of the catalogs themselves, and the culture and language in which they had been developed. In this sense, it would be advisable to propose a taxonomy and analyze which approach is the most suitable to meet the needs of most (if not all) publishers.

Taking into account previous work on ontology localization (Montiel-Ponsoda et al. 2010, Cimiano et al. 2010), we envision two possibilities:

1. To map the different categorizations by means of a mapping model
2. To maintain one categorization and to represent cultural issues in an external linguistic model or as specific language modules or extensions in the ontology

The first approach allows for each publisher maintaining its own categorization, and all of them being mapped or linked to a central categorization (see Figure 9 from Montiel-Ponsoda, 2011). However, the mapping establishment may be a tough task, and some scalability issues may also appear as more and more datasets use the DCAT vocabulary.

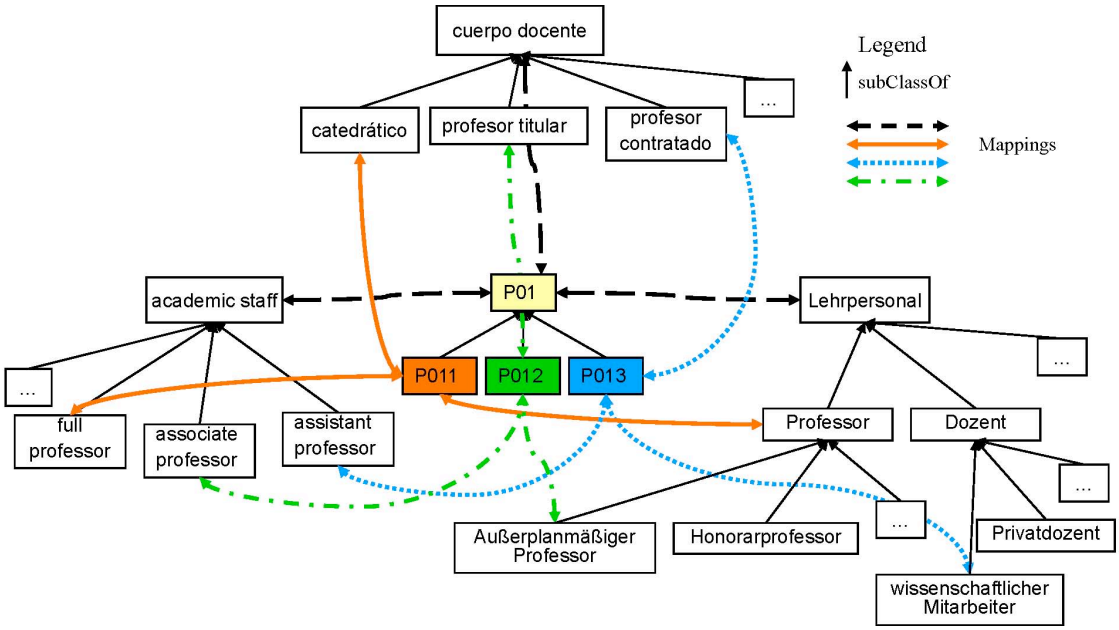

Figure 9. Mapping model

As for the second option, Figure 10, one categorization would be shared by all publishers, and in case of cultural issues, these could be kept in the linguistic model, or, if needed, "specific cultural modules" could be proposed to extend the original categorization. The main advantage of this latter approach is that it contributes to interoperability, but without forgetting culturally bound issues. In the case of the DCAT vocabulary, we would be in favor of this latter option.

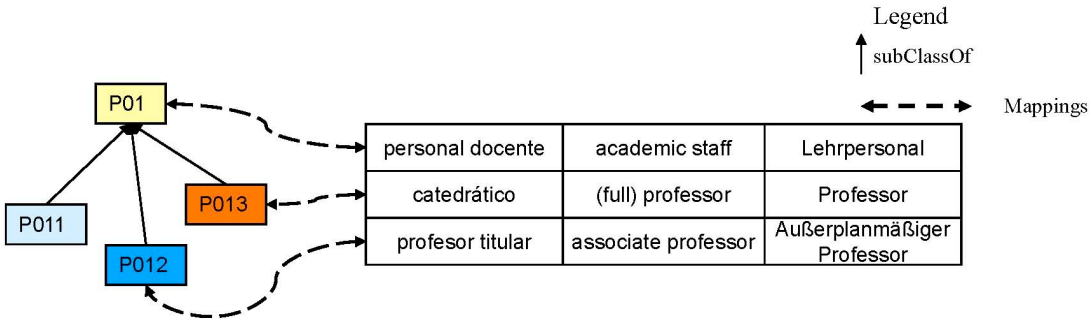

Figure 10. Vocabulary linked to an external model

Again, the *lemon* model described in section 4 (or the model that will result from the W3C Onto-Lexica Community Group) would come to solve the modelling issues involved in this latter model.

7 Conclusions and recommendations

The number of data catalogs in Europe is increasing. Lately, there is a trend in public administrations (regional, local, national and European) to public government data in data catalogs. DCAT, a vocabulary for representing metadata of data catalogs, is being developed within the Government Linked Data W3C Working Group. Thanks to DCAT, publishers increase discoverability and enable applications to easily consume metadata from multiple catalogs.

In this report we have presented:

- an overview of DCAT;
- a set of data portals that rely on DCAT for describing their metadata;
- an analysis on how the several portals actually use DCAT, which classes and properties they use, and, in particular, how they represent the themes of their catalog;
- the possible options of enriching DCAT with multilingual information, to be able to represent data catalogues in different languages

Our main recommendation is to consider the multilingualism aspect in any vocabulary, since, on the one hand, it may contribute to its global adoption, and, on the other, it may also add to interoperability. To this respect we have proposed *lemon*, a model for the representation of linguistic information relative to an ontology or RDF vocabulary that is currently being reviewed for standardization purposes.

Ideally, multilingualism should be considered as early as possible, so that specificities of certain languages could be approached as soon as possible. This would also allow for a prescriptive approach, in which publishers are said which labels to use in each case. However, the process rarely follows this order. As vocabularies gain popularity, their adoption increases and multilingual needs appear to support interoperability. In fact, widespread adoption comes first, and, then, one realizes the benefits of the multilingual aspect. For these reasons, models such as *lemon* allow to maintain the model or vocabulary "as it is", and enrich it with multilingual information at any stage of the process. In the specific case of the DCAT vocabulary, and taken into account its general adoption, the next step would involve an analysis of the catalogs and portals that implement it to identify the labels used by the various publishers in different languages. All those labels, or preferably, the ones that better express the meaning of the vocabulary terms should be captured in the linguistic model and recognized as preferred labels

in each language. The benefit of this approach is that the model would take advantage of labels (variants or translations) that are popular and accepted by publishers, and would not "impose" the use of some labels that may end up not being meaningful for users. The model would also "leave the door open" for new linguistic needs without interfering with the original vocabulary. Moreover, we believe that it should be made following a conciliatory approach in which different options are welcomed and integrated, and in which different communities can participate in proposing terms and translations in their own languages, thus building it in a cooperative way. All in all, the enrichment of the vocabulary with multilingual linguistic information would contribute to a wider adoption and increased understanding and interoperability.

About the Authors

Elena Montiel Ponsoda is Lecturer at the Universidad Politécnica de Madrid, in Madrid, Spain, and member of the Ontology Engineering Group at the same university. She received her M.A. in Conference Interpreting and Translation (September 2000) by Universidad de Alicante, her B.A. in Technical Interpreting (February 2003) by Hochschule Magdeburg-Stendal, Germany, and her PhD on Applied Linguistics (January 2011) by Universidad Politécnica de Madrid. Her current research activities include, among others: Terminology and Translation in the field of Information Technology and Natural Language Processing (NLP), in which she has participated in different international projects concerning terminology, ontologies and multilingualism and its application to the Semantic Web. She has published the book "Multilingualism in Ontologies. Building Patterns and Representation Models", and numerous papers in journals, conferences and workshops in the areas of Applied Linguistics, Semantic Web, and NLP.

Boris Villazón-Terrazas is Linked Data Researcher Manager at iSOCO. He holds a PhD in Artificial Intelligence from Universidad Politécnica de Madrid. He has previously worked as Post-Doc at the Ontology Engineering Group. Before he was a researcher and software developer at the Research Institute of Informatics at the Universidad Católica Boliviana San Pablo. His research interests are focused on Linked Data, Semantic Web and Ontology Engineering, among others. He has participated in several European research projects such as Knowledge Web, OntoGrid, SEEMP, NeOn, SemsorGrid4Env, PlanetData, and Parlance, as well as in national projects such as Reimdoc, Servicios Semánticos, Plata, Gis4Gov, WebN+1, Buscamedia and Ciudad2020. Moreover, he was leading the Spanish Linked Data initiatives, such as GeoLinkedData, datos.bne.es, AEMETLinkedData, and El Viajero. Finally, he has published more than 40 papers in journals, conferences and workshops, and currently he is actively participating in the RDB2RDF, and Government Linked Data W3C Working Groups.

References

1 Maali, F. & Cyganiak, R. & Peristeras, V. (2010). **Enabling Interoperability of Government Data Catalogues**. Electronic Government 10th International Conference

2 Maali, F. & Erickson, J. & Archer, P. (2013). **Data Catalog Vocabulary (DCAT), W3C Last Call Working Draft.**

3 Buitelaar, P., Declerck, T., Frank, A., Kiesel, M., Sintek, M., Romanelli, M., Sonntag, D., Loos, B., Micelli, V., & Porzel, R. (2006). **LingInfo: Design and Applications of a Model for the Integration of Linguistic Information in Ontologies**. In Proceedings of Ontolex 2006.

4 Cimiano, P., Haase, P., Herold, M., Mantel, M., and Buitelaar, P. (2007). **LexOnto: A Model for Ontology Lexicons for Ontology-based NLP**. In Proceedings of the OntoLex07 Workshop at the ISWC07.

5 Peters, W., Montiel-Ponsoda, E., Aguado de Cea, G., and Gómez-Pérez, A. (2007). **Localizing ontologies in OWL**. In From text to knowledge, the lexicon/ontology interface, proceedings of the Ontolex07 workshop. Busan, South Corea.

6 Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., and Peters, W. (2010). **Enriching Ontologies with Multilingual Information**. Journal of Natural Language Engineering, 17 (3), 283-309.

7 Buitelaar, P., Cimiano, P., Haase, P., and Sintek, M. (2009). **Towards linguistically grounded ontologies**. In Proceedings of the 6th European Semantic Web Conference (ESWC09), 111-125.

8 Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., and Gómez-Pérez, A. (2010). **A note on ontology localization**. Journal of Applied Ontology, 5(2), 127-137.

9 McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Goméz-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T. (2011). **Interchanging lexical resources on the Semantic Web**. En Language Resources and Evaluation, 46, 701-719.

10 Montiel-Ponsoda, E. (2011). **Multilingualism in Ontologies. Building Patterns and Representation Models**. LAP - Lambert Academic Publishing.

.

Copyright information