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

Elena Montiel-Ponsoda[1], Boris Villazón-Terrazas[2]
1 Ontology Engineering Group, Universidad Politécnica de Madrid, Spain
2 Intelligent Software Components, iSOCO, Madrid, Spain

Abstract. Within the European Union, member states are setting up official data catalogues as entry points to access PSI (Public Sector Information). In this context, it is important to describe the metadata of these data portals, i.e., of data catalogs, and allow for interoperability among them. To tackle these issues, the Government Linked Data Working Group developed DCAT (Data Catalog Vocabulary), an RDF vocabulary for describing the metadata of data catalogs. This topic report analyzes the current use of the DCAT vocabulary in several European data catalogs and proposes some recommendations to deal with an inconsistent use of the metadata across countries. The enrichment of such metadata vocabularies with multilingual descriptions, as well as an account for cultural divergences, is seen as a necessary step to guarantee interoperability and ensure wider adoption.

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 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 and has seen widespread adoption at the time of this publication. The original vocabulary was further developed by the eGov Interest Group , before being brought onto the Recommendation Track by the Government Linked Data (GLD) Working Group.

2 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[1]
- The data catalog of the Local Government of Gijón[2], in Spain.
- Gencat, the data catalog of the Regional Government of Catalonia[3], in Spain.
- The data catalog of the Local Government of Zaragoza[4], in Spain.

[1] http://publicdata.eu/
[2] http://datos.gijon.es/
[3] http://www20.gencat.cat/portal/site/dadesobertes
[4] http://www.zaragoza.es/ciudad/risp/

3 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.

- 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 `dct:mediaType` defined in the current version of the vocabulary.
- 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.
- 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. 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.
- 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.
- The last issue which we came across regarding the use of DCAT by different publishers in Europe is 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.

4 Enriching RDF vocabularies with multilingual information

As mentioned in the fourth point 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. 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.

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.

5 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 the fifth issue, 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 1 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.

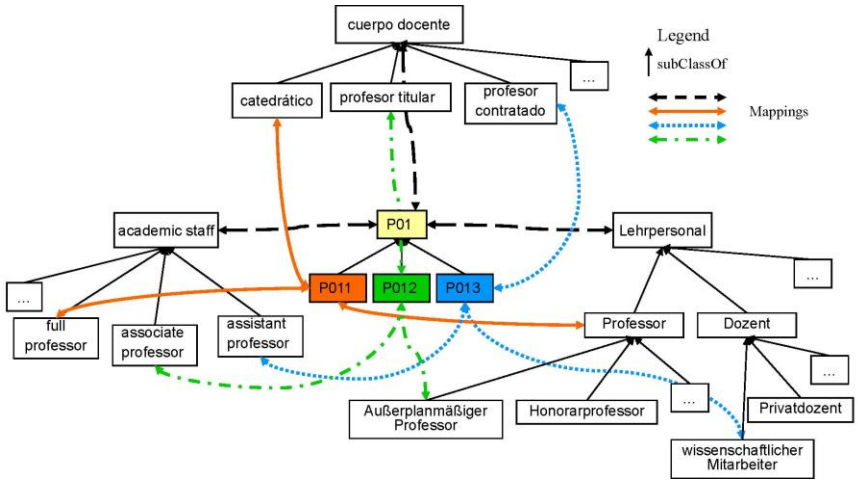

Figure 1. Mapping model

As for the second option, Figure 2, 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.

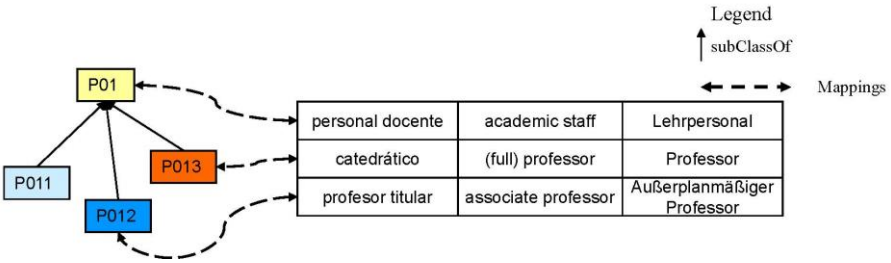

Figure 2. Vocabulary linked to an external model

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.

6 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.

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.

7 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.

8 Type of the presentation proposed

Research contribution

9 Contributors

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.