Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 212 (2015) 151 – 158

Procedia
Social and Behavioral Sciences

MULTIMODAL COMMUNICATION IN THE 21ST CENTURY: PROFESSIONAL AND ACADEMIC CHALLENGES. 33rd Conference of the Spanish Association of Applied Linguistics (AESLA), XXXIII AESLA CONFERENCE, 16-18 April 2015, Madrid, Spain

Lexicalizing Ontologies: The Issues Behind The Labels

Guadalupe Aguado-de-Cea*, Elena Montiel-Ponsoda, María Poveda-Villalón, Olga Ximena Giraldo-Pasmin

Ontology-Engineering Group, Universidad Politécnica de Madrid,Campus de Montegancedo, Boadilla del Monte 28660, Madrid, España

Abstract

In information science, ontologies are used to capture knowledge about some domain of interest, by formally naming and defining the types, properties and interrelationships of the concepts that describe that domain. They are the building blocks of the Linked Data initiative in which datasets of related domains are linked to each other, and also to more general datasets, resulting in a huge space of interconnected data. It is precisely in this linking step that the linguistic descriptions used to name or label ontology entities (i.e., concepts, properties, attributes) become undeniably significant. This has also an impact in the subsequent process of ontology localization, thus turning it into a critical process. Based on our experience in the lexicalization and localization of several well adopted ontologies (FOAF, GoodRelations, the Organization ontology, among others) from English into Spanish, we propose a preliminary set of guidelines with a twofold goal. First, to guide users in the process of assigning labels and descriptions to ontology entities; second, to help terminologists and translators in the translation of these specific resources by providing them with coherent, user-friendly examples of how to apply the above mentioned guidelines

Peer-review under responsibility of the Scientific Committee of the XXXIII AESLA CONFERENCE

Keywords: ontology lexicalization; ontology localization; ontology entity labels

* Corresponding author. Tel.: +34913367415; fax: +34913524819.
*E-mail address:*lupe@fi.upm.es

Peer-review under responsibility of the Scientific Committee of the XXXIII AESLA CONFERENCE
doi:10.1016/j.sbspro.2015.11.313

1. Introduction

In information science, ontologies are used to capture knowledge about some domain of interest, by formally naming and defining the types, properties and attributes of the concepts that describe that domain. Ontology entities are generally represented using machine-readable languages such as RDFs - Resource Data Framework Schema (Brickley & Guha, 2014), SKOS - Simple Knowledge Organization System (Miles & Bechhofer, 2008), or OWL - Web Ontology Language (McGuinness & van Harmelen, 2004). These languages provide mechanisms to associate natural language descriptions to ontology entities, a process termed "ontology lexicalization" (Cimiano et al., 2014). For instance, in Fig. 1 we can see an excerpt from the GoodRelations ontology, an ontology about e-commerce, in which the concept `gr:BusinessEntity` is labelled "Business entity" by means of the `rdfs:label` annotation, and is also provided with a definition by means of the `rdfs:comment` annotation, both in the English language, as specified by the XML annotation, `xml:lang="en"`. (A table with all the prefixes related to the ontologies mentioned in the paper is included in the Appendix.)

```
<owl:Class
    rdf:about="http://purl.org/goodrelations/v1#BusinessEntityType">
  <rdfs:label xml:lang="en">Business entity</rdfs:label>
  <rdfs:comment xml:lang="en">An instance of this class represents the
legal agent making (or seeking) a particular offering. This can be a
legal body or a person. A business entity has at least a primary
mailing address and contact details. (…)
  </rdfs:comment>
</owl:Class>
```

Fig. 1. Excerpt from the GoodRelations ontology

In the same way, other ontology languages or vocabularies have specific annotations for these natural language descriptions (e.g., `skos:prefLabel` in SKOS or `lemon:LexicalEntry`, in the *lemon* -Lexical Model for Ontologies- vocabulary, accessible at http://lemon-model.net/). Such descriptions not only facilitate the understanding of the knowledge represented by the ontology, but they also contribute to a quick adoption of the ontology by application developers.

Recently, ontologies have also become the building blocks of the Linked Data initiative (Bizer et al., 2009). This initiative is based on the publication, sharing and interlinking of data, in the same way as we publish and link documents on the Web. The vocabularies used to publish data sets are represented as ontologies, and they allow data sets of related domains to be linked to each other, and also to more general data sets, resulting in a huge space of interconnected data.

It is precisely in the linking step that the natural language descriptions used to term or label ontology entities become undeniably significant. Moreover, those data sets may not be in the same language, nor the ontologies used to structure and represent them. Consequently, some vocabularies have to be translated or localized to ensure a smooth linking process. The process of adapting an ontology to a particular linguistic and cultural community is known as "ontology localization" (Espinoza et al. 2012, Gracia et al. 2012).

Therefore, be it for the purposes of contributing to the linking process in the context of Linked Data or for the adoption by users from different linguistic and cultural communities, ontology lexicalization and ontology localization have become a priority in the current ontology engineering research. In this sense, there is a lack of guidelines or recommendations for the performance of both tasks from a linguistic perspective that needs to be addressed to meet the challenges mentioned.

Based on our experience in the lexicalization of ontologies (AEMET, http://aemet.linkeddata.es/ontology/, El Viajero, http://webenemasuno.linkeddata.es/ontology/OPMO) and the localization of several well adopted ontologies (FOAF, http://xmlns.com/foaf/spec/, GoodRelations, http://www.heppnetz.de/projects/goodrelations/, and the ORG Organization ontology, http://www.w3.org/TR/vocab-org/, amongst others) from English into Spanish, as well as on previous literature in the field (Noy & McGuiness, 2001; Schober et al., 2009; Montiel-Ponsoda et al., 2011), we propose a preliminary set of principled reflections with a twofold goal. Firstly, to guide users in the label assignment task within the ontology lexicalization process. Secondly, to help general users, terminologists and translators in the translation/localization of these specific resources by providing them with coherent, user-friendly

examples on how to apply the above mentioned principles.

In this paper, we focus on certain linguistic problems that currently occur in both processes, since they can be considered as the two sides of the same coin. Some of these problems are derived from the concision typically specific of ontology labels, due to the use of URIs (uniform resource identifier), i.e., the identifier given to ontology entities. Others have to do with different conceptualizations in several natural languages. The paper is organized as follows. In Section 2, we review the principles or guidelines when providing URIs. Then, we analyze current practices in ontology lexicalization and localization in the light of real examples (section 3). After this, we propose a preliminary set of best practices in ontology lexicalization and in the ontology localization process (section 4 and 5, respectively). Finally, we conclude with some reflections on both processes (section 6).

2. State of the art

Terms or labels traditionally used in ontologies can be characterized by its conciseness. We argue that this concision has its origin in the URI or (uniform resource identifier), i.e., a string of characters used to name and uniquely identify a resource, or in this case, an ontology entity. If we take the example in Figure 1, the URI for the ontology entity labelled "Business entity" is http://purl.org/goodrelations/v1#BusinessEntity. This is a unique identifier of the concept business entity on the Web.

Usually, the last part of the URI consists of a descriptive name of the ontology entity in natural language. This type of URI has received the name of "meaningful URI" (Montiel-Ponsoda et al., 2011) or "descriptive URI", according to the terminology used in Multilingual Linked Open Data Patterns document at http://www.weso.es/MLODPatterns/. However, this is not always the case. In some ontologies URIs are "opaque", and specific annotations (labels) are used to provide a human readable version. See for example the code for the class or concept "spatial region" from the OBO (Open Biological and Biomedical Ontologies, http://www.obofoundry.org/) ontologies. Its URI is http://purl.obolibrary.org/obo/BFO_0000006, and a label has to be added to make it readable.

```
<!-- http://purl.obolibrary.org/obo/BFO_0000006 -->

<owl:Class rdf:about="http://purl.obolibrary.org/obo/BFO_0000006">
   <rdfs:label xml:lang="en">spatial region</rdfs:label>
</owl:Class>
```

Fig. 2. Excerpt from the OBO library ontology

In the case of meaningful URIs, the descriptive part of the identifier has been traditionally adopted to lexicalize the ontology entity as well (by means of a specific annotation, such as `rdfs:label`, as already mentioned). Since some guidelines have been suggested for the provision of meaningful URIs, they have also been followed when assigning labels to ontology entities.

In this sense, literature on the use of URIs has always recommended the use of "short URIs", as suggested by one of the fathers of the WWW, Tim Berners-Lee, in "Cool URIs don't change" (http//www.w3.org/Provider/Style/URI.html). Noy and McGuiness in their seminal paper (2001) claim the use of lower case for ontology properties and upper case for concepts. To continue with our example of the GoodRelations ontology, the ontology property "category" would be represented by the following identifier: http://purl.org/goodrelations/v1#category

We can observe how the final part of the URI contains the word category in lower case, whereas "BusinessEntity" is capitalized.

Later, Schober et al. (2009) offered a comprehensive set of URI naming conventions that can be classified into two types: conventions on content and conventions on format. Regarding the former, the authors encourage the use of homonyms and conjunctions, and advise the use of positive names (without a negative particle), and the recycling of strings rather than using synonyms. As for the latter, the following ones are relevant for our paper: (1) use explicit and concise names ("wall of esophagus" instead of "the wall of the esophagus"); (2) prefer singular nominal forms; (3) use space as word separators preferably, and if not possible, use underscores; (4) expand abbreviations and

acronyms; (5) prefer lower case beginnings for concept and property names “as they would appear in normal English written text”; (6) use plain ASCII format and avoid accents.

Should we follow those conventions, we would have to re-write the URI for business entity in the GoodRelations ontology as: http://purl.org/goodrelations/v1#business_entity

As we can see, the adoption of these guidelines for URIs justifies somehow the conventions followed when lexicalizing ontologies. In the ontology lexicalization context, some pioneering guidelines were provided by Montiel-Ponsoda et al., 2011. These authors recommend to (1) use the singular form for nouns that describe classes; (2) use verbal phrases and the predicate or range in object properties for disambiguation purposes; (3) use spaces as word delimiters, since it supports readability; (4) use upper or lower case according to the language conventions of each language; (5) include as many labels as needed (synonyms) to describe ontology entities.

An interesting aspect of these recommendations is that they foresee the possibility of lexicalizations in languages other than English. Moreover, the authors make a clear distinction between meaningful URIs and labels, thus, proposing specific guidelines for the latter.

3. Some issues in ontology lexicalization and ontology localization

Before we analyze some examples of labels given to ontology entities, we should make a distinction between labels for describing concepts, and labels for naming properties of concepts or relations between concepts. We claim that whereas some recommendations would apply to both, others would be specific to each type.

Let us examine some labels from the ORG Organization ontology which are representative of ontology labels in general. There we find labels for classes such as:

- `rdfs:label "Organization"@en`
- `rdfs:label "OrganizationalUnit"@en`
- `rdfs:label “OrganizationalCollaboration”@en`
- `rdfs:label “ChangeEvent”@en`

And the following labels for properties:

- `rdfs:label "identifier"@en`
- `rdfs:label "memberOf"@en`
- `rdfs:label "headOf"@en`
- `rdfs:label "hasSite"@en`
- `rdfs:label "basedAt"@en`

As for concept labels, it is usual to find nouns in the singular form (Organization), adjective noun collocations (Organizational Unit, Organizational Collaboration), and compounds (Change Event).

In the case of labels for properties, common labels are nouns (identifier), noun plus preposition combinations (member of, head of), verb plus noun combinations (has site), and past participle plus preposition combinations (based at). So, the type of labels for properties is even wider. In the GoodRelations ontology, we additionally find:

- `rdfs:label "color"@en`
- `rdfs:label "hasValue"@en`
- `rdfs:label "serialNumber"@en`
- `rdfs:label "closes"@en`
- `rdfs:label "isSimilarTo"@en`

Thus, we should add adjective plus noun collocations (serial number), verbs in the third person singular (closes), and verb plus adjective plus preposition combinations (is similar to).

Without claiming to be exhaustive, what these examples aim to show is that no patterns or guidelines are followed in the ontology lexicalization task, but it is left to the individual criteria of ontology engineers or ontology

engineering teams. So within the same ontology we may find differences even in the same syntagmatic structure. How could we explain otherwise that “identifier”, “color” and “serial number” are not preceded by “has”, as in the case of “site” or “value”? Why “member of” or “head of” appear without the complete verbal phrase, but “is similar to” contains the whole verbal phrase?

As has been shown, inconsistencies are found even within the same ontology, let alone in different ontologies, and this is a further demonstration that guidelines are needed. If we now think of the localization process of these ontologies, we immediately realize that different syntagmatic structures may be more natural in some languages other than English. What should be done in that case? Should we stick to the English structure despite being less common in the target language?

A further problem arises when providing a translation for a label, if no definition or comment accompanies that label. Even if the rest of labels of related concepts provide some sort of context, it may not be sufficient. As can be seen in Figure 1, a definition explains the meaning of the concept, but this is not always the case, not to say, rarely. In the ontology engineering jargon, it would be said that the ontology is badly documented.

At a more conceptual level, we encounter difficulties with no easy solution that are also related to the ontology localization process. In other words, those related to the way in which a certain language and culture understands and structures reality. Let us illustrate this idea with an example from the FOAF ontology. There we find six concepts related to the name of a person: name, first name, last name, given name, surname, and family name. When translating these into Spanish, the difficulties are not so much a question of labels, but of conceptual structures.

Last but not least, we would like to mention that in the case of labels for properties defining relations between concepts, we may need to resort to the concepts those relations are linking in order to understand the label and provide a suitable translation. Figure 3 illustrates this with two examples from the FOAF ontology: the relations “made” and “maker”. These are two inverse relations. “made” relates `foaf:Agent` to an `owl:Thing` made by it, as explained in the `rdfs:comment`. Whereas, “maker” relates `owl:Thing` to the `foaf:Agent` who made it. In a Spanish translation we could opt for a literal translation *hizo* or *creó* and *creador*, or for more elaborated structures that give an idea of the directionality of the relation and correspond to natural syntagmatic structures in Spanish, for example, *es creador/a* de and *ha sido creado/a por*.

```
<rdf:Property rdfs:label="made" rdfs:comment="Something that was made
by this agent.">
   <rdf:type
       rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
   <rdfs:domain rdf:resource="http://xmlns.com/foaf/0.1/Agent"/>
   <rdfs:range rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
   <rdfs:isDefinedBy rdf:resource="http://xmlns.com/foaf/0.1/"/>
   <owl:inverseOf     rdf:resource="http://xmlns.com/foaf/0.1/maker"/>
</rdf:Property>

<rdf:Property rdfs:label="maker" rdfs:comment="An agent that   made
this thing.">
   <owl:equivalentProperty
       rdf:resource="http://purl.org/dc/terms/creator"/>
   <rdf:type
        rdf:resource="http://www.w3.org/2002/07/owl#ObjectProperty"/>
   <rdfs:domain rdf:resource="http://www.w3.org/2002/07/owl#Thing"/>
   <rdfs:range rdf:resource="http://xmlns.com/foaf/0.1/Agent"/>
   <rdfs:isDefinedBy rdf:resource="http://xmlns.com/foaf/0.1/"/>
   <owl:inverseOf rdf:resource="http://xmlns.com/foaf/0.1/made"/>
</rdf:Property>
```

Fig. 3. Excerpt from the FOAF ontology

Having analyzed some of the main problems in the process of ontology lexicalization and ontology localization, in the following, we present a set of principled reflections that may help ontology engineers or linguists involved in any of the two processes.

4. Best practices for ontology lexicalization

All As in previous sections, the distinction between labels for classes and labels for properties is maintained in the set of the best practices listed/proposed below:

- Labels for classes should be as short as possible, self-contained, meaningful, and concise (capturing or summarizing in a concise manner the meaning of the class). However, we should respect the common or natural syntagmatic structures of each language. For example, in the case of compounds, in English we would have "event change", whereas in Spanish it would be *evento de cambio*.
- We advise against the use of compact abbreviated words or terms, because the main objective of labels is to be as descriptive and meaningful of the ontology entity as possible, in order to guarantee an appropriate use and, hence, the adoption of the ontology by final users. For example, for the "classification" property label, we would suggest, *se clasifica de acuerdo con*.
- We are in favor of including several labels (synonyms or term variants) if used as equivalents in a certain domain.
- Terms proposed as labels for classes should follow the specific conventions of term formation accepted for each language. For example, in Spanish common nouns are written in small letter or lower case letter.
- Labels for classes should be in the singular. It would be advisable to choose a label that accepts suffixes, thus complying with a derivative paradigm: creator, creation, is creator of, created by.
- In the case of labels for properties, we propose labels forming a syntagmatic pattern, i.e., a syntactic unit composed of at least one verb and the syntagmatic unit(s) that accompany that verb, usually the object representing the nearest argument of the verb and or preposition. For example, *tiene* (has) or *tiene sede en* (has site in). In this way we are able to "read" the triple as a sentence, or with a syntactic structure quite close to the natural language interpretation. For instance, *Organización – tiene sede en – Sede* (Organization- hasSite-Site). Again, labels for properties should also follow the conventions of the language we are translating them into.
- When facing gender-based languages, labels should reflect gender for several reasons. In Spanish, gender should be considered in past participles in verbal patterns, since the subject of the relation can refer to nouns in feminine or masculine. For example, *está ubicada en*, *está ubicado en*, referring to an institution, *organización*, which has feminine gender, but we can also find *negocio* which is masculine.

All definitions (comments, usage notes) should follow the same format. In the case of Spanish, the format followed is the one proposed by the UNE ISO 1087-1, following ISO 1087-1 and ISO 1087-2 in which an intensional definition states the superordinate concept and the delimiting characteristics. This means that definitions should not start with a verb as was the case for the English version of the ORG ontology.

5. Recommendations for ontology localization

As can be seen, some of the best practices mentioned in Section 4 can also be applied in the ontology localization process. Therefore, in this section we focus on some tasks typical in ontology localization.

- Before localizing an ontology, all the documentation related to the ontology needs to be read thoroughly in order to understand the use and purpose of the ontology in general, as well as the "meaning" of the classes and properties that make up the ontology (By documentation we refer to the ontology specification document. In the case of the ORG ontology, see http://www.w3.org/TR/vocab-org/).
- The most updated version and the latest recommendation of the ontology specification should be examined to check if there is a schema available, aligned with the specification document.
- It is also advisable to look for "normative" translations of the ontology. They may also help in the translation process, especially when translating languages that belong to the same family. As for the ORG ontology, we could rely on the French and Italian translations, as both languages come from Latin, like Spanish.
- In order to understand and correctly interpret the meaning of the ontology entities, it is recommendable to rely on natural language descriptions of those entities, be it in the form of comments, glosses, definitions or usage notes. Without those descriptions we may interpret the meaning and usage of classes and properties wrongly, thus misleading people in their data annotation process.

- Before translating the labels of classes and properties, we should check that ontology entities are accompanied by natural language descriptions. In case there is no description, we should write those descriptions and check with the authors of the ontology that they match the purpose and use of the ontology entities. Otherwise, i.e. if the ontology already includes them, they should be translated before proposing a label for an ontology entity. These steps will help to find the best term as label, since it is more natural and easy to translate a sentence or a body of text than to provide an appropriate label that captures (summarizes) the meaning of that body of text in one word/term or multi-word expression.
- Point 8 in Section 4 would also apply to ontology localization. This means that definitions should not start with a verb as was the case for the English version of the ORG ontology. In this sense, we believe that we should use the accepted format or conventions of the language we are translating into. However, some other conceptual relations can be present when defining a concept, such as Part_of, or Set_of. In these cases, the superordinate could be substituted by the most appropriate collective noun, i.e. fleet: a group of military ships that are controlled by one leader, *flota: conjunto de barcos/aviones/vehículos*.
- As a final recommendation, we suggest that once all ontology entity descriptions and labels have been translated, the resulting text should be reviewed by another translator/linguist, and afterwards by an ontology engineer or user of the ontology in the target language, if possible. In this way, the final product complies with the standard UNE EN-15038:2006.

6. Conclusions

As shown in this paper, the problem of ontology localization and its counterpart, ontology lexicalization is critical when building ontologies. These two processes acquire special relevance in the Linked Data paradigm since the underpinnings of this trend are to link datasets which should be described in ontologies. Our contribution aims to provide some preliminary guidelines for the process of lexicalization and localization in order to help translators, terminologists and users in general in both processes.

Acknowledgements

This work has been supported by the LIDER Project, a support action (SA) funded by the European Commission under the Frame Program 7, and the Spanish project "4V: volumen, velocidad, variedad y validez en la gestión innovadora de datos" (TIN2013-46238-C4-2-R), funded by Ministry of Economy and Competitiveness.

Ontology prefix	Ontology URI
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs	http://www.w3.org/2000/01/rdf-schema#
owl	http://www.w3.org/2002/07/owl#
foaf	http://xmlns.com/foaf/0.1/
gr	http://purl.org/goodrelations/v1#
lemon	http://lemon-model.net/lemon#
skos	http://www.w3.org/2004/02/skos/core#

References

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American, 284(5),* 34-43.

Bizer, C., Heath, T, & Berners-Lee, T. (2009) Linked data - the story so far. Special Issue on Linked Data, *International Journal on Semantic Web and Information Systems (IJSWIS). 5(3)*, 1-22.

Brickley, D., & Guha, R.V. (2014). *RDF Schema 1.1*. Accessible at http://www.w3.org/TR/rdf-schema/

Cimiano, P., Unger, Ch. & McCrae, J. (2014). *Ontology-based interpretation of natural language*. Toronto: Morgan & Claypool.

Espinoza Mejía, M., Montiel-Ponsoda, E., Aguado de Cea, G., & Gómez-Pérez, A. (2012). Ontology localization. In M. C. Suárez-Figueroa, A. Gómez-Pérez, E. Motta, & A. Gangemi (Eds.), *Ontology Engineering in a Networked World* (pp. 171-191). Berlin-Heidelberg: Springer.

Fliedl G., Kop, Ch. & Vöhringer, J. (2007). From owl class and property labels to human understandable natural language. *Natural Language Processing and Information Systems*,156–167.

Gracia, J., Montiel-Ponsoda, E., Cimiano, P., Buitelaar, P., Gómez-Pérez, A., & McCrae, J. (2012). Challenges for the multilingual web of data. *Journal of Web Semantics,* 11, 63-71.

McGuinness D.L. & van Harmelen F. (2004). *OWL Web Ontology Language.* Accessible at http://www.w3.org/TR/owl-features/

Miles, A. & Bechhofer, S. (2008). *SKOS Simple Knowledge Organization System RDF Schema*. Accessible at http://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html

Montiel-Ponsoda, E., Vila Suero, D., Villazón-Terrazas, B., Dunsire, G., Escolano Rodríguez, E. & Gómez-Pérez, A. (2011). Style guidelines for naming and labeling ontologies in *The multilingual web, International Conference on Dublin Core and Metadata*, Dublin, Irland.

Noy N. F. & McGuiness, L. (2001). Ontology development 101: A guide to creating your first ontology. Accesible at http://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Schober, D., Smith, B., Lewis, S. E., Kusnierczyk, W., Lomax, J., Mungall, C., Taylor, C. F., Rocca-Serra, P. & Sansone, S.A. (2009). Survey-based naming conventions for use in obo foundry ontology development. *BMC Bioinformatics, 10: 125*.

UNE EN-15038:2006 *Norma de calidad para servicios de traducción*. AENOR: Madrid.

UNE-ISO 1087-1. (2009). *Trabajos terminológicos. Teoría y aplicación*. AENOR: Madrid.

UNE-ISO 1087-2. (2009) *Trabajos terminológicos. Aplicaciones informáticas*. AENOR: Madrid.