Representing Translations on the Semantic Web

Elena Montiel-Ponsoda, Jorge Gracia, Guadalupe Aguado-de-Cea, and Asunción Gómez-Pérez

Ontology Engineering Group, Dpto. Inteligencia Artificial
Facultad de Informática, Universidad Politécnica de Madrid
Campus de Montegancedo s/n, 28660 Boadilla del Monte, Madrid, Spain
jgracia, emontiel, lupe, asun@fi.upm.es
http://www.oeg-upm.net/

Abstract. The increase of ontologies and data sets published in the Web in languages other than English raises some issues related to the representation of linguistic (multilingual) information in ontologies. Such linguistic descriptions can contribute to the establishment of links between ontologies and data sets described in multiple natural languages in the Linked Open Data cloud. For these reasons, several models have been proposed recently to enable richer linguistic descriptions in ontologies. Among them, we find *lemon*, an RDF ontology-lexicon model that defines specific modules for different types of linguistic descriptions. In this contribution we propose a new module to represent translation relations between lexicons in different natural languages associated to the same ontology or belonging to different ontologies. This module can enable the representation of different types of translation relations, as well as translation metadata such as provenance or the reliability score of translations.

Keywords: multilingual Semantic Web, multilingual Linked Data, *lemon* model, translation relations

1 Introduction

The Linked Open Data [1, 2] initiative has triggered the publication and linking of data sets in the RDF [13] format, contributing in this way to semantically structuring huge amounts of data on the Web. Thanks to the representation format propounded by Linked Data, concepts are connected across resources, breaking down the barriers imposed by data silos, and enabling machines to smartly navigate the Web as a big data set. Currently, more than 250 data sets containing more than 30 billion triples are available in the Linked Open Data (LOD) cloud[1], ranging from domains as far apart as biomedicine, music or geography. Governmental institutions, enterprises and the private sector have realized the benefits and potential of such an initiative and have made their data sets available for linking and exploitation by third parties.

[1] http://www4.wiwiss.fu-berlin.de/lodcloud/state/

The launching phase of the LOD was led by English speaking countries, but in recent years, the LOD cloud has also seen an increase in resources documented in languages other than English. By having a quick look at the CKAN[2] catalogue of data sets, we come across the *data.bnf.fr* data set from the French National Library, the *GeoLinkedData.es* data set of Spanish geographical data, *Rechtspraak.nl* from the Netherlands Council of the Judiciary, or the *FAO geopolitical ontology* with labels in English, French, Spanish, Arabic, Chinese, Russian and Italian.

This proliferation of semantic data described in several natural languages evidences the need for accounting for the linguistic information relative to ontologies and linked data because of several reasons. One of the main reasons is that the linguistic descriptions of these resources help in finding and establishing mappings between concepts and individuals of different ontologies and data sets [22]. Another evident reason is that such descriptions contribute to a better exploitation of the data sets by tasks such as information extraction [19], natural language generation [3], or multilingual data access [7], to mention but a few.

Several formats and annotation properties have been developed in the Semantic Web to represent natural language descriptions associated to ontologies and linked data, such as the *rdfs:label* [13] or *skos:prefLabel* [15] properties. Their limitations have been discussed in several fora [5, 18, 14], and extensions or new models have been proposed in the last years for the representation of linguistic descriptions relative to ontologies and linked data in more principled ways. Some of these models are SKOS-XL [16], LexInfo [5], LIR [18], or the recently appeared *lemon* model [14]. Most of these models also provide some mechanisms to allow for the representation of multilingual descriptions associated to the same ontological representation. However, we argue that explicit relations between descriptions in different languages, i.e., translation relations, as well as translation descriptive metadata, would help in a more efficient exploitation of these multilingual annotations. Moreover, they would also contribute to the establishment of principled links between ontologies and data sets described in multiple natural languages in the LOD cloud.

In this paper, we propose a representation mechanism of translations between labels in different languages associated to ontology terms. To that end, we propose a metamodel in OWL which extends the *lemon* ontology, and which is offered as a module of the *lemon* model. *lemon* is a linguistic model developed in the framework of the Monnet[3] project to represent lexical and terminological descriptions relative to an ontology. The *lemon* extension we propose in this paper enables the representation of translations in a separate linguistic layer, thus leaving the original ontologies or data sources untouched. It also contributes to the linking of ontologies and data sets described in different natural languages in the Web of Data.

The rest of the paper is organized as follows. Section 2 summarizes the mechanisms that some Semantic Web formats or models have for linking linguistic

[2] http://ckan.net/
[3] http://www.monnet-project.eu/

descriptions in several natural languages. In section 3, we analyze the problem of translation relations in the context of the Semantic Web. After that, in section 4, we briefly present the *lemon* model. Thanks to the modular conception of this model, we are now able to propose a translation module, i.e., a module to explicitly represent translations in *lemon*. Section 5 will be devoted to a detailed description of the translation module, and some examples will be provided to illustrate the use of this module. Finally, we conclude the paper in section 6.

2 Related work

As it is well known, RDFS [13] and SKOS [15] rely on limited annotation properties to represent labels or linguistic descriptions associated to ontologies and linked data. They also enable a simple form of multilingual labeling by using language tags to restrict the scope of a label to a particular language (e.g., *skos:prefLabel "bank"@en*). This representation allows for *indirect* or *non-explicit links* between or among multilingual labels, when associated to the same resource in the data set.

Conscious of these limitations, SKOS developers worked on an extension of SKOS called SKOS-XL [16] , that allows to make links explicit between labels associated to the same concept. This extension introduces a *skosxl:Label* class that allows labels to be treated as first-order RDF resources, and a *skosxl:labelRelation* property that provides links between the instances of *skosxl:Label* classes. In this way, we can specialize the *skosxl:labelRelation* into a translation relation and explicitly link *skosxl:Label* instances in different natural languages.

The LIR [18] model also focuses on the representation of links between labels within and across natural languages. This model was created with the purpose of keeping the ontology and the linguistic information independent from each other, so that lexical and terminological properties of labels could be further described (e.g., part-of-speech, gender, terminological variants). The relations provided by LIR to labels within the same natural language have lexical (*hasSynonym*, *hasAntonym*) or terminological nature (*hasVariant*, *hasAbbreviation*, *hasTransliteration*, etc.). And the ones between labels across different natural languages have a translational nature (*hasTranslation* or *hasScientificName*).

Now, the relations provided by the SKOS-XL and LIR models, though being useful for certain applications because of the explicitness of the *hasTranslation* relation between labels in different natural languages, do not allow to account for some aspects of the translation process that may also be relevant for certain applications. For instance, the difference between original and target label. This may be interesting in the case that we have an ontology documented in four natural languages, and we want to specify which labels (or which linguistic descriptions) have been taken as the source in the translation process. Another aspect to be considered could be the type of translation relation existing between labels (we will come back to this in section 3). Moreover, the provenance, i.e., the resource from which translations have been obtained may also be the kind of metadata that enriches the information about translation. Finally, it is

important to account for the adequacy and reliability of the translation in the specific context of the ontology. An extension of these models would be required to represent further translation metadata. However, we have chosen the *lemon* model for this purpose, because its design principles make it specially appropriate for the Web of Data scenario. Firstly, *lemon* introduces a 'well-defined lexical-conceptual' path between linguistic descriptions and ontology elements. Secondly, *lemon* has been designed as a concise RDF model that captures complex linguistic descriptions by dereferencing resources that contain them. And thirdly, it is an extensible and modular model, which allows the use or inclusion of certain modules if so required by the final application. These and other features of the model will be further detailed in section 4.

Finally, we will refer to the *LOD in Translation work*[4], in which a model has been created to describe and retrieve translations in the LOD cloud relying on resources that contain labels in different natural languages. This model takes advantage of multilingual labels associated to resources by means of language tags (as in *rdfs:label "bank"@en*, *rdfs:label "Bank"@de*, *rdfs:label "banco"@es*) and retrieves available translations. Our purpose, on the other hand, is to contribute to the creation of explicit translation links within the same data source and across data sources, so that this and other systems can benefit from the multilingual data in the LOD cloud.

3 Translation relations in the Semantic Web

Ontology localization [21, 8, 6] has been defined as the activity of adapting an ontology to the needs of a particular (linguistic and cultural) community. Methodological guidelines, tools and models have been developed to support the ontology localization activity, which normally results in an ontology in which labels are documented in multiple natural languages, what is the same, a multilingual ontology [6]. Since the different linguistic versions are assumed to be pointing to the same ontology concepts, it could be derived that they are all translations of each other. However, if we have several terms in each language (synonyms or term variants), we may want to unambiguously express which term variant in language A is translation of which term variant in language B. At this point, translation relations acquire significance.

Let us illustrate this with a simple example. In the FAO geopolitical ontology mentioned in the introduction, one ontology term may describe the organization as such and have the labels "Food and Agriculture Organization" and "FAO". Translations of full form and acronym will be provided in the rest of languages, and, ideally, explicit links will be created between the full forms and the acronyms, respectively.

However, translation relations are not always so direct and simple. As claimed in [8, 6], depending on the type of conceptualization represented in the ontology, direct translations in the target language will be available or not. A distinction

[4] http://sites.google.com/site/pierreyvesvandenbussche/apps/lod-in-translation

is made between the so-called *internationalized or standardized conceptualizations*, and conceptualizations more prone to reproduce the vision of the world of a certain community, the so-called, *culturally-influenced domains*. When localizing ontologies of these two types, translation relations may also need to be of different types. To put it in other words, when dealing with internationalized domains, i.e., *technical or specialized domains of knowledge such as engineering or medicine that have standards for processes and descriptions, and whose categorizations usually reflect the common view of different cultures* [17], we may find translations for all terms describing the concepts in the ontology, since the same conceptualization is shared among the languages represented in the ontology. Contrary to that, when localizing ontologies representing culturally-influenced domains, in which the granularity level of some concepts may differ from culture to culture, we may come across mismatches that need to be solved to provide adequate translations. Under this group we include domains such as law, geography or the political and administrative organization of countries, universities, and so on.

Imagine an ontology of financial institutions in Germany. One of the concepts represented in the ontology may be *Sparkasse* (which we could generally translate as *savings bank* in English). However, there may be differences between these concepts concerning business purpose, ownership or governance of the institution. So, maybe, a more adequate translation of *Sparkasse* could be *German savings institution*, although we usually tend to look for the *closest equivalent concept* in the target language and get the term used to refer to it, i.e., *savings bank* in this case. This simple example aims at illustrating the difference between 'literal or documentary translations', and 'functional translations'[5]. The first type usually describes the concept in the target language, because there is no *exact equivalence* in the target language. The second type looks for the *closest equivalence* -though being conscious of the existence of disparities- because it may be convenient for practical reasons. For instance, when aiming at interoperability (at a European or international level), near-equivalents are assumed to match although a complete overlap between them does not exist.

According to this, we make a distinction between *literal translations* and *cultural equivalences*. In the context of the Semantic Web, this distinction may be quite simple to make. The literal translation would be pointing to the same ontology concept, whereas the cultural equivalent would most probably belong to an equivalent ontology documented in the target language. See figure 1 for an illustration of this. Ontology A is an ontology of German credit institutions in which labels have been translated into English, whereas Ontology B conceptualizes the structuring of British credit institutions in English. It would be highly interesting to specify the links between these terms in a multilingual scenario. For these reasons, we claim that further specifications of the translation relation would contribute to envisage a true Multilingual Semantic Web.

[5] Many practitioners and translation theorists agree on this difference and speak about *overt* vs. *covert* translation [11], or *documentary* vs. *instrumental* or *functional* translation [20], respectively.

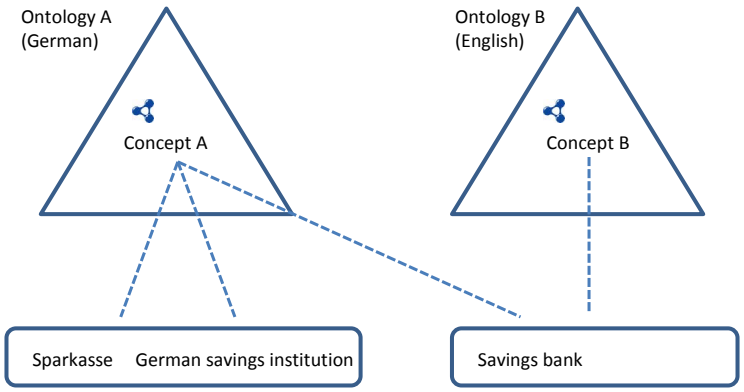

Fig. 1. Oversimplified example of literal translation and cultural equivalence links

4 *lemon*, an interchange model for the Multilingual Semantic Web

The *lemon* model (*le*xicon *m*odel for *on*tologies) [14] is an RDF model of linguistic descriptions that has been designed to a) be published with ontologies, b) extend their lexical layer with as much linguistic information as needed, and c) exchange the resulting lexical resources on the Web. Technical details and usage of the model can be found at http://lexinfo.net/lemon-cookbook.pdf The main features of the model can be summarized as follows:

- Linguistic descriptions are kept separated from the ontology, but their semantics are defined by pointing to the corresponding semantic objects in the ontology (what has been called 'semantics by reference' [4]).
- The model consists of a core set of classes (as described below) and several modules capturing different types of lexical and terminological descriptions.
- Rich lexical and terminological descriptions are grouped into five modules: linguistic properties (part-of-speech, gender, number...), lexical and terminological variation, decompositions of phrase structures (representation of multi-word expressions), syntactic frames and their mappings to the logical predicates in the ontology, and morphological decomposition of lexical forms.
- Linguistic annotations (data categories or linguistic descriptors) are not captured in the model, but have to be specified for each lexicon by dereferencing their URIs as defined in the repositories that contain them (for instance, the ISOcat repository [12]).

The different types of linguistic descriptions captured by the model and its main classes can be seen in figure 2. The core classes of the model are the ones that form the main path between the *Ontology* and the lexical variants represented in the *LexicalEntry* class. The *LexicalSense* class provides a principled link between an ontology concept and its lexical materialization (*LexicalEntry*).

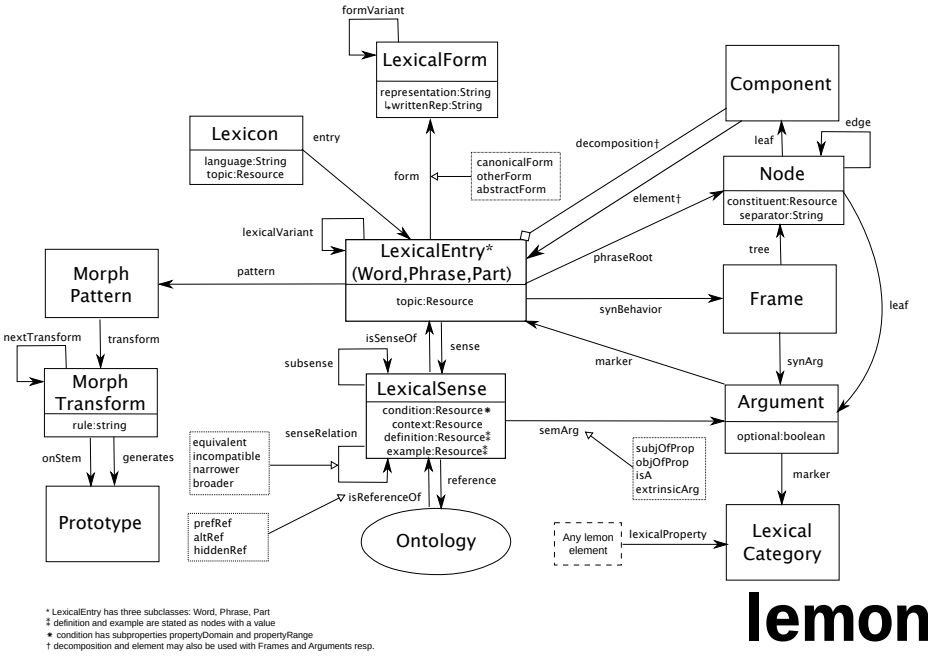

Fig. 2. Core classes and modules of the lemon model

Since 'concepts', as defined in ontologies, and 'lexical entries', as defined in lexicons, cannot be said to overlap [10], the *LexicalSense* class provides the adequate restrictions (usage, context, register, etc.) that make a certain lexical entry appropriate for naming a certain concept in the specific context of the ontology being lexicalized.

LexicalSense is also the class that is foreseen to provide the links between lexical entries within and across languages. Four specializations of this relation are provided: equivalent, incompatible, narrower and broader, as illustrated in figure 2. As the *lemon* model defines one lexicon per language, translation relations could be inferred as lexical entries in different languages would be all pointing to the same ontology reference. However, it is also foreseen to make this type of relation explicit between lexical senses, in the case that, for instance, lexical entries are not pointing to the same ontology reference, but belong to the linguistic descriptions associated to other ontologies.

As such, the translation relation between lexical senses is a powerful mechanism to represent translations. Nevertheless, and as already pointed out in section 1, when dealing with translations, additional properties of the translation relation need to be made explicit, such as reliability score, provenance, or type of translation relation, as already introduced in section 2. In this sense, the flexibility provided by the *lemon* model by means of modules allows us to propose a so-called 'translation module', by reifying a translation relation be-

tween lexical senses into a class. The use of such a module could be exploited by applications that require multilingual ontologies and want to keep track of the relations between the lexical entries in different languages. This information would be very valuable if translations have been automatically generated via an ontology localization system (e.g., LabelTranslator NeOn Toolkit plug-in [9]).

5 *lemon* module for translations

In this section we describe the entities of the translation module in *lemon*[6] and illustrate its use by means of some examples. Figure 3 shows the class diagram of the translation ontology. Some classes are imported from the core of the *lemon* ontology, namely *Lexicon*, *LexicalEntry*, *Form*, and *LexicalSense*.

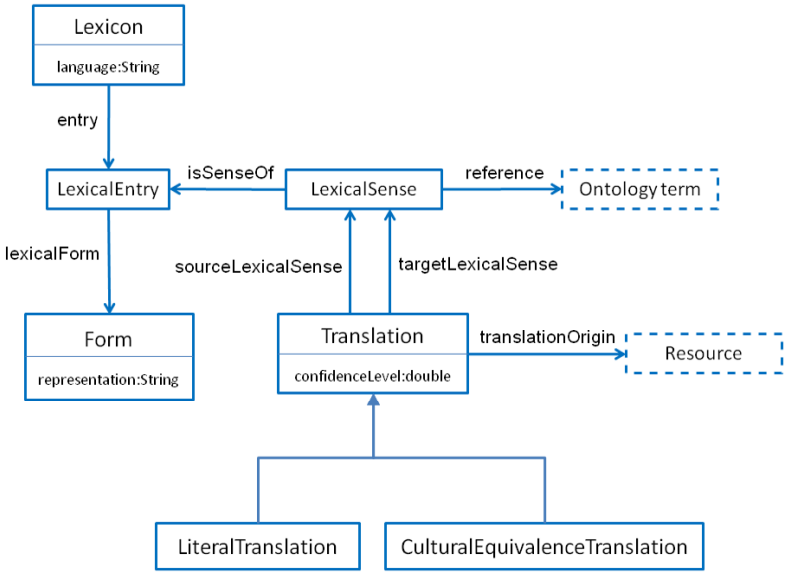

Fig. 3. *lemon* Translation module

– *Translation*. This is the central class of the translation module. It mediates the translation relation between lexical senses, and contains also information that characterizes the translation process, such as a *confidence level*. This confidence level will ultimately depend on the translation tools and translation resources employed to obtain translations. We do not deal here with the algorithms used for its computation, but it will typically combine different features such as probabilities of translation systems, reliability of translations resources, scores of disambiguation methods, etc.

[6] It will be available at http://www.monnet-project.eu/lemon_translation.owl

- *Literal Translation*. It is a subtype of the translation class that corresponds to the idea of literal translations mentioned above.
- *Cultural Equivalence Translation*. A subtype of the translation class that covers translations that are not literal, but close cultural equivalences between the languages considered.
- *Resource*. It represents resources from which translations have been obtained.
- *Lexical Sense*. A sense links a lexical entry to the reference (ontology term) used to represent its meaning.
- *Lexical Entry*. It is a container of the different forms and meanings of a lexeme.
- *Form*. An inflectional form of an entry. It admits several representations (written, phonetic, etc.).
- *Lexicon*. This class represents the whole lexicon. It has a language associated, so it is assumed to be monolingual. Translations will typically connect entries between different monolingual lexicons.

5.1 Examples of use of the *lemon* translation module

In order to illustrate the usage of the translation module, in this section we provide some examples of the financial and politics domains.

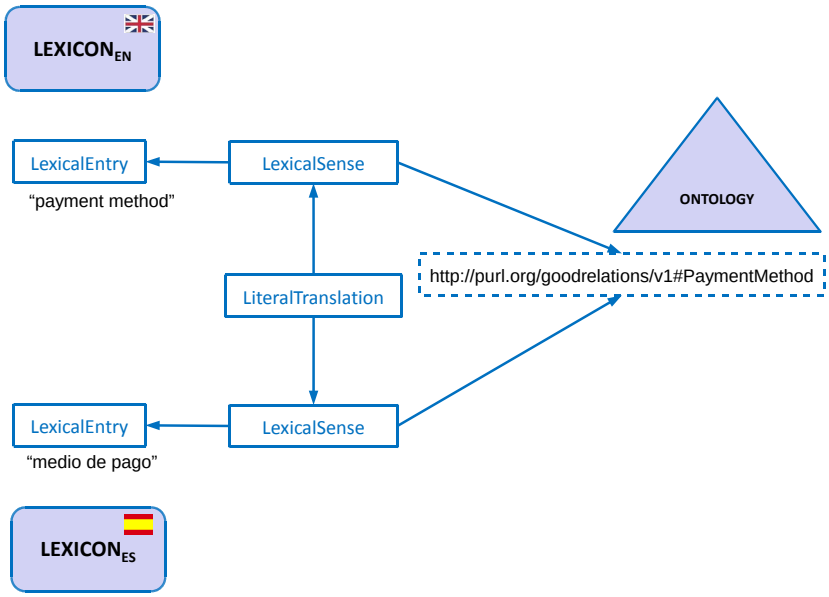

Fig. 4. Example of literal translation

Figure 4 represents an ontology term extracted from the GoodRelations ontology[7]. In *lemon* we would be able to associate as many lexicons in different languages to the ontology as wished. In the figure, we show two lexicons that have been associated to the ontology: one lexicon with English descriptions and the other with Spanish descriptions. Both lexicalize in different languages the same ontology concept, namely, *http://purl.org/goodrelations/v1#PaymentMethod*. Each lexicon contains a lexical entry and a lexical sense representing the ontology concept in each language. The lexical sense belonging to the English lexicon would be the *sourceLexicalSense*, and the one of the Spanish lexicon would be the *targetLexicalSense*, since the ontology was conceived in English. The provenance of the translation would be specified at the *Resource* class. It could be an on-line resource (machine translation service), a lexicon or terminology of the domain, or even a human translator. A confidence value could also be assigned to the translation by means of the *confidenceLevel* property of the *Translation* class. Finally, we would relate these two translations by means of the *LiteralTranslation*, subclass of the *Translation* class. **This would mean that in the specific context of the ontology being lexicalized and localized, the target lexical sense provides a description or literal translation of the term, which is to be used in the context of the original ontology.** It is highly probable that the Spanish translation "medio de pago" is also its cultural equivalent, which would mean that the same concept exists in the Spanish financial system and has been termed as the literal translation. So in this case, both translation relations would be valid.

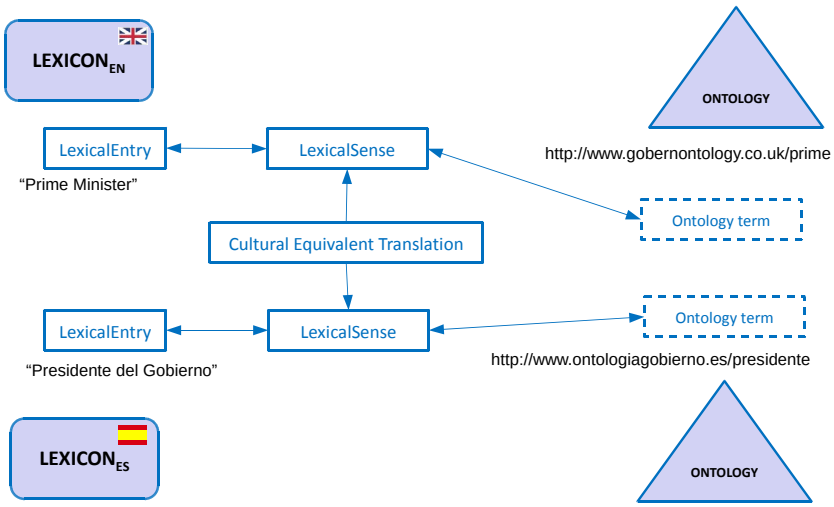

Fig. 5. Example of Cultural Equivalence

[7] http://www.heppnetz.de/ontologies/goodrelations/v1

Now, let us have a look at figure 5. This aims at illustrating cultural equivalents between political systems. Here we have two ontologies, each one representing a different political system, and each one documented in a different natural language. The concept of "Prime Minister" in the British political system and the concept of "Presidente del Gobierno" in the Spanish political system are not exact equivalents, but can be considered the closest equivalents in the respective cultures. This is why we would use the class *CulturalEquivalenceTranslation* to relate the two lexical senses that we assume would belong to two lexicons associated to two different ontologies. **Such a relation would indicate that these two terms are substitutable or translations of each other, when looking for interoperability and referring to (close) equivalents in different languages and cultures, whose extention may not completely overlap.** In this case, we could also include literal translations of each lexical entry in the respective lexicons. In the English lexicon we could include the Spanish lexical entry "Primer Ministro Británico", which would be a literal translation in Spanish. In the same way, we could also add the lexical entry "Spanish President" or "Spanish President of the Government" in the Spanish lexicon. These translations would be related to each other by the *LiteralTranslation* class.

6 Conclusions

The publication of ontologies and data sets in multiple natural languages has raised some issues related to the representation of the linguistic descriptions relative to ontologies. In the context of Linked Data, this takes on more importance since ontologies and data sets described in different natural languages have to be linked to each other. Moreover, such natural language descriptions have proven essential in enabling the exploitation of semantically structured knowledge by language-based tasks. With the purpose of establishing explicit links between the linguistic descriptions associated to ontologies and linked data in several natural languages, in this paper we propose an extension of the *lemon* model to represent translation relations. This translation module allows us to differentiate between *literal* and *cultural equivalence* translations. In addition to that, we can provide metadata relevant to the localization process that may be of great interest when relying on the automatic translation of ontologies.

As future work we plan to carry out some experiments to provide statistics on the impact of such translation relations in the Multilingual Semantic Web, specifically the distinction between literal translations and cultural equivalences. We also aim at investigating the implementation of algorithms that would automatize this process.

Acknowledgments. This work is supported by the EU project Monnet (FP7-248458), and by the Spanish national project BabeLData (TIN2010-17550).

References

1. Bizer, C., Heath, T., Berbers-Lee, T.: Linked data - the story so far. International Journal on Semantic Web and Information Systems. 5, 1–22 (2009)
2. Bizer, C., Heath, T., Idehen, K., Berners-Lee, T.: Linked data on the web (LDOW2008). In: 17th international conference on World Wide Web, pp. 1265–1266 (2008)
3. Bontcheva, K.; The Semantic Web: Research and Applications. In: Generating tailored textual summaries from ontologies Springer, pp. 531–545 (2005)
4. Buitelaar, P.: Ontology-based Semantic Lexicons: Mapping between Terms and Object Descriptions. In: Ontology and the Lexicon, pp. 212–223. Cambridge University Press (2010)
5. Buitelaar, P., Cimiano, P., Haase, P., Sintek, M.: Towards Linguistically Grounded Ontologies In: 6th European Semantic Web Conference (ESWC09), pp. 111–125 (2009)
6. Cimiano, P., Montiel-Ponsoda, E., Buitelaar, P., Espinoza, M., Gómez-Pérez, A.: A Note on Ontology Localization. Journal of Applied Ontology, 5(2), pp. 127–137 (2010)
7. Declerck, T., Krieger, H.-U., Thomas, S. M., Buitelaar, P., ORiain, S., Wunner, T., Maguet, G., McCrae, J., Spohr, D., Montiel-Ponsoda, E.: Ontology-based Multilingual Access to Financial Reports for Sharing Business Knowledge across Europe. In Rooz, J., Ivanyos, J. (eds.) Internal Financial Control Assessment Applying Multilingual Ontology Framework, HVG Press Kft, pp. 67-76 (2010)
8. Espinoza, M., Montiel-Ponsoda, E., Gómez-Pérez, A.: Ontology Localization. In: 5th International Conference on Knowledge Capture (KCAP09), pp.33–40 (2009)
9. Espinoza, M., Gómez-Pérez, A., Mena, E.: Enriching an Ontology with Multilingual Information. In: 5th Annual of the European Semantic Web Conference (ESWC08), pp. 333–347 (2008)
10. Hirst, G.: Ontology and the Lexicon. In: Staab, S., Studer, R. (eds.) Handbook on Ontologies, International Handbooks on Information Systems, Springer, pp. 209–230 (2004)
11. House, J.: A Model for Translation Quality Assessment, Narr, (1977)
12. Kemps-Snijders M., Windhouwer M., Wittenburg P., Wright S.: ISOcat: Corralling data categories in the wild. In: International Conference on Language Resource and Evaluation (LREC) (2008)
13. Manola, F., Miller, E.: RDF Primer. Technical report, W3C Recommendation World Wide Web Consortium (W3C) (2004)
14. McCrae, J., Aguado-de-Cea, G., Buitelaar, P., Cimiano, P., Declerck, T., Gómez-Pérez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., Wunner, T.: Interchanging Lexical Resources in the Semantic Web. Language Resources and Evaluation, in press (2011)
15. Miles, A., Bechhofer, S.: SKOS-Simple Knowledge Organization System Reference, W3C, Retrieved April 11, 2011, from http://www.w3.org/TR/skos-reference/ (2009)
16. Miles, A., Bechhofer, S.: SKOS Simple Knowledge Organization System eXtension for Labels (SKOS-XL) Namespace Document - HTML Variant , W3C, Retrieved June 21, 2011, from http://www.w3.org/TR/skos-reference/skos-xl.html (2009)
17. Montiel-Ponsoda, E.: Multilingualism in Ontologies. Building Patterns and Representation Models. LAP LAMBERT Academic Publishing, Germany (2011)

18. Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., Peters, W.: Enriching Ontologies with Multilingual Information. Journal of Natural Language Engineering, 17 (3), 283–309 (2010)
19. Müller, H.-M., Kenny, E. E., Sternberg, P. W.: Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. PLoS Biol, 2, e309 (2004)
20. Nord, Ch.: Loyalitt statt Treue. Lebende Sprachen, 34, pp. 100–105 (1989)
21. Surez-Figueroa, M. C., Gómez-Pérez, A.: Towards a Glossary of Activities in the Ontology Engineering Field In: 6th Language Resources and Evaluation Conference (LREC08) (2008)
22. Svab-Zamazal, O., Svatek, V.: Analysing Ontological Structures through Name Pattern Tracking. In: EKAW 2008 - 16th International Conference on Knowledge Engineering and Knowledge Management, pp. 213–228 (2008)
23. Volz, J., Bizer, C., Gaedke, M., Kobilarov, G.: Silk A Link Discovery Framework for the Web of Data. In: 2nd Workshop about Linked Data on the Web (LDOW2009) (2009)