Modelling Multilinguality in Ontologies

Elena Montiel-Ponsoda, Guadalupe Aguado de Cea,
Asunción Gómez-Pérez
Ontology Engineering Group
Universidad Politécnica de Madrid
emontiel@delicias.dia.fi.upm.es,
{lupe,asun}@fi.upm.es

Wim Peters
Sheffield Natural Language
Processing Group
University of Sheffield
w.peters@dcs.shef.ac.uk

Abstract

Multilinguality in ontologies has become an impending need for institutions worldwide with valuable linguistic resources in different natural languages. Since most ontologies are developed in one language, obtaining multilingual ontologies implies to *localize* or adapt them to a concrete language and culture community. As the adaptation of the ontology conceptualization demands considerable efforts, we propose to modify the ontology terminological layer, and provide a model called *Linguistic Information Repository* (LIR) that associated to the ontology meta-model allows terminological layer localization.

1 Introduction

Multilinguality in ontologies is nowadays demanded by institutions worldwide with a huge number of resources in different languages. One of these institutions is the FAO[1]. Within the NeOn project[2], the FAO is currently leading a case study on fishery stocks in order to improve the interoperability of its information systems. The FAO, as an international organization with five official languages -English, French, Spanish, Arabic and Chinese- deals with heterogeneous and multilingual linguistic resources with different granularity levels. This scenario is an illustrative example of the need for semantically organizing great amounts of multilingual data. When providing ontologies with multilingual data, one of the activities identified in the NeOn ontology network development process is the Ontology Localization Activity, that consists in *adapting an ontology to a concrete language and culture community*, as defined in (Suárez-Figueroa et al., 2007). In particular, our aim is to obtain multilingual ontologies by localizing its terminological layer (terms or labels that name ontology classes), rather than modifying its conceptualization. Thus, we propose to link ontologies with a linguistic model, called LIR, whose main feature is that it is holistic in the sense that it (1) provides a complete and complementary amount of linguistic data that allows localization of ontology concepts to a specific linguistic and cultural universe, and, (2) provides a unified access to aggregated multilingual data. The model we present in this paper is an enhanced version of the one introduced in (Peters et al., 2007).

2 Related work

The most widespread modality for introducing multilingual data in ontology meta-models consists in using some ontology properties (`rdfs:label` and `rdfs:comment`[3])that define labels and descriptions in natural language of ontology classes. In this system, information is embedded in the ontology. In a similar way, the Simple Knowledge Organization System (SKOS[4]) data model for semantically structuring thesauri, taxonomies, etc., permits the labelling of ontology classes with multilingual strings, and even the establishment of some relations between labels (*preferred* label against the *alternative* one). In any case, however, both modelling modalities restrict the amount of linguistic data that can be included in the ontology, and assume full synonym relations among the multilingual labels associated to one and the same concept.

A further multilingual model is one adopted by the general purpose lexicon EuroWordNet [5]

[1] http://www.fao.org/

[2] http://www.neon-project.org/

[3] www.w3.org/TR/rdf-schema/

[4] http://www.w3.org/2004/02/skos/specs

[5] http://www.illc.uva.nl/EuroWordNet/

Coling 2008: Companion volume – Posters and Demonstrations, pages 67–70
Manchester, August 2008

(EWN). EWN consists of monolingual ontologies, each one reflecting the linguistic and cultural specificities of a certain language, linked to each other through an interlingual set of common concepts that allows building equivalences among ontologies. Although concept equivalences among localized ontologies are reliable and reflect cultural differences, the quantity of linguistic information is also limited to labels and definitions attached to concepts.

Finally, we come to the upward trend in recent research for providing ontologies with linguistic data, which is the association of the ontology meta-model to a linguistic model keeping both separate. The model for representing and organizing the linguistic information can be a data base (as in GENOMA-KB[6] or OncoTerm[7]), or an ontology (as in the case of LingInfo (Buitelaar et al., 2006) or LexOnto (Cimiano et al. 2007)). The main advantage of this modeling modality is that it allows the inclusion of as much linguistic information as wished, as well as the possibility of establishing relations among linguistic elements. Thus, conceptual information is greatly enriched with linguistic data. Additionally, these systems are considered domain independent, and can be linked to any domain ontology.

The differentiating aspect among the mentioned systems is determined by the kind of linguistic classes that make up each model. Depending on the linguistic needs of the end user, some models will be more suitable than others. LingInfo or LexOnto can offer not only multilingual strings to classes and properties of the ontology, but also a deeper morphosyntactic decomposition of linguistic elements, in the case of LingInfo, or a greater focus on syntactic structures by means of *subcategorization frames*, in LexOnto. Our LIR model, however, is more in the line of GENOMA-KB or OncoTerm, in the sense that they follow localization or translational approaches. The main objective of the LIR is to localize a certain ontology category to the linguistic and cultural universe of a certain natural language and to capture translation specificities among languages. Morphosyntactic information is left in the background, although interoperability with ISO standards for representing that sort of information is foreseen. Contrary to GENOMA-KB or OncoTerm, the LIR is represented as an ontology and will be provided with the necessary infrastructure to access external resources for obtaining linguistic data and maintaining links to supplier resources (cf. 5).

2.1 Interoperability with existing standards

Lexical knowledge is expressed in various ways in terminological and linguistic resources. There is a wealth of proposals for enhancing the interoperability of lexical knowledge by encoding it following standard models. As the most important initiatives we take into account two ISO (International Organization for Standardization[8]) standards: The Terminological Markup Framework (TMF[9]) (and the associated TermBase eXchange format; TBX[10]), which captures the underlying structure and representation of computerized terminologies, and the Lexical Markup Framework (LMF) (Francopoulo et al., 2006), an abstract meta-model that provides a common, standardized framework for the construction of computational lexicons.

The LIR model adopts a number of data categories from these standards in order to guarantee interoperability with them. For instance, the notion of *lexical entry* or *lexeme*, in itself a well known central linguistic notion of a unit of form and meaning, has been taken from LMF, whereas the attribute *term type*, which covers representational aspects such as full forms versus abbreviations, has been taken from TMF.

3 Linguistic Information Repository

As shown in Figure 1, the linguistic information captured in the LIR is organized around the `LexicalEntry` class. A lexical entry is *a unit of form and meaning in a certain language* (Saloni et al., 1990). Therefore, it is associated to the classes `Language`, `Lexicalization` and `Sense`. A set of related lexicalizations or term variants shares the same meaning within the specific context of a certain cultural and linguistic universe. E.g., *Food and Agriculture Organization* and *FAO* would be two lexicalizations linked to the same sense. Thanks to the expressiveness of the `hasVariant` relation, it is possible to say that the one is acronym of the other.

The `Language` class at the `LexicalEntry` level allows launching searches in which just those lexical entries related to one natural language are shown to the user, thus displaying the ontology in the selected language.

[6] http://genoma.iula.upf.edu:8080/genoma

[7] http://www.ugr.es/~oncoterm/alpha-index.html

[8] www.iso.org

[9] http://www.loria.fr/projets/TMF/

[10] http://www.lisa.org/standards/tbx/

Sense is considered a *language-specific unit of intensional lexical semantic description* (*ibidem*), which comes to fruition through the Definition class expressed in natural language. Therefore, Sense is an empty class realized by means of the Definition. By keeping senses in the linguistic model independent from ontology concepts, we allow capturing cultural specificities that may slightly differ from the concept expressed in the ontology. Definition has a pointer to the linguistic resource it has been obtained from. In this way reliability and authority of definitions are guaranteed.

Then, Lexicalization is related to its Source or provenance, to a Note class and to a UsageContext class. The Source class aims again at being a pointer to the resource where the information has been extracted from. Note is here linked to Lexicalization, but it could be linked to any other class in the model. It allows the inclusion of supplemental information; e.g., usage specificities of a certain lexicalization within its language system. By linking Note to the Sense or Definition classes we could make explicit possible differences among senses in different languages. The UsageContext class provides information about the behaviour of a certain lexicalization within the language system it belongs to. Finally, lexical semantic equivalences are established among lexical entries within the same language (hasSynonym or hasAntonym), or across languages (hasTranslation). Note that we use the latter label to establish equivalences between lexicalizations in different languages, although it is assumed that words identified as translation equivalents are rarely identical in sense. As Hirst (2004) stated, *more usually they are merely cross-lingual near-synonyms*, but this approach is adopted for the practical reason of providing multilinguality.

The LIR is linked to the OntologyElement class of the OWL meta-model permitting in this way the association of multilingual information to any element of the ontology. Finally, it is left to say that the rationale underlying LIR is not to design a lexicon for different natural languages and then establish links to ontology concepts, but to associate multilingual linguistic knowledge to the conceptual knowledge represented by the ontology. What the LIR does is to associate *word senses* –as defined by Hirst (2004)- in different languages to ontology concepts, although word senses and concepts can not be said to overlap since they are tightly related to the particular vision of a language and its culture, whereas ontology concepts try to capture objects of the real world, and are defined and organized according to expert criteria agreed by consensus.

4 Application of the LIR in NeOn

The LIR has been developed within the NeOn project and is currently being implemented. In order to check its suitability, it was evaluated against the linguistic requirements of the use cases participating in this project (see Note 2): the Spanish pharmaceutical industry, and the Fisheries Stock system of the FAO. Both use cases are working in the development of ontologies for organizing the information they have in several languages. As a consequence, one of the requirements for the NeOn architecture was to support multilingual ontologies.

As already introduced, the LIR not only provides multilingual information to any ontology element, but it also enables unified access to aggregated multilingual data, previously scattered in heterogeneous resources. In that way, it integrates the necessary linguistic information from use case resources and offers a complete and complementary amount of linguistic data.

Regarding the FAO use case, the LIR was evaluated against the recent developed model for the AGROVOC thesaurus, the AGROVOC Concept Server (Liang et al., 2008). This is a concept-based multilingual repository, which, compared to a traditional Knowledge Organization System, allows the representation of more semantics such as specific relationships between concepts and relationships between multilingual lexicalizations. It serves as a pool of agricultural concepts and is a starting point in the development of domain ontologies. The adequacy of the LIR model was positively evaluated against the linguistic requirements of the Concept Server in terms of flexible association of language specific lexicalizations with agricultural domain concepts, and compatibility with TBX.

5 Conclusions and future research

In this contribution we have raised the impending need of international organizations dealing with multilingual information for representing multilinguality in ontologies. In order to obtain multilingual ontologies, we have proposed the association of the ontology meta-model to a linguistic

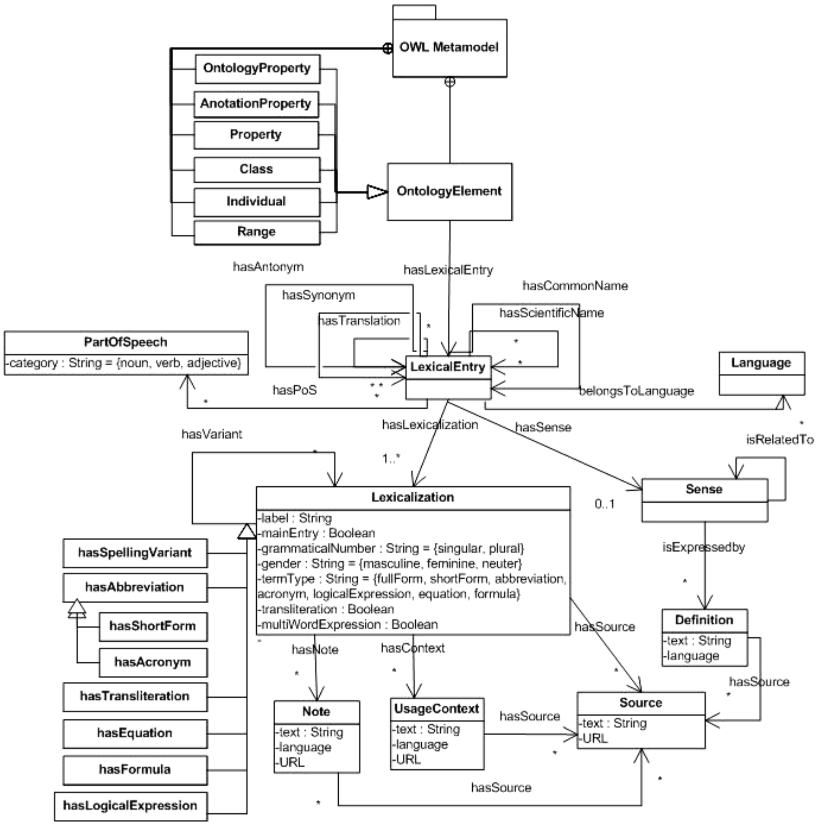

Figure 1. LIR model.

model, the LIR. The LIR has proven to be a holistic linguistic information repository with the following benefits:

- Provision of a complete and complementary set of linguistic elements in each language for localizing ontology elements
- Homogeneous access to linguistic information distributed in heterogeneous resources with different granularity
- Establishment of relations between linguistic elements, and solution to conceptualization mismatches among different cultures

Besides, within NeOn there is a current research regarding the integration of the LIR with the *LabelTranslator* tool (Espinoza et al., 2007), that allows: (1) quick access to external multilingual resources, (2) an automatic translation of the ontology terminological layer, (3) an automatic storage of the resulting multilingual information in the LIR, and (4) convenient editing possibilities for users in distributed environments.

Acknowlegements. This research was supported by the *NeOn* project (FP6-027595)

References

Buitelaar, P., M. Sintek, M. Kiesel. 2006. *A Multilingual/Multimedia Lexicon Model for Ontologies*. In Proc. of ESWC, Budva, Montenegro.

Cimiano, P., P. Haase, M. Herold, M. Mantel, P. Buitelaar. 2007. *LexOnto: A Model for Ontology Lexicons for Ontology-based NLP*. Proc. of OntoLex.

Espinoza, M., A. Gómez-Pérez, and E. Mena. 2008. *Enriching an Ontology with Multilingual Information*. Proc. of ESWC, Tenerife, Spain.

Francopoulo, G, M. George, N. Calzolari, M. Monachini, N. Bel, M. Pet, C. Soria. 2006. *Lexical Markup Framework (LMF)*. Proc. of LREC.

Hirst, G. 2004. Ontology and the Lexicon. In S. Staab, and R. Studer (eds.) *Handbook on Ontologies and Information Systems*. Springer, Berlin.

Liang, A.C., B. Lauser, M. Sini, J. Keizer, S. Katz. 2008. *From AGROVOC to the Agricultural Ontology Service/Concept Server. An OWL model for managing ontologies in the agricultural domain*. In Proc. of the "OWL: Experiences and Directions" Workshop, Mancherter, U.K.

Peters, W., E. Montiel-Ponsoda, G. Aguado de Cea. 2007.*Localizing Ontologies in OWL*.Proc. OntoLex

Saloni, Z., S. Szpakowicz, M. Swidzinski. 1990. The Design of a Universal Basic Dictionary of Contemporary Polish. *International Journal of Lexicography*, vol. 3 no1. Oxford University Press.

Suárez-Figueroa, M.C. (coord.) 2007. *NeOn Development Process and Ontology Life Cycle*. NeOn Project D5.3.1