LexOMV: an OMV extension to capture multilinguality

Elena Montiel-Ponsoda[1], Guadalupe Aguado de Cea[1],
Mari Carmen Suárez-Figueroa[1], Raúl Palma[1], Asunción Gómez-Pérez[1]
and Wim Peters [2]

[OEG]
[1] Ontology Engineering Group, Universidad Politécnica de Madrid
Campus de Montegancedo s/n
28660 Boadilla del Monte, Madrid, Spain
{emontiel, rpalma}@delicias.dia.fi.upm.es {mcsuarez, lupe}@fi.upm.es

[USFD]
[2] Sheffield Natural Language Processing Group, Department of Computer Science
Regent Court, 211 Portobello Street
Sheffield S1 4DP, UK
w.peters@dcs.shef.ac.uk

Abstract. In this paper we aim at proposing metadata for expressing information about multilingual data in ontologies. Nowadays, and thanks to ontology metadata standards as the Ontology Metadata Vocabulary (OMV), ontology developers can have an effective access to all available ontological resources in the web. However, this metadata does not offer information about the linguistic properties of the resource. Our proposal comes to solve this problem by adding an extension to the current OMV that documents about the linguistic or multilingual data contained in the ontology.

Keywords: multilingual ontologies, linguistic representation standards, OMV, LexOMV, ontology localization

1 Introduction

Since the appearance of ontologies as a computational resource in the 90s, their field of application has widened and they are currently considered as the mainstay in the construction of the Semantic Web. Ontologies enable a better understanding between humans and computers, and between computers themselves as they are meant to represent agreed and shared domain knowledge. For this reason, great efforts are devoted by researchers in the Ontology Engineering field to develop models and tools that enhance reusability and sharing of this type of domain knowledge-based applications integrating semantics.

The enormous increase in the development and use of ontological resources by experts in different countries and in the most different domains has shown the need of ontologies undergoing a localization process, which has as result multilingual

ontologies. Interest in multilinguality issues is growing within the scientific community from various perspectives: multilingual information retrieval, query answering systems, machine translation, etc. [1]. OntoSelect [2], an online ontology library that registers ontologies published in the web in RDF(S), DAML and OWL formats, reports the existence of 36 multilingual ontologies out of the total amount of 1420 ontologies that it contains, i.e., 2.5%. Nevertheless, and although this number is expected to rise in the immediate future, multilinguality in ontologies has not been deeply analyzed from a conceptual perspective, and current solutions to localize ontologies have been applied *ad hoc* in each specific case. Moreover, we have been able to state that from those ontologies which contain multilingual labels, most of them lack consistency in the languages which are not the original language of the resource, which is English in most of the cases, i.e., not all concepts in multilingual ontologies have lexicalizations in all the languages the ontology lexicalizations cover.

The Ontology Engineering Group (OEG[2]) at the Universidad Politécnica de Madrid and the Natural Language Processing Group[3] at the University of Sheffield have been working for more than ten years with ontologies and other applications based on ontologies, within the Semantic Web and in other domains. The OEG has worked in the analysis of the theoretical and practical aspects of ontologies, covering the main activities of the ontology lifecycle [3]. The NLP Group at the University of Sheffield has been mainly devoted to the creation of the Gate system [4] which contains tools for ontology editing, and for associating concepts with spans in textual data. Thus, and in view of the emerging need of handling with multilingual knowledge, both groups have carried out a deep survey of the main implications of adding multilinguality to knowledge based applications in the framework of the European project NeOn[4] [7].

In the rest of this paper we present a proposal for providing a set of terms and definitions that will serve the objective of describing the multilingual information contained in ontologies. With this aim, we will offer an overview of the Ontology Metadata Vocabulary (Section 2), on which our extension for the description of multilinguality is based. A general outline of the main ISO standards for modeling multilingual information has been included in Section 3. Then, the qualitative and quantitative step that has to be taken to move from monolingual to multilingual ontologies is as well explained. An extension to the current OMV, called LexOMV that provides information about the linguistic or multilingual data contained in the ontology is presented in section 4, and its appropriateness justified. Finally, Section 5 gives a conclusion to the paper.

2 The Ontology Metadata Vocabulary (OMV)

The OMV is a standard for describing ontologies developed by the joint work of researchers at the Institute AIFB[5] and at the OEG. The main purpose of this research was to create an ontology metadata standard "reflecting the most relevant properties

[2] http://parla.dia.fi.upm.es/oeg/jsp/frames.jsp
[3] http://nlp.shef.ac.uk/
[4] http://www.neon-project.org/web-content/
[5] http://www.aifb.uni-karlsruhe.de/

of ontologies for supporting their reuse" [5]. By means of this standard, ontologies are annotated, which in turn implies the existence of tools and metadata repositories that support the "engineering process, maintenance and distribution of ontologies" (*ibidem*).

As in every process of proposing and approving a standard, the requirements the ontology metadata should comply with were analyzed in the first place. Those requirements took into consideration that the metadata should be "understood" by humans (by usage of natural language concepts) as well as by machines (by usage of Semantic Web languages). It should cover the needs of the majority of ontologies without loosing sight of particular application scenarios in which extensions should also be possible. Furthermore, in order to make the reuse and exchange of ontologies effective and efficient, ontology metadata should provide not only general information of the ontology (e.g. name, description, date of creation, etc.) but also statistical metrics such as the size and structure of the ontology, applicability information (i.e. intended usage or scope), location (e.g. URL), information about the physical representation such as the language and syntax of the formalization, provenance and information about relationships with other resources (e.g. import ontology). Finally, to ensure and facilitate the interoperability of OMV among machines and applications, it is itself represented as an ontology in OWL.

Therefore, and taking all these requirements into account, OMV was designed modularly. It defines a core and allows the creation of various extensions. Some of the main classes and properties of the OMV Core can be observed in Figure 1. As we can see in the figure, besides the main class `Ontology` we have various additional classes (and properties) that allow us to describe the aspects considered during the analysis phase. Therefore, in a general way, we can state that according to this metadata vocabulary ontology we can get, for example, information about the *Person* or *Organization* that created the ontology; the *Type* of ontology, the *Ontology Language* or the *Methodology* used for its development, as well as data about the *Representation Paradigm*, the *Engineering Tool* or the *Task* for which the ontology was originally created.

As already mentioned, the OMV Core covers the majority of available information about ontologies. Nevertheless, OMV can also reflect the specificities of a particular ontology task or application by the development of OMV extension modules. In Section 4, thus, we propose one of such extensions for covering information about linguistic and multilingual data.

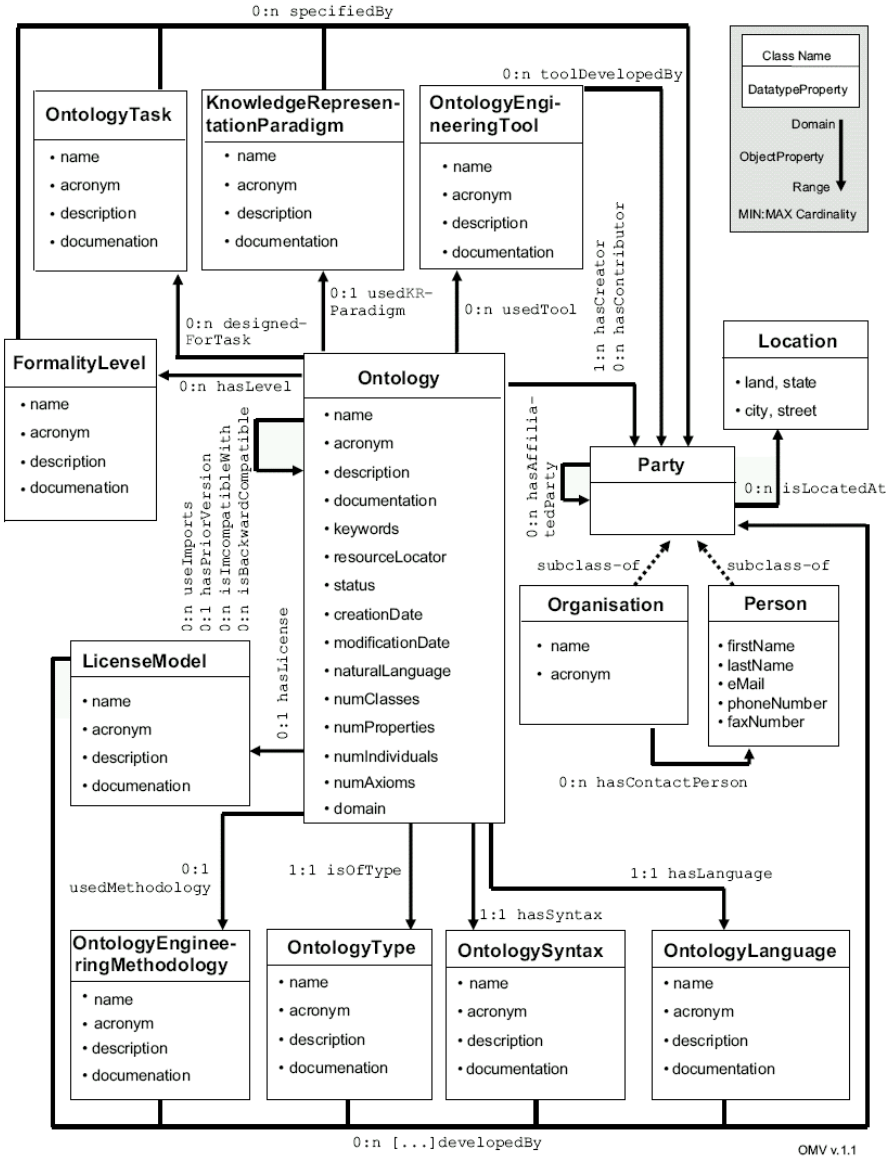

Fig. 1. OMV Core [5]

3 Providing multilinguality to ontologies

Recently, the need for providing multilinguality to ontologies has emerged as one of the main priorities in the Ontology Engineering research. The incremental use of knowledge based systems has raised the need for expressing knowledge in a way that can be understood by people coming from different cultures and speaking different languages, i.e, the need for having to adapt knowledge for specific cultural and language universes. The process of adapting an ontology to a concrete language and culture community has received the name of *ontology localization*, as has been defined in Deliverable 5.3.1 of the NeOn project, in which the main activities to be performed in the ontology network development process are described [6]. Again, within the framework of the NeOn project, the OEG has also carried out a detailed survey of the different localization strategies or approaches that can be used in the localization task of ontologies [8]. However, for the purpose of this paper, we are not so much interested in the steps that have to be carried out in the process of adapting an ontology to a certain language, but in the representational aspects that have to be considered in the ontology meta-model and, more specifically, in the ontology meta-data model.

Until now, research was mainly based on ontology metamodels that represented domain ontologies in only one natural language. However, the trend now is to develop multilingual ontologies that require the representation of multilingual ontology meta-models in which linguistic information is part of the ontology meta-model, as illustrated by Figure 2. Different possibilities in order to integrate multilingual information in the ontology meta-model have been explored by the OEG and the Sheffield NLP Group in the framework of the NeOn project [7]. Those possibilities depend mainly on the quantity of linguistic information that is supposed to be included in the ontology, and in the place where that information is to be stored in the ontology. The main data categories a multilingual ontology should have, as defined in ISO 12620 [9], are *labels* or *lexicalizations* in different natural languages, e.g. in English, French and Spanish. We could also expect to have natural language definitions of concepts in those languages, or other types of linguistic information as context use examples, part-of-speech, etc. All these elements, which are considered linguistic data, should be encoded following standard models in order to guarantee interoperability with existing and proposed standards for the representation and integration of terminological and linguistic knowledge.

For this purpose, the International Organization for Standardization (ISO)[6] has been working on the development of different linguistic representation standards, mainly depending on the purpose of the resource, and the sort and quantity of linguistic information to be represented. It is worth mentioning the existing standardization efforts that have been done to this respect, and which are:

- ISO 16642, *Computer applications in terminology – TMF (Terminological Markup Framework)* [10]
- ISO 24613, *Language Resource Management – LMF: Lexical Markup Framework* [11]
- ISO 12620, *Terminology and other language resources: Data categories* [9]

[6] www.iso.org

Within the ISO, the Technical Committee 37/SC 4 is in charge of the "Language resource management", and Work Group 3 (WG 3) of this committee is currently dealing with the representation of multilingual information, as well as with localization and internationalization issues, among others. For such purposes, this WG has already proposed a standard named MLIF [12] (*Multi Lingual Information Framework*) whose objective is to provide "a common conceptual model and platform allowing interoperability among several translation and localization formats (…)". In this sense, MLIF could be considered a "meta-standard" that allows for the interaction of different representation models in which the designer can select which models to use or combine depending on the linguistic needs of the end resource.

The above described standards are essential in the quantitative and qualitative move from monolingual to multilingual ontologies, because they provide some useful hints of how to represent the basic structure that the linguistic information to be included in the ontology may adopt. Once the linguistic model has been designed, the next step is to associate it to the ontology meta-model, so that it now results in a multilingual ontology meta-model with a common API', as has been represented in Figure 2, that allows a transparent access of the user to the multilingual information. The last point now is to reflect this change at the ontology meta-data level to accomplish the main objective of that level, which is to guarantee a description of the main properties of an ontology, and therefore, its reuse.

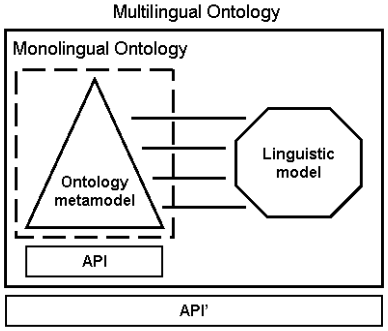

Fig. 2.: Ontology with multilingual information

4 Proposed extension for capturing information about multilinguality in ontologies: LexOMV

According to the OMV philosophy, as already mentioned in section 2, the purpose of the metadata collected in the OMV is to offer ontology users a general description of available ontologies to enable an efficient identification of what they are looking for. In that sense, the foreseen increase of multilingual ontologies needs also to be reflected at this metadata level. Hence, our proposed extension to the OMV Core,

LexOMV, in which we aim at capturing the general linguistic information present in the ontology.

The quantity and quality of linguistic and terminological information available in any ontology in different languages can vary enormously from ontology to ontology. Most of the current available ontologies present multilingual labels for ontology classes or concepts. However, some recently developed knowledge bases that incorporate ontologies (as GENOMA-KB[7] [13] or OncoTerm[8]) have seen the benefits of exploiting ontologies from the terminological or translational perspectives, and include additional linguistic data, i.e. natural language definitions or comments on the usage of the term. In this sense, we should not forget the impending need for international organizations that want to introduce ontologies in their information systems, where several languages are official. The linguistic and semantic part of those ontologies is of great interest for those organizations, and they expect to be able to include as many linguistic data as possible to improve not only indexing and information retrieval tasks, but also translational issues. Therefore, additionally to knowledge structuring, ontologies are required to organize linguistic information.

LexOMV comes to solve this obstacle by informing people searching for multilingual ontologies of the quantity of linguistic and terminological data associated to the ontology. Traditionally, multilingual information has been associated to classes or concepts in the ontology, but classes are not the only ontology elements we can add linguistic data to. Properties (relations and attributes) of classes can also be expressed in different languages.

Therefore, in order to embrace at the metadata level the different possibilities of adding multilingual data to ontologies, we have proposed the following OMV extension, called LexOMV, as represented by Fig. 3. First, we create a new class called `OntologyElement`, so that we are able to make statements separately about the different elements in an ontology. In our example, we follow the DL paradigm, and therefore ontology elements will be *classes*, *properties*, *individuals*, etc. However, it is important to note that our model foresees the description of ontologies following other paradigms. Then, we define a class called `LinguisticElement`, in which we have included the attributes *name* -referring to the name of the linguistic element: *definition*, *lexicalization*, *usage context*, or *part-of-speech*, for example-, and *description* –including an explanation of what is understood under *definition*, *lexicalization*, *usage context*, or *part-of-speech*, or any kind of information we should find in those parts of the ontology-. As it is expected, we also define a class called `NaturalLanguage` with attributes such as *name*, *description* and *ISOcode* that allow us to refer to the different languages as defined by the ISO standard 639 [14]. Finally, we define the class `LinguisticData` in order to associate the multilingual information with the rest of the ontology metadata.

Thus, to express that the piece of linguistic data in question (let us say, *Definition*) is expressed in three languages (e.g. *English*, *Spanish* and *French*) for a certain type of ontology element (e.g. *Class*) in a given ontology, we link the ontology (described in the OMV Core) via the `hasAssociated` relation to the `LinguisticData` class where we integrate all the necessary information using:

[7] http://genoma.iula.upf.edu:8080/genoma/index.jsp
[8] http://www.ugr.es/~oncoterm/

`hasOntologyElement` property to relate the *Class* ontology element, `hasLinguisticElement` property to relate the *Definition* linguistic element and `isExpressedIn` to relate the *English, Spanish* and *French* languages.

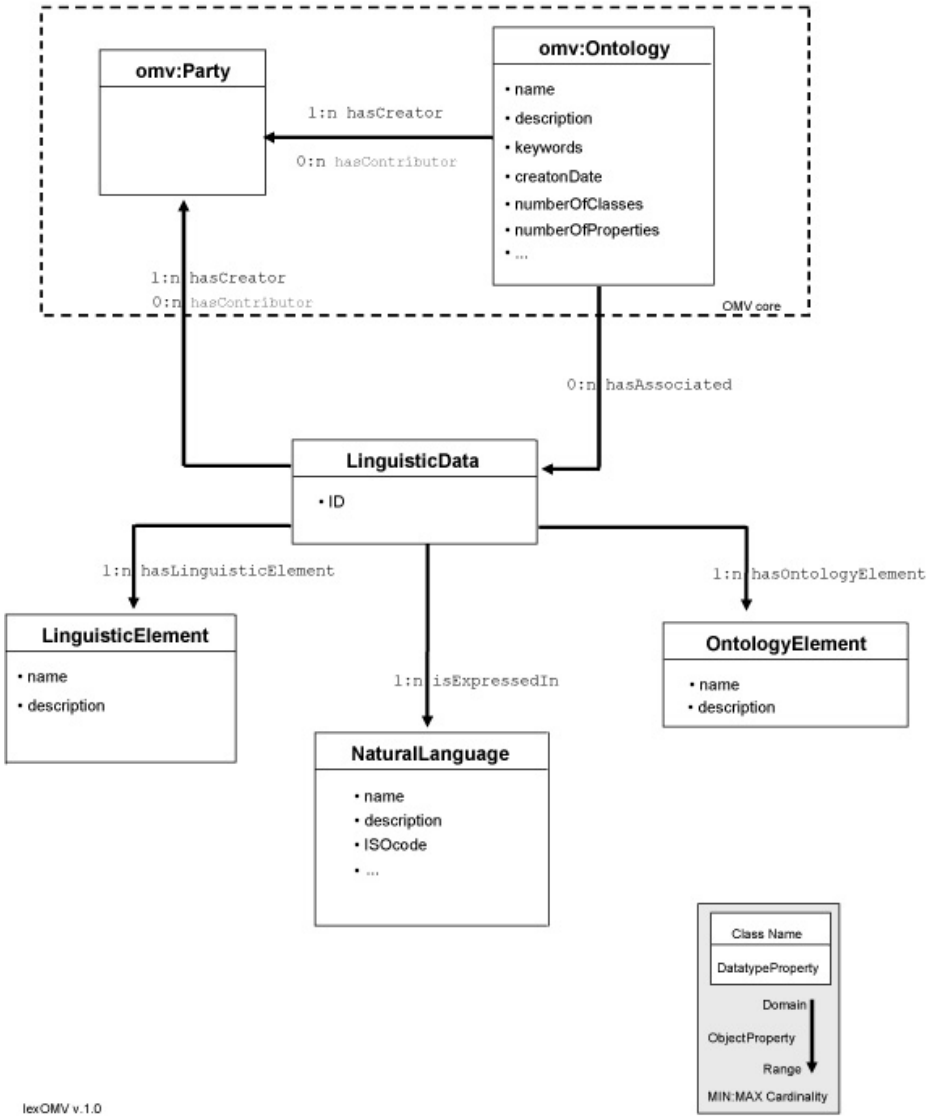

Fig. 3.: LexOMV

The *description* property of the `LinguisticElement` class offers the possibility of defining the quantity and quality of linguistic data provided by the linguistic element in question. For instance, and following with our example of the

`LinguisticElement` *Definition*, it could be defined as "terminological/linguistic sense description" in a certain Linguistic Model, as has been the case in the LIR (Linguistic Information Repository) model designed within the framework of the NeOn project in order to store the linguistic information associated to ontology networks [15]. In the same sense, *Part-of-Speech* could be defined as "the grammatical class of the lexicalization", and so on. By means of that *description* property in natural language, the user is made aware of the scope and coverage of the linguistic information offered by the `LinguisticElement` class.

Thanks to LexOMV, we inform the user searching for ontologies with linguistic information, of the various types of linguistic data included in the ontology in the different languages. Furthermore, our extension allows us to describe who the authors and contributors of those linguistic data were by relating the `LinguisticData` class to the `Party` class of the OMV Core. According to this extension, we can now capture the *author name* or *date of creation* of the ontology next to information like "this ontology includes lexicalizations and definitions of ontology classes in English, Spanish and French". Moreover, and as a result of the general approach of this extension, we are able to capture any kind of linguistic information depending on the Linguistic Model adopted for the ontology.

5 Conclusions

This paper was dedicated to the extension of the standard metadata vocabulary for describing ontologies, OMV, with metadata aimed at accounting for the linguistic data attached to ontologies. The purpose of this extension, that we have renamed LexOMV, is to enable ontology selection on the basis of multilingual and other linguistic information required by the user. This work was motivated by the imminent need of providing ontologies with multilingual data, which is the result of the so-called ontology localization process, and which we have described in Section 3. Ontology localization is one of the activities in the ontology development process receiving most attention in the recent years by the Ontology Engineering community. As a result of this prioritized activity, different research groups as the OEG and the Sheffield NLP group have devoted many efforts to analyze ontology localizing strategies and provide ontology engineers with models for representing multilinguality. At the metadata level in ontology architecture, one missing aspect in the ontology architecture was the possibility of reporting about multilingual data, which we have come to solve with the presented extension of the metadata vocabulary: the LexOMV.

Acknowledgments. The research described in this paper is supported by the European Commission's Sixth Framework Program under the project name: *Lifecycle support for networked ontologies (NeOn)* (FP6-027595). We would like to thank José Ángel Ramos Gargantilla for his feedback.

References

1. Peñas, A., Gonzalo, J.(eds.): Acceso a información multilingüe. Número monográfico de la Revista Iberoamericana de Inteligencia Artificial, Vol. 8. nº 22. (2004)
2. Buitelaar, P., Eigner, T., Declerck, T.,: OntoSelect: A Dynamic Ontology Library with Support for Ontology Selection. In: Proc. of the Demo Session at the International Semantic Web Conference, Hiroshima, Japan (2004) http://www.dfki.de/~paulb/iswc04.demo.OntoSelect.pdf
3. Gómez-Pérez, A., Fernández-López, M., Corcho, O.: Ontological Engineering. Springer, New York (2003)
4. Cunningham, H., Maynard, D., Bontcheva, K., Tablan, V.: GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. ACL'02. Philadelphia (2002)
5. Hartmann, J., Palma, R., Paslaru Bontas, E.: OMV-Ontology Metadata Vocabulary for the Semantic Web. V.2.0. OMV Report (2006) http://ontoware.org/frs/download.php/336/OMV-ReportV2.1.pdf
6. Suárez-Figueroa, M.C. (coordinator): NeOn Development Process and Ontology Life Cycle. NeOn Project Deliverable 5.3.1 (2007)
7. Montiel-Ponsoda, E., Peters, W. (coordinators): Multilingual ontology support. NeOn Project Deliverable 2.4.1 (2007)
8. Aguado de Cea, G. Montiel-Ponsoda, E. Ramos Gargantilla, J.A.: Multilingualidad en una aplicación basada en el conocimiento. In TIMM, monográfico para la revista SEPLN, nº 38, pp. 77-97 (2007)
9. Data Categories. In ISO 12620, Terminology and other language resources.(2003) http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=32347
10. TMF-Terminological Markup Framework. In ISO 16642, Computer applications in terminology. (2003) http://www.loria.fr/projets/TMF/
11. LMF-Lexical Markup Framework . In ISO 24613, Language Resource Management. (2006) http://lirics.loria.fr/doc_pub/LMF%20rev9%2015March2006.pdf
12. Cruz-Lara, S., Bellalem, N., Ducret, J., Kramer, I.: Standarizing the management and the representation of multilingual data: the MultiLingual Information Framework. LR4Trans-III, Genoa, Italy (2006) http://hal.inria.fr/inria-00105653/en/
13. Cabré, M.T., Bach, C., Estopà, R., Feliu, J., Martínez, G., and Vivaldi, J.: The GENOMA-KB project: towards the integration of concepts, terms, textual corpora and entities. LREC (2004)
14. Codes for the representation of names of languages. In ISO 639 (2002) http://www.iso.org/iso/en/commcentre/news/archives/2002/iso639_1.html
15. Peters, W., Montiel-Ponsoda, E., Aguado de Cea, G. Gómez-Pérez, A.: Localizing Ontologies in OWL (2007). Ontolex 2007, Busan, Korea (accepted as full paper).