Rivière or *Fleuve*? Modelling Multilinguality in the Hydrographical Domain

Guadalupe Aguado-de-Cea, Asunción Gómez-Pérez,
Elena Montiel-Ponsoda, and Luis M. Vilches-Blázquez
Ontology Engineering Group
Dpto. de Inteligencia Artificial
Facultad de Informática, Universidad Politécnica de Madrid
28660, Boadilla del Monte, Madrid, Spain
{lupe, asun, emontiel, lmvilches}@fi.upm.es

ABSTRACT
The need for interoperability among geospatial resources in different natural languages evidences the difficulties to cope with domain representations highly dependent of the culture in which they have been conceived. In this paper we characterize the problem of representing cultural discrepancies in ontologies. We argue that such differences can be accounted for at the ontology terminological layer by means of external elaborated models of linguistic information associated to ontologies. With the aim of showing how external models can cater for cultural discrepancies, we compare two versions of an ontology of the hydrographical domain: *hydrOntology*. The first version makes use of the labeling system supported by RDF(S) and OWL to include multilingual linguistic information in the ontology. The second version relies on the *Linguistic Information Repository* model (LIR) to associate structured multilingual information to ontology concepts. In this paper we propose an extension to the LIR to better capture linguistic and cultural specificities within and across languages.

Categories and Subject Descriptors
I.2.4 [**Artificial Intelligence**]: Knowledge Representation Formalisms and Methods [semantic networks]

General Terms
Documentation, Design

Keywords
multilingual ontologies, hydrographical domain, LIR, ontology localization

1. INTRODUCTION
The symbiosis between ontologies and natural language has proven more and more relevant on the light of the growing interest and use of Semantic Web technologies. Ontologies that are well-documented in a natural language not only provide humans with a better understanding of the world model they represent, but also a better exploitation by the systems that may use them. This "grounding in natural language" is believed to provide improvements in tasks such as ontology-based information extraction, ontology learning and population from text, or ontology verbalization, as pointed out in [4].

Nowadays, there is a growing demand for ontology-based applications that need to interact with information in different natural languages, i.e., with multilingual information. This is the case of numerous international organizations currently introducing semantic technologies in their information systems, such as the Food and Agriculture Organization or the World Health Organization, to mention just a few. Such organizations have to manage information and resources available in more than a dozen of different natural languages, and have to customize the information they produce to a similar number of linguistic communities.

In the present research, we are concerned with a further use case: the geospatial information. The importance of multilingualism in this domain lies in the need for interoperability among multiple geospatial resources in different languages, and a flexible human interaction with multilingual information. For many years, geospatial information producers have focused on the collection of data in one or different languages without considering the interoperability level among them. If we take as example the case of Spain, data have been collected from multiple producers at different levels (national, regional and local), and in the different languages that are official in the country (Spanish, Catalan, Basque and Galician), but there have been no efforts to make these data interoperable.

Today, the widespread use of geospatial information and the globalization phenomenon have brought about a radical shift in the conception of this information. In this context, multilingualism has reached a pre-eminent position in the international scene. The rapid emergence of international projects such as EuroGeoNames[1] confirms this trend. The main goal of EuroGeoNames is to implement an interoperable internet service that will provide access to the official, multilingual geographical name data held at national level and make them available at European label.

Ideally, the "meaning" expressed by ontologies would provide the "glue" between geospatial communities [22] by capturing their knowledge and facilitating the alignment of heterogeneous and multilingual elements. However, this still remains an open issue because of the cultural and subjective discrepancies in the representation of geospatial information. This domain is a good

[1] http://www.eurogeographics.org/eurogeonames

Other international projects in a similar line of research are: eSDI-NET+ (http://www.esdinetplus.eu) or GIS4EU (http://www.gis4eu.org/)

exponent of what has been called *culturally-dependant domains* [8] that is, domains in which their categorizations tend to reflect the particularities of a certain culture. The geospatial domain has to do with the most direct experiences of humans with their environment, and it has, therefore, a very strong relation with how a certain community perceives and interacts with a natural phenomenon. A good example of these experiences can be found in [13]. This is inevitably reproduced in the different viewpoints and granularity levels represented by conceptualizations in this domain, which are, in its turn, reflected in the language.

However true that may be, we believe that interoperability is still possible by assuming a trade-off between what is represented in the ontology and what is captured in the ontology terminological (or lexical) layer [2]. Up to now, the representation of multilingualism in ontologies has not been a priority [1], and very few efforts have been devoted to the representation of linguistic information in ontologies, let alone multilingual information. We believe that a sound lexical (and terminological) model independent from the ontology that could capture cultural discrepancies, would pave the way for solving this problem.

In this paper, our purpose is to show how such an external and portable model created to associate lexical and terminological information to ontologies may account for categorization mismatches among cultures. This is the purpose of the *Linguistic Information Repository* (LIR) [15][19], a model created to capture specific variants of terms within and across languages. With the aim of showing this, we will compare the functionalities offered by two representation modalities to link linguistic and multilingual information with ontologies: the labelling system of RDF(S) and OWL vs. the LIR model. This comparison will be done on the basis of an ontology of the hydrographical domain: *hydrOntology*. Additionally, an extension of the LIR model to better account for categorization mismatches among cultures will be proposed.

The rest of the paper is structured as follows. In section 2 we present the state of the art on formalisms and models to represent linguistic information in ontologies. Then, in section 3 we try to characterize the problem of conceptual mismatches or discrepancies among conceptualizations in multilingual knowledge resources. Section 3 is devoted to a brief description of *hydrOntology*. The inclusion of linguistic information in the ontology by means of the RDF(S) labels is described in section 4. Then, the LIR model is presented in section 5, and its instantiation with the linguistic information related to *hydrOntology* is detailed in section 6. By describing the two versions of *hydrOntology*, we aim at showing the main benefits and drawbacks of each modelling modality. Finally, we conclude the paper in section 7.

2. The linguistic-ontology interface

Most of the ontologies available nowadays in the Web are documented in English, i.e., the human-readable information associated to ontology classes and properties consists of terms and glosses in English. Most of these ontologies, not to say all of them, make use of the *rdfs:label* and *rdfs:comment* properties of the RDF Schema vocabulary, a recommendation of the W3C Consortium to provide "a human-readable version of a resource's name"[3].

It is also specified that labels can be annotated using the "language tagging" facility of RDF literals[4], which permits to indicate the natural language used in a certain information object. The RDF(S) properties can be complemented by Dublin Core metadata [5] that have been created to describe resources of information systems. Examples of the Dublin Core Metadata elements are: title, creator, subject or description. Since it is possible to attach as many metadata as wished, this has been used to associate the same metadata in different natural languages to obtain an ontology documented in different natural languages, in other words, to obtain a multilingual ontology. This is precisely one of the main advantages of this representation modality, namely, associating as much information in different languages as wished.

However, we identify several drawbacks for an appropriate exploitation of the resulting multilingual ontologies:

(1) All annotations are referred to the ontology element they are attached to, but it is not possible to define any relation among the linguistic annotations themselves. This results in a bunch of unrelated data whose motivation is difficult to understand even for a human user.

(2) When different labels in the same language are attached to the same ontology element, absolute synonym or exact equivalence is assumed among the labels. As reported in [6] "identical meaning" among linguistic synonyms is rarely the case. It could be argued that in technical or specialized domains, absolute synonymy exists, but even in those domains, labels usually differ in "denotation, connotation, implicature, emphasis or register" [5], what sometimes is reflected in the subcategorization frames they select (syntactic arguments they co-occur with). We will try to illustrate this in section 6.

(3) A similar situation arises when labels in different languages are attached to the same ontology element. In some cases, they will share the common meaning represented by the ontology element (Figure 1). However, the problem appears when a language understands a certain concept with a different granularity level to the one represented by the ontology concept, as illustrated in Figure 2 and Figure 3. In this case, if more fine-grained equivalents exist in one of the languages represented by several labels, it will be interesting to make those differences explicit for a suitable treatment of multilinguality.

(4) Finally, scalability issues should also be mentioned. If only a couple of languages are involved and not much linguistic information is needed, the RDF(S) properties can suffice. But if a higher number of languages are required, as seems to be the trend in the current demand, the linguistic information will become unmanageable.

On the light of the drawbacks outlined, additional approaches have been proposed to connect linguistic and ontological information. In this sense, we will first refer to the Linguistic Watermark initiative [17]. The Linguistic Watermark is a framework or metamodel for describing linguistic resources and

[2] In [3] the *terminological layer* in an ontology is defined as the terms or labels selected to name ontology elements.

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

[4] http://www.isi.edu/in-notes/rfc3066.txt

[5] http://dublincore.org

using their content to "enrich and document ontological objects". The authors already propose a description for WordNet and a set of dictionaries called DICT. Their idea would be to directly import the linguistic information contained in those resources and integrate it in the ontology. However, it seems as if the reused information is included in the ontology by making use of the RDF(S) properties, and this shows the same disadvantages presented above. This approach is technologically supported by the OntoLing Protégé plugin[6].

A further effort to associate linguistic information to ontologies is represented by the LexInfo [4] model. This model is more in line with what we propose in this paper to enrich ontologies with a linguistic model that is kept separated from the ontology. LexInfo is a joint model that brings together two previous models LingInfo and LexOnto, and builds on the Lexical Markup Framework or LMF, an ISO standard created to represent the linguistic information in computational lexicons. As already mentioned, LexInfo offers an independent portable model that is to be published with arbitrary domain ontologies. LexInfo combines the representation of deep morphological and syntactic structures (segments, head, modifiers), as contained in the LingInfo model, with linguistic predicate-argument structures (subcategorization frames) for predicative elements such as verbs, as captured by LexOnto. Since its main objective is to provide an elaborate model to increase the expressivity of ontological objects in a certain language, it cares less for multilingual aspects and categorization discrepancies among languages.

Finally, we will briefly mention the Simple Knowledge Organization System (SKOS) [12], a model to represent the concept schema of thesauri in RDF(S) and OWL. This model also accounts for the representation of multilingual terms, but does not offer a complex machinery to deal with cultural discrepancies. As it has not been created with the purpose of associating linguistic information to ontologies, the semantic relations captured in the model are limited to hierarchical and associative relations among concepts.

3. Characterization of the multilingual representation problem

The reconciliation of different representations (within the same natural language) can be solved by establishing mappings among those representations. When facing representations in different languages, the mapping process results in a multilingual system. A collection of mapping approaches with monolingual and multilingual resources can be found in [9]. Our approach to tackle multilinguality, however, takes as a starting point one conceptualization to which information in different languages is attached. From the development viewpoint, reusing an existing conceptualization in the domain to transform it into a multilingual resource that can be shared among different speaking communities demands less time and efforts than having to conceptualize the same domain from scratch in each natural language, and then find the mappings or correspondences among concepts. Both approaches have to deal with the differences in conceptualizations that each culture makes. In the mapping approach, it is the mapping itself the one that establishes the equivalence links among ontologies, whereas in the second option, this can be solved at the terminological layer or by modifying the conceptualization (for a detailed analysis of modeling modalities to represent multilinguality in knowledge-based systems, see [1]).

To the best of our knowledge, the most recurrent conceptual discrepancies could be systematically classified as follows:

(a) 1:1 or exact equivalence (as illustrated in Figure 1)

(b) n:1 subsumption relation (isSubsumedBy) (illustrated in Figure 2)

(c) 1:n subsumption relation (subsumes) (represented by Figure 3)

In case (a) both conceptualizations or world views share the same structure and the same granularity level. This is normally reflected in the language by means of a word or term that designates that concept. In the situation represented by (b), the original conceptualization (the one belonging to the English language) makes a more fine grained distinction of a certain reality that does not correlate with the granularity level in the target representation of the same reality. In that case, the target concept is slightly more general, and it could be understood as encompassing the *n* concepts in the original conceptualization. This results in two terms in the English language, for instance, to designate those two concepts, whereas in the target culture, only one term is available. The last case (c) depicts the same situation as in (b) but exactly the other way round.

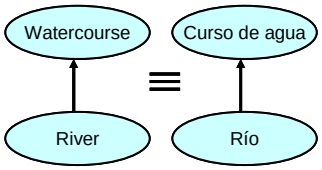

Figure 1. 1:1 or exact equivalence between conceptualizations

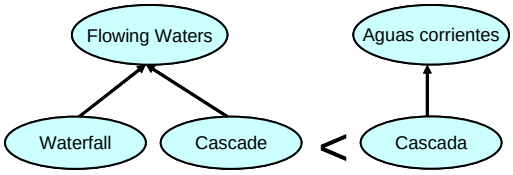

Figure 2. n:1 subsumption relation

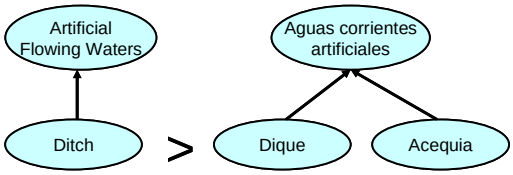

Figure 3. 1:n subsumption relation

However, if our objective is to rely on one ontology to "glue" the different conceptualizations of reality that cultures make, we will need to assume a trade-off between what is represented in the ontology and what is left out, so that every culture can feel that conceptualization as its own, and can meet its representation needs.

Coming back to case (c), if we agree on representing the view of the English culture in the ontology, we will be missing the granularity level of the Spanish world view. We think that those cultural discrepancies could still be reported at the terminological layer of the ontology. A further option would be to integrate the granularity level of the target culture in the common ontology,

[6] http://art.uniroma2.it/software/OntoLing/

but, here again, a certain compromise would be necessary. However, if there are more than two or three cultures and languages involved in the multilingual ontology the suggested option will not be an optimal one. In that case, one possible solution could be to include specific language modules in the ontology, and support different linearizations or visualizations of the same ontology according to the language selected. To the best of our knowledge, currently there is no system to support this latter option. Therefore, our proposal is to account for those categorization mismatches in an elaborated model of lexical and terminological information, separated from the ontology.

4. *hydrOntology*: an ontology of the hydrographical domain

hydrOntology [24] is an ontology in OWL that follows a top-down development approach. Its main goal is to harmonize heterogeneous information sources coming from several cartographic agencies and other international resources. Initially, this ontology was created as a local ontology that established mappings between different data sources (feature catalogues, gazetteers, etc.) of the Spanish National Geographic Institute (IGN-E). Its purpose was to serve as a harmonization framework among Spanish cartographic producers. Later, the ontology has evolved into a global domain ontology and it attempts to cover most of the concepts of the hydrographical domain.

hydrOntology has been developed according to the ontology design principles proposed by [10] and [2]. Some of its most important characteristics are that the concept names (classes) are sufficiently explanatory and are correctly written. Thus each class tries to group only one concept and, therefore, classes in brackets and/or with links ("and", "or") are avoided. According to certain naming conventions, each class is written with a capital letter at the beginning of each word, while object and data properties are written with lower case letters.

In order to develop this ontology following a top-down approach, different knowledge models (feature catalogues of the IGN-E, the Water Framework European Directive, the Alexandria Digital Library, the UNESCO Thesaurus, Getty Thesaurus, GeoNames, FACC codes, EuroGlobalMap, EuroRegionalMap, EuroGeonames, several Spanish Gazetteers and many others) have been consulted; additionally, some integration issues related to geographic information and several structuring criteria [25] have been considered. The aim was to cover most of the existing GI sources and build an exhaustive global domain ontology. For this reason, the ontology contains one hundred and fifty (150) relevant concepts related to hydrography (e.g. river, reservoir, lake, channel, and others), 34 object properties, 66 data properties and 256 axioms.

Currently, the *hydrOntology* ontology is available in two versions. The first one in Prótége makes use of the RDF(S) labeling model to document the ontology in natural language. In a subsequent stage, the ontology was associated to the Linguistic Information Repository (LIR) model, currently supported in the NeOn Toolkit. The first version of the ontology is available in Spanish and English, whereas in the second version two more languages were added: French and Catalan, as will be reported in section 6.

Regarding the first version of the ontology, *hydrOntology* was originally developed in Spanish, and therefore, the *labels* given to the concepts in the original ontology were in Spanish. Later on, English *labels* were also related to ontology concepts, and the language of those labels was specified by means of language tags. Definitions or glosses describing the concepts were also included in Spanish and English, if available, by making use of the *comment* property. Finally, one metadata element of Dublin Core (*source*) and one additional annotation (*provenance*) were used to report about the resources from which the different definitions (*comments*) and *labels* had been obtained, respectively. It must be noted that the process of documentation was not systematically carried out for different reasons, and not all types of annotations are available for every concept.

A snapshot of the class hierarchy of *hydrOntology* in the Protégé ontology editor can be seen in Figure 4. The concept *Río* (River) has been chosen for illustration. It has nine annotations related to it: three *provenance* annotations, two *comment* annotations, three *label* annotations, and one *source* annotation. As already reported, the *provenance* annotation gives information about the linguistic resources (glossaries, thesauri, dictionaries, etc.) labels have been obtained from. Since there are no mechanisms for relating the *label* (e.g. River) with its source of *provenance* (e.g. Water Framework Directive), the authors have decided to include the *label* in the provenance text for the sake of clarity (e.g.; "River – Water Framework Directive. European Union"@en).

Two *comments* are included, one in Spanish, and one in English, though no relation to any of the *labels* is given. Finally, three *label* annotations are given: two in Spanish (in addition to the one given in the URI, i.e., *Río*) and one in English. The two additional labels are *Curso de agua principal* (Main Watercourse), and *Curso fluvial* (Watercourse). According to the authors, the main difference among the three synonyms is the discourse register. The label *Río* is the general word, and would appear in general documents, whereas the other two additional labels would only come up in technical documentation managed by experts in the domain. It is worth noting that such fine-grained aspects could be relevant for certain indexing or information extraction tasks, but cannot be made explicit in the RDF(S) labeling model.

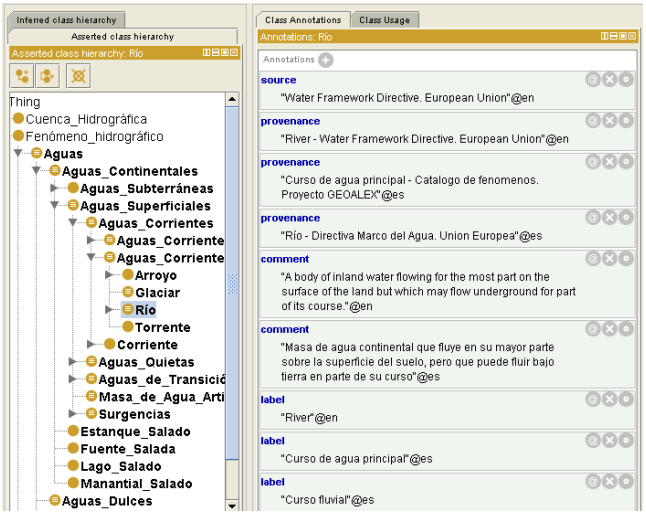

Figure 4. Snapshot of *hydrOntology* and the linguistic information associated to the *Río* ontology concept

Regarding the English translation, *River*, it is not possible to know to which of the Spanish labels is related to or is translation of. *River* is considered to be in a complete equivalence relation with *Río*, which would be appropriate in this case, but it is rarely the case, as explained in section 2. However, the RDF(S) labeling model does not offer any means to report about those cultural

differences that, more often than not, occur between two languages.

Because of these deficiencies in the representation of multilinguality in ontologies in OWL, and with the aim of giving response to the increasing demand for multilingual ontologies, the *Linguistic Information Repository* (LIR) model was developed. In the next section, we present the LIR model and how it aims at solving some of the representation problems identified so far.

5. LIR, a model for structuring the linguistic information associated to ontologies

The *Linguistic Information Repository* or LIR is a proprietary model expected to be published and used with domain ontologies. In itself, it has also been implemented as an ontology in OWL. Its main purpose is not to provide a model for a lexicon of a language, but to cover a subset of linguistic description elements that account for the linguistic realization of a domain ontology in different natural languages. A complete description of the current version of the LIR can be found in [14].

The lexical and terminological information captured in the LIR is organized around the Lexical Entry class. Lexical Entry is considered a union of word form (Lexicalization) and meaning (Sense). This ground structure has been inspired by the Lexical Markup Framework (LMF). The compliance with this standard is important for two main reasons: (a) links to lexicons modeled according to this standard can be established, and (b) the LIR can be flexibly extended with modular extensions of the LMF (or standard-compliant) modelling specific linguistic aspects, such as deep morphology or syntax, not dealt by LIR in its present stage. For more details on the interoperability of the LIR with further standards see [18].

The rest of the classes that make up the LIR are Language, Definition, Source, Note and Usage Context (see Figure 5). These can be linked to the Lexicalization and Sense classes. Each Lexicalization is associated to one Sense. The Sense class represents the meaning of the ontology concept in a given language. It has been modelled as an empty class because its purpose is to guarantee interoperability with other standards. The meaning of the concept in a certain language (which may not completely overlap with the formal description of the concept in the ontology) is "materialized" in the Definition class, i.e., is expressed in natural language. The Usage Context gives us information about how a word behaves syntactically in a certain language by means of examples. Source information can be attached to any class in the model (Lexicalization, Definition, etc.), and, finally, the Note class has been meant to include any information about language specificities, connotations, style, register, etc., and can be related to any class. By determining the Language of a Lexical Entry, we can ask the system to display only the linguistic information associated to the ontology belonging to a given language.

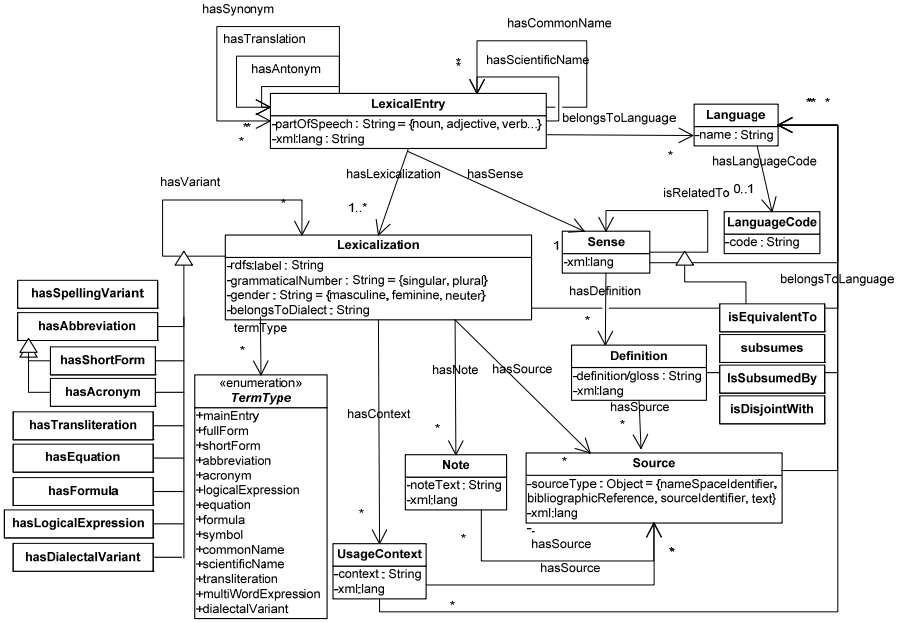

Figure 5. Overview of the LIR model with extensions to the isRelatedTo relation

Thanks to this set of linguistic descriptions, the LIR is capable of managing lexicalizations within one language, and their translations to other languages. Relations of synonymy can be expressed among lexicalizations in the same language, and the preferred lexicalization can be determined (main Entry), as well as other term variant relations (such as Acronym, Multi Word Expression or Scientific Name). Finally, relations of translation equivalence can be established among lexicalizations in different languages.

However, as we stated previously, more often than not lexicalizations in different languages are not exact equivalents, because the senses they represent do not completely overlap in their intensional and/or extensional descriptions. In order to account for cultural and linguistic specificities of languages, we propose an extension of the LIR to allow declaring semantic relations among the senses (Sense) of lexicalizations within and across languages. The semantic relations identified with this purpose are: equivalence (isEquivalentTo), subsumption (subsumes or isSubsumedBy), or disjointness (isDisjointWith).

So, the relation *isRelatedTo* that currently links senses (Sense) in the model is further specified.

6. Modeling multilinguality in *hydrOntology* with LIR

The current version of the LIR is supported by the LabelTranslator system[7], a plug-in of the NeOn Toolkit[8].As soon as an ontology is imported in the NeOn Toolkit, the whole set of classes captured in the LIR is automatically associated to each Ontology Element, specifically, to ontology classes and properties, by means of the relation "has Lexical Entry". In this way, the rest of linguistic classes organized around the Lexical Entry class are linked to an ontology element.

LabelTranslator [8] has been created for automating the process of ontology localization. Ontology Localization consists in adapting an ontology to the needs of a concrete linguistic and cultural community, as defined in [20]. Currently, the languages supported by the plug-in are Spanish, English and German. Once translations are obtained for the labels of the original ontology, they are stored in the LIR model. However, if the system does not support the language combination we are interested in, we can still use it to take advantage of the LIR API implemented in the NeOn Toolkit. In this sense, we can manually introduce the linguistic information necessary for our purposes.

As already mentioned, in the second version of *hydrOntology*, our purposes were to enrich the ontology in French and Catalan. With this aim, we imported the ontology originally documented in Spanish in the NeOn Toolkit, and automatically, all the linguistic classes of the LIR were associated to the concepts and properties in the ontology. The linguistic information associated to the URI of ontology concepts and properties in the original ontology automatically instantiated the LIR classes, i.e., a Lexical Entry was created for each ontology element, with its corresponding identifier (e.g., LexicalEntry-1), the Language of the label was identified and instantiated (e.g., Spanish), and a Lexicalization related to the Lexical Entry was also instantiated with the label in Spanish (e.g., Río). The rest of the linguistic information contained in the original ontology was not imported by the tool, and this fact was reported to the developers.

The next step was to manually introduce the labels in English, already available in the Protégé version of *hydrOntology*,. Since not all the concepts had been originally translated into English, we decided to make use of the LabelTranslator system to semi-automatically obtain translations for the original labels. The process was carried out in a semi-automatic way, and the translation candidates returned by LabelTranslator were evaluated by a domain expert. Since the purpose of this paper is not to evaluate the LabelTranslation plug-in, we will only refer to some of the results by way of example. To obtain more information about the experimental evaluation conducted with this tool, we refer to [8]. A table summarizing the results has been included in the Annex section[9] (see Table 1).

Then, the following step was the enrichment of the ontology with information in French and Catalan. Since these languages are not supported by LabelTranslator, we resorted to authoritative terminological resources in the domain [10], and manually introduced the information in the LIR by means of the LIR API (see Figure 6). For the sake of comparison, we will illustrate the results by taking the concept *River* as example, as in the case of the Protégé version of *hydrOntology*.

As shown in Figure 6, seven Lexical Entries with Part of Speech *noun* were associated to the concept *Río*: three in Spanish, one in English, one in Catalan and two in French. By clicking on each Lexical Entry we are able to visualize the rest of linguistic information associated to it: Lexicalizations, Senses, Usage Contexts, Sources and Notes.

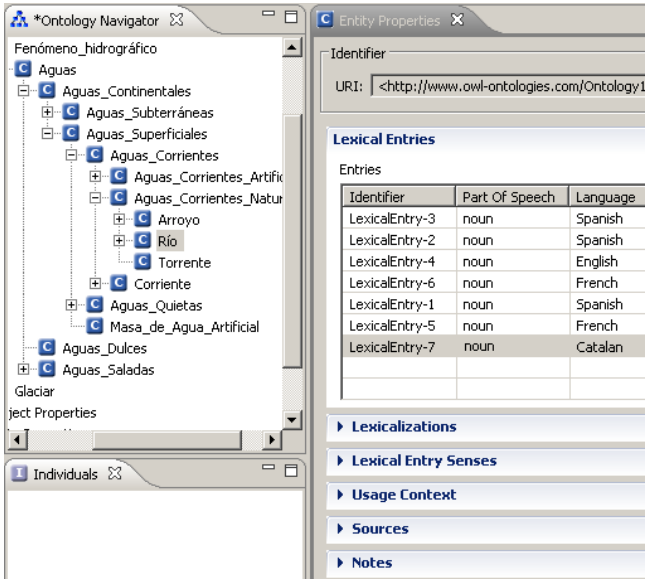

Figure 6. Linguistic Information associated to the concept *Río* in the LIR API (supported by LabelTranslator)

The three Lexical Entries in Spanish (*Río*, *Curso de agua principal*, and *Curso fluvial*) are related by means of the hasSynonym relation (see Figure 7 for Lexical Entry Relationships). The differences in use depending on register (formal vs. informal) are explained in the Note class. With the new extension to the LIR that we propose in this paper, the Senses of these Lexical Entries could additionally be related by an equivalence relation (isEquivalentTo).

Then, the three Lexical Entries in Spanish are related to the Lexical Entry in English (*River*), the one in Catalan (*Riu*), and the last two in French (*Rivière* and *Fleuve*) by means of the hasTranslation relation (see Figure 7). The Lexical Entry in English and the Lexical Entries in Spanish are considered equivalents in meaning, and the same happens with the Catalan equivalent. Therefore, their senses could also be related by the equivalence relation (isEquivalentTo).

[7] http://neon-toolkit.org/wiki/LabelTranslator

[8] http://neon-toolkit.org/

[9] The reason for not obtaining correct translations for some ontology terms may be due to the fact that the resources currently accessed by the system are quite general.

[10] For instance, the *Diccionari de l'Enciclopèdia Catalana* for the Catalan language (http://www.enciclopedia.cat), and the *Dictionnaire français d'hydrologie* for the French language (http://www.cig.ensmp.fr/~hubert/glu/indexdic.htm)

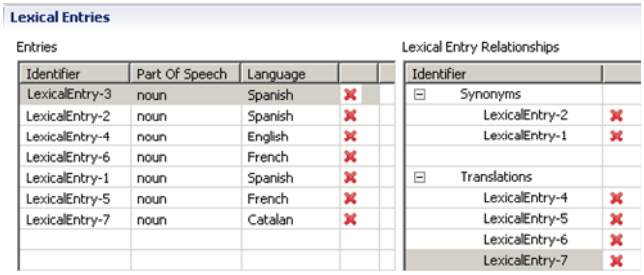

Figure 7. *Synonymy* and *Translation* Relationships among Lexical Entries

However, the two French Lexical Entries represent two more specific concepts that would stay in a relation of subsumption with the Spanish *Río*, the Catalan *Riu*, and the English *River*. This is an example of conceptual mismatch. The French understanding of river has a higher granularity level and identifies two concepts which are intensionally more specific, and extensionally do not share instances. These concepts are *Rivière* and *Fleuve*. According to the specialized resources accessed, *Rivière* is defined as a stream of water of considerable volume that flows into the sea or into another stream, and *Fleuve* is defined as a stream of water of considerable volume and length that flows into the sea. Therefore, in order to make explicit those differences in meaning, we relate them to two different Senses, and provide a definition in natural language for each of them (see Figure 6 for the Definition of *Rivière* in French). Then, with the new functionality of the LIR, we would establish a relation of subsumption between these two senses and the Spanish, English, and Catalan senses for *Río*, *River*, and *Riu* (isSubsumedBy).

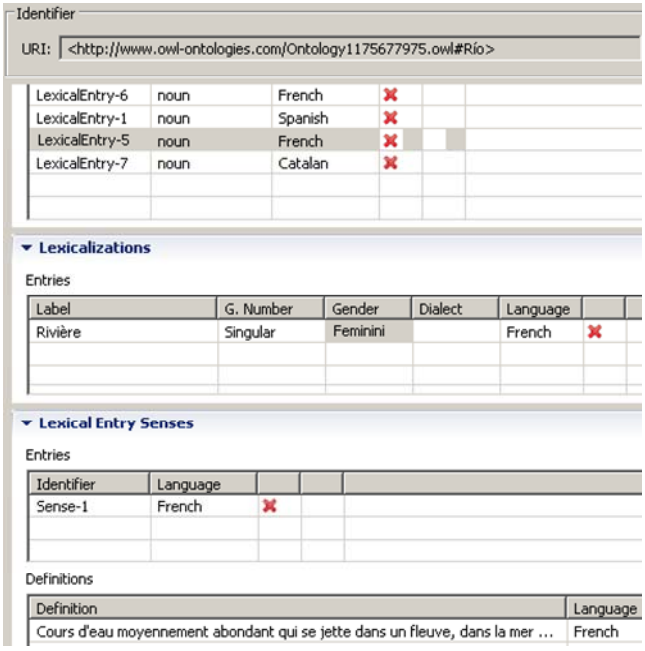

Figure 8. Lexicalization *Rivière* and its related Sense-1 and Definition in French

This further specification of the *isRelatedTo* relation among Senses allows accounting for categorization discrepancies among languages, which are not simply motivated by the fact that there are more lexicalizations in one language than in another, but by the different granularity levels that cultures make of the same world phenomenon. One could argue that these language specificities are only captured in the terminological layer of the ontology, but not in the conceptual model. However, this may suffice for certain ontology-based tasks such as information extraction or verbalization, whereas it may be insufficient for others. In that sense, a modification of the conceptualization to adapt the specificities of a certain language could be directly carried out by considering the lexical and terminological information contained in the LIR.

7. Conclusions

The aim of this paper has been twofold. On the one hand, we have discussed the difficulties involved in the interoperability of resources in the same domain created in different cultural settings, specifically because of the different granularity levels in which world phenomena are dealt with. We have described and illustrated the problematic issues of so called cultural dependent domains taking as example concepts of the hydrographical domain. On the other hand, our objective has been to compare two modalities for the representation of multilingual information in ontologies, with the aim of emphasizing the benefits of associating complex and sound lexical models to ontological knowledge. To achieve this we have presented in detail two versions of a multilingual ontology of the hydrographical domain, *hydrOntology*. The first version shows the representation possibilities offered by the OWL formalism to account for the multilingual information associated to ontology concepts in two languages: English and Spanish. The second version describes the representation possibilities of the Linguistic Information Repository (LIR), a proprietary model designed to associate lexical and terminological information in different languages to domain ontologies. Thanks to such a portable model, the lexical information can be structured for better exploitation purposes of ontology-based applications, and can account for linguistic and cultural discrepancies among languages.

8. ACKNOWLEDGMENTS

This work is supported by the Spanish R&D Project *Geobuddies* (TSI2007-65677C02) and the European Project *Monnet* (FP7-248458). We would also like to thank Óscar Corcho for valuable comments on a draft version of the paper.

9. REFERENCES

[1] Aguado de Cea, G., Montiel-Ponsoda, E., Ramos, J. A. (2007) Multilingualidad en una aplicación basada en el conocimiento TIMM, Monográfico para la revista SEPLN

[2] Arpírez JC, Gómez-Pérez A, Lozano A, Pinto HS (ONTO)2Agent: An ontology-based WWW broker to select ontologies. In: Gómez-Pérez A, Benjamins RV (eds) ECAI'98 Workshop on Applications of Ontologies and Problem-Solving Methods. Brighton, (UK), 1998, pp 16–24.

[3] Barrasa, J. Modelo para la definición automática de correspondencias semánticas entre ontologías y modelos relacionales. PhD Thesis, UPM, Madrid, Spain. 2007.

[4] Buitelaar, P., Cimiano, P., Haase, P. and Sintek, M. Towards Linguistically Grounded Ontologies. In Proceedings of the 6th Annual European Semantic Web Conference (ESWC2009), 111-125, 2009.

[5] DiMarco, Ch. Hirst, G. and Stede, M. The semantic and stylistic differentiation of synonyms and near-synonyms. In AAAI Spring Symposium on Building Lexicons for Machine Translation, 114–121, Stanford, CA, 1993.

[6] Edmonds, P. and Hirst, G. Near-synonymy and lexical choice. In Computational Linguistics, 28, 2, MIT Press, 105-144, 2002.

[7] Espinoza, M. Montiel-Ponsoda, E. and Gómez-Pérez, A. Ontology Localization. In Proceedings of the 5th Fifth International Conference on Knowledge Capture (KCAP), 33-40, 2009.

[8] Espinoza, M. Gómez-Pérez, A. and Mena, E. "Enriching an Ontology with Multilingual Information", Proc. ESWC'08, Tenerife (Spain), Springer LNCS, pp. 333-347, 2008.

[9] Euzenat, J. et al., Results of the Ontology Alignment Evaluation Initiative' 09. ISWC workshop on Ontology Matching (OM-2009). 2009.

[10] Gruber T.R. Toward principles for the design of ontologies used for knowledge sharing, International Journal of Human-Computer Studies, 1995, v.43 n.5-6.

[11] Guarino, N.: Formal Ontology and Information Systems, in Guarino, N. (ed.), Proceedings of FOIS98, 1998.

[12] Isaac, A., Summers, E. (Eds.) SKOS Simple Knowledge Organization System Primer. W3C, 2009. http://www.w3.org/TR/skos-primer/

[13] Mark, D. M., and Turk, A. G. Landscape Categories in Yindjibarndi: Ontology, Environment, and Language. In Kuhn, W., Worboys, M., and Timpf, S., Editors, Spatial Information Theory: Foundations of Geographic Information Science, LNCS No. 2825, Springer. 2003, pp. 31-49.

[14] Montiel-Ponsoda, E., Peters, W., Aguado de Cea, G., Espinoza, M. Gómez-Perez, A. and Sini, M. Multilingual and Localization support for ontologies. Technical report, D2.4.2 NeOn Project Deliverable, 2008.

[15] Montiel-Ponsoda, E., Aguado de Cea, G., Gómez-Pérez, A., and Peters, W. Modelling multilinguality in ontologies. In Coling 2008: Companion volume – Posters and Demonstrations, Manchester, UK, 67-70, 2008.

[16] Nowak, J., Nogueras-Iso, J., Peedell, S. Issues of multilinguality in creating a European SDI – The perspective for spatial data interoperability, in: Proceedings of the 11th EC GI & GIS Workshop, ESDI Setting the Framework. Alghero, Italy. 2005.

[17] Oltramari, A and Stellato, A. Enriching ontologies with linguistic content: An evaluation framework. In Proceedings of OntoLex 2008 Workshop at 6th LREC Conference in Marrakech, Morocco, 2008

[18] Peters, W. Gangemi, A. and Villazón-Terrazas, B. Modelling and re-engineering linguistic/terminological resources. Technical report, D2.4.4 NeOn Project Deliverable, 2010.

[19] Peters, W., Montiel-Ponsoda, E., Aguado de Cea, G., and Gómez-Pérez, A. Localizing ontologies in OWL. In Proceedings of the OntoLex07 Workshop at the ISWC in Busan, South Corea, 2007.

[20] Súarez-Figueroa, M.C. and Gómez-Pérez, A. A First Attempt towards a Standard Glossary of Ontology Engineering Terminology. In Proceedings of the 8th International Conference on Terminology and Knowledge Engineering (TKE2008), Copenhagen, 2008.

[21] Suarez-Figueroa, M.C. (coordinator). NeOn Development Process and Ontology Life Cycle. NeOn Project Deliverable 5.3.1 (2007).

[22] Tanasescu, V. Spatial Semantics in Difference Spaces, COSIT 2007, Melbourne, Australia. 2007.

[23] Thomson, M.K. and Béra, R. Relating Land Use to the Landscape Character: Toward an Ontological Inference Tool. In Winstanley, A. C. (Ed): GISRUK 2007, Proceeding of the Geographical Information Science Research UK Conference, Maynooth, Ireland, 2007, pp.83-87

[24] Vilches-Blázquez, L. M., Ramos, J. A., López-Pellicer, F. J., Corcho, O., Nogueras-Iso, J. An approach to comparing different ontologies in the context of hydrographical information. Popovich et al., (eds.): IF&GIS'09. LNG&C Springer. Pages: 193-207, 2009 St. Petersburg, Russia.

[25] Vilches-Blázquez L.M., Bernabé-Poveda M.A., Suárez-Figueroa M.C., Gómez-Pérez A., Rodríguez-Pascual A.F. "Towntology & hydrOntology: Relationship between Urban and Hydrographic Features in the Geographic Information Domain". In Ontologies for Urban Development. Studies in Computational Intelligence, Springer. 2007, vol.61, pp73–84

ANNEX

Table 1. Results of the semi-automatic translation with LabelTranslator

LabelTranslator	1st Candidate Translation	Candidate Translation	No Candidate Translation	Correct Translation
Aguas Continentales		✓		Inland waters
Aguas Subterráneas	✓			Groundwater
Acuífero	✓			Aquifer
Aguas superficiales		✓		Surface waters
Aguas corrientes			×	Flowing waters
Aguas Corrientes artificiales			×	Artificial flowing waters
Acequia		✓		Irrigation ditch
Canal	✓			Channel
Conducto			×	Pipe
Aguas Corrientes naturales			×	Natural flowing water
Arroyo	✓			Brook
Glaciar	✓			Glacier
Charca		✓		Pond
Torrente	✓			Torrent
Embalse	✓			Reservoir
Afluente		✓		Tributary
Guadi			×	Wadi
Aguas de transición			×	Transitional waters
Terma			×	Hot spring
Aguas costeras		✓		Coastal water
Mar	✓			Sea
Ribera	✓			Riverbank