Multilingual Lexical Support for the SEMbySEM project.

Ingrid Falk, Samuel Cruz-Lara, Nadia Bellalem, Tarik Osswald, Vincent Herrmann,

Multilingual Lexical Support for the SEMbySEM Project

Ingrid Falk, Samuel Cruz-Lara, Nadia Bellalem, Tarik Osswald, Vincent Herrmann

Centre de Recherche INRIA Grande Est, Nancy Universités, LORIA
{ingrid.falk, samuel.cruz-lara, nadia.bellalem}@loria.fr

Abstract

In this paper we describe how multilingual linguistic and lexical information is stored and accessed within the framework of the SEMbySEM project. The SEMbySEM project is dedicated to defining tools and standards for the supervision and management of complex and dynamic systems by using a semantic abstract representation. To provide the project with multilingual linguistic and lexical information and in order to achieve an appropriate, flexible, reusable and accurate representation of this information we chose the Linguistic Information Repository representation (Peters et al., 2009) model and adapted it to our needs. In this paper we discuss the rationale for this choice, describe its implementation and also the employment of other linguistic standards.

1. The SEMbySEM project.

The SEMbySEM project aims at providing a framework for universal sensors management using semantic representations. A detailed description can be found in (Brunner et al., 2009b), here we give a brief overview and concentrate on the aspects related to language and linguistic information. A sensor system supervises and manages the data coming from various sensors with varying technical specifications and placed on various objects. The sensors collect and transmit data and a sensor management system must make sense of and visualise this data. To achieve this the SEMbySEM system is organised in a three layered architecture. The interaction with the sensors (registering and processing events from the sensors) is done in the basic layer, the *Façade Layer*. The information from the sensors is unified and processed and may then trigger an update of the semantic model of the system. The semantic model together with a rule system make up the middle layer, the *Core Layer*. End-users connect to the system through the top layer, the *Visualisation Layer*. They have access to tailored view points designed by expert users and HMI experts through which the data from the semantic model is displayed. From the linguistic point of view the relevant modules are the *Core* and the *Visualisation Layer*. The semantic representation is based on a business-oriented model, the **MicroConcept** model (Brunner et al., 2009a). It was decided against OWL and Description Logic which are habitually employed to represent semantic information in this setting (Brunner et al., 2009b) because of its beeing difficult to handle by business users and its deficiencies in expressing some specific business needs. However, the **MicroConcept** model also uses existing standards and it is therefore possible to leverage standards and methods developed for OWL as for *eg.* the lexicalisation tools to be discussed later in this paper.

1.1. Linguistic needs in SEMbySEM

SEMbySEM needs (multilingual) linguistic information

- on the conceptual level, the *Core Layer* (cf. 2.) and
- on the GUI or visualisation level (cf. 3.)

2. Linguistic information on the conceptual level.

The most common way to provide linguistic and lexical information to a conceptualisation is by using the `rdfs:label` and `rdfs:comment` tags with `xml:lang` attributes. However, this approach, albeit presumably sufficiently expressive for SEMbySEM needs

- is only suitable when there are one to one equivalents for the ontology elements in each language and can not account for any conceptualisation mismatches,
- is not user friendly,
- is hardly reusable.

We identified two recent models for representing linguistic information for ontologies: LIR, (Peters et al., 2009) and LexInfo, (Buitelaar et al., 2009). In both models the linguistic information is stored in a lexical ontology and elements of the domain conceptual representation are linked via an ontology relation (or property) to concepts of the lexical ontology. Both lexical ontologies use LMF (the Lexical Markup Framework, (The LMF Working Group, 2008)) as building blocks. However the resulting ontological structures differ not only from a syntactic point of view but also semantically: LexInfo rather emphasises the representation of properties (relations) and in particular the syntax ↔ semantics interface whereas LIR adopts a more traditional lexicographic position, describing translation (partial) equivalents and linguistic phenomena as synonymy.

We finally opted for LIR as representation model for SEMbySEM for the following reasons:

- LIR's lexicographic point of view seemed to fit the SEMbySEM needs better,
- the project seemed more advanced and tested than LexInfo,
- LIR's alignment with other linguistic and lexicographic standards in addition to LMF: TMX, MLIF and XLIFF.

However, due to time constraints and also LIR's complexity the model finally integrated into SEMbySEM had to be further simplified.

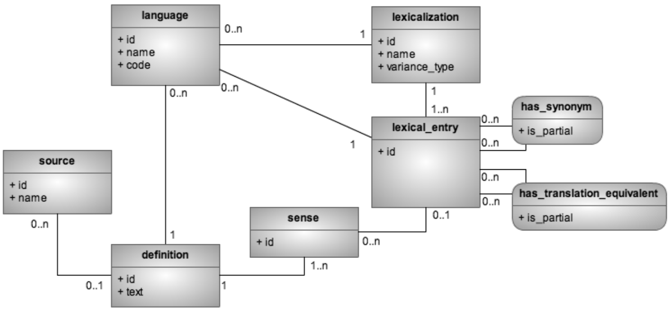

Figure 1: Merise diagram for a simplified LIR-like database.

2.1. Structure of the lexical ontology.

The structure of the simplified ontology is shown in Fig. 1. We (re)used the following ontology classes: LexicalEntry, Lexicalization, Language, Sense, Definition and Source and the following relations (properties): belongsToLanguage, hasSynonym, hasTranslation, hasLexicalization, hasSense, hasDefinition, hasSource. It is possible to express that the synonymy or translation relations hold only partly via the `is_partial` attribute (datatype property). We only reduced the classes and properties of LIR in number, we did not change their semantics.

The classes Lexicalization, Sense, Definition and Source and the relations hasSynonym, hasTranslation and belongsToLanguage are equivalent to elements of the LMF model, whereas the LMF LexicalEntry is more general than the LIR LexicalEntry. The LIR LexicalEntry and hasTranslation are also equivalent to MLIF components.

We will illustrate the model and some of its benefits and limitations in a few examples. First consider the concept *Wagon* as defined in the following snippet:

```
<smc:Concept rdf:about="&sembysem;#AssetTracking/Wagon"/>
```

This concept is linked to the LIR lexical ontology as shown in the following:

```
<smc:Concept
    rdf:about="&sembysem;#AssetTracking/Wagon">
    <lir:hasLexicalEntry rdf:resource="&lexo;#LE-1-En"
                         xml:lang="eng"/>
    <lir:hasLexicalEntry rdf:resource="&lexo;#LE-1-Fr"
                         xml:lang="fr"/>
</smc:Concept>
```

Here the `hasLexicalEntry` elements point to the elements with identifier *LE-1-En* and *LE-1-Fr* in the lexical ontology. These could be represented as follows in the lexical ontology:

```
<lir:LexicalEntry rdf:about="&lexo;#LE-1-En">
 <lir:partOfSpeech>noun</lir:partOfSpeech>
 <lir:belongsToLanguage rdf:resource="&lexo;#English"/>
 <lir:hasLexicalization rdf:resource="&lexo;#Lex-1-En"/>
 <lir:hasSense rdf:resource="&lexo;#Sense-1-En"/>
 <lir:hasTranslation rdf:resource="&lexonto;#LE-1-Fr"/>
</lir:LexicalEntry>
```

This lexical entry describes the word *wagon*, it states that it is an English noun. It's sense is given in a *Sense* instance of the lexical ontology by a definition. The actual lexicalisation (the word string *wagon*) together with possibly other linguistic and terminologic properties is given in the Lexicalization instances of the lexical ontology. In addition, a translation is given through the `hasTranslation` relation, in this case it is the lexical entry *LE-1-Fr*.

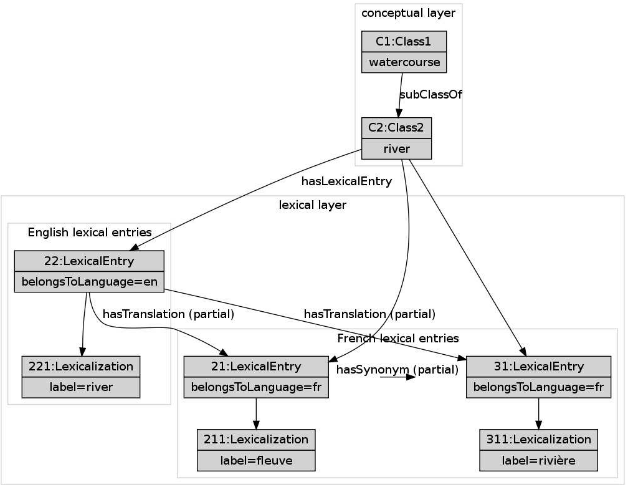

Figure 2: Example of a localisation in case of conceptual mismatches between English and French.

In this simple example, the mapping between ontology elements and lexical entries in several languages is straight forward. However, in cases where the conceptual mapping is different across different languages, the model allows to account for certain discrepancies, as shown in the following (fictitious) example, where one would like to localise to French the concept labeled by the English word *river* (Fig. 2).

The localisation choices made explicit here are the following: The English label *river* is lexicalised in English by the lexical entry *river* and in French by the two lexical entries *fleuve* and *rivière*. However, *fleuve* and *rivière* don't have exactly the same meaning in French, they are both more specific than *river*. This is expressed through the partial synonymy relation and by the fact that the translation relation between *river* and *rivière* and *fleuve* is marked as partial. Note that both synonymy and translation are relations in the lexical ontology. These localisation choices can be easily adapted, refined or reverted.

The next example shows the lexicalisation of a concept where the label consists of several words:

```
<smc:Concept rdf:about=
 "&sembysem;#AssetTracking/WagonMovement_Notification"/>
```

In our simplified model this concept would be associated to one lexical entry corresponding to the entire expression and would also be marked as *mwe*. The latest version of LIR represents a multi-word expression and its components using the LMF ListOfComponents constructs. It is thus possible to link the components to the corresponding lexical entries. However, this multi-word expression also contains relational information reflected in the syntax of the noun phrase: it represents the action of issuing a notification about the movement of a wagon. Within LIR it is currently not possible to capture and represent accurately the corresponding interactions between the words forming the multi-word expression. As this is possible within the LexInfo model, it would be profitable if the two models could be made compatible and merged. Such efforts are currently under way within the Monnet project[1].

2.2. Implementation

The NeOn project also proposes an API which allows to automatically generate a scheletton of the lexical ontology from the domain ontology labels and then to enhance and maintain the lexical ontology. Unfortunately it was not possible to reuse this API due to its complexity and our tight schedule. Therefore, the lexical ontology is currently developed and maintained at the LORIA as a database which can be exported to the OWL or MLIF format. The designer of the conceptual SEMbySEM model is in most cases located elsewhere and uses a web service to require lexical information from the lexical ontology. More specifically, the designer enters a word in natural language and is returned, via the web service the identifiers of the *LexicalEntrie*s in the lexical ontology for the corresponding word. This information is returned in the MLIF format.

3. Linguistic information on the visualisation level.

While on the conceptual level the linguistic and lexical information provides multilingual support, on the visualisation level lexicalisation and translation activities pertain to a more traditional localisation task. SEMbySEM's visualisation layer consists of end-user interfaces displaying and giving access to elements of the core semantic representation. The end-user interfaces are designed by HMI experts in a language independent way. Currently the data format used is XUL, the XML User Interface Language developed by the Mozilla project. It has no formal specification and does not inter-operate with non-Gecko implementations. However, it uses an open source implementation of Gecko and relies on multiple existing web-standards and web-technologies. It was chosen because there was no other suitable standard or norm available.

Language dependant data (ie. the strings labeling and describing the elements of the visual user-interface) are provided in a file in the MLIF format (The MLIF Working Group, 2010). The Multi Lingual Information Framework (MLIF) is a standard unter development with the ISO/TC37/SC4 group. Its objective is to provide a generic platform for modelling and managing multilingual information in various domains while also providing strategies for the inter-operability and/or linking of other formats of interest for localisation and translation including for example TMX and XLIFF.

Finally, at run time the XUL description containing links to the corresponding MLIF components and the MLIF information are combined to render the user-interface in the end-user's language.

4. Conclusion

In this paper we report about efforts to provide linguistic and lexical information to the SEMbySEM project, whose aim it is to implement a sensor supervision and management framework based on an semantic representation. Linguistic and lexical information intervenes at two levels: First it is attached to the conceptual representation through a lexical ontology based on LMF and aligned with other linguistic and lexical standards. Thus conceptual and lexical representations can be developed and maintained separately while allowing for a flexible and accurate coupling. Second, language support is necessary at the visualisation level for the localisation of the end-user interfaces. Here the user-interface itself is specified in a language independent manner using XUL and linguistic information is provided through the MLIF format. We describe when and where it was possible to use existing or emerging standards or best practices and discussed arising issues.

5. References

J. S. Brunner, J. Beck, P. Gatellier, J. F. Goudou, I. Falk, S. Cruz-Lara, and N. Bellalem. 2009a. Micro-concept: Model reference. Technical report, D2.3v1.4 SEMbySEM working draft.

Jean-Sébastien Brunner, Jean-François Goudou, Patrick Gatellier, Jérôme Beck, and Charles-Eric Laporte. 2009b. SEMbySEM: a Framework for Sensors Management. In *1st International Workshop on the Semantic Sensor Web (SemSensWeb 2009)*, June 1st.

Paul Buitelaar, Philipp Cimiano, Peter Haase, and Michael Sintek. 2009. Towards Linguistically Grounded Ontologies. In *The 6th Annual European Semantic Web Conference (ESWC2009)*, Heraklion, Greece.

W. Peters, M. Espinoza, E. Montiel-Ponsoda, and M. Sini. 2009. Multilingual and localization support for ontologies. Technical report, D2.4.3 NeOn Project Deliverable.

The LMF Working Group. 2008. Language Resource Management - Lexical Markup Framework (LMF). Technical report, ISO/TC 37/SC 4 N453 (N330 Rev. 16).

The MLIF Working Group. 2010. MultiLingual Information Framework. Technical report, ISO/DIS 24616. `http://mlif.loria.fr/`.

[1] `http://cordis.europa.eu/fp7/ict/language-technologies/project-monnet_en.html`